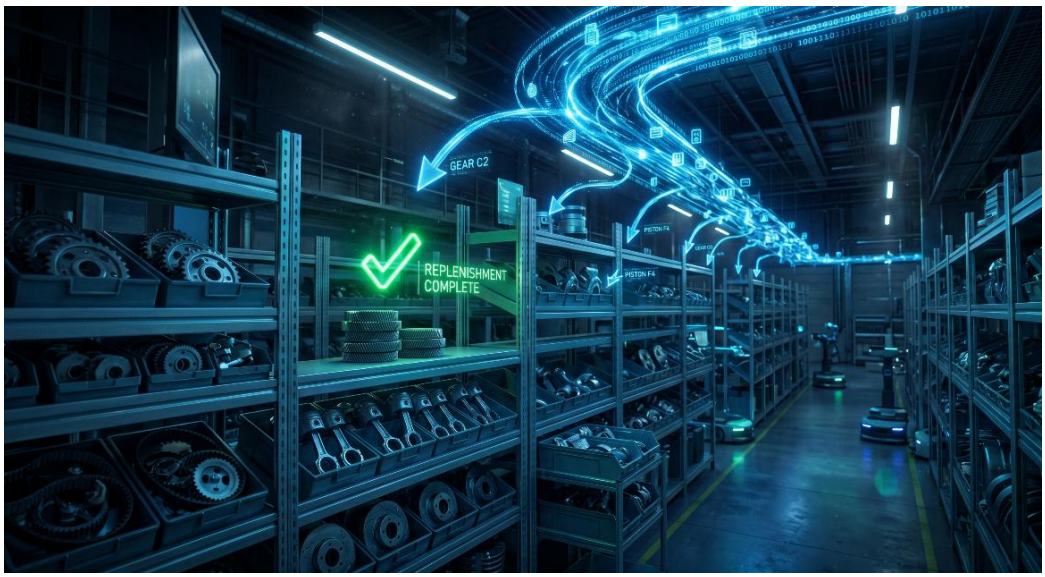
ΔΙΕΘΝΕΣ ΠΑΝΕΠΙΣΤΗΜΙΟ ΤΗΣ ΕΛΛΑΔΟΣ

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΗΛΕΚΤΡΟΝΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«ΠΡΟΒΛΕΨΗ ΖΗΤΗΣΗΣ ΚΑΙ ΒΕΛΤΙΣΤΗ ΑΝΑΠΛΗΡΩΣΗ ΑΠΟΘΕΜΑΤΩΝ ΣΤΗΝ ΑΓΟΡΑ ΑΝΤΑΛΛΑΚΤΙΚΩΝ ΑΥΤΟΚΙΝΗΤΟΥ ΜΕΣΩ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ»



Των φοιτητών
Κωνσταντίνου Αϊτσίδη,
Περικλή Βουτσά
Αρ. Μητρώου: 03/2024, 05/2024

Επιβλέπων
Περικλής Χατζημίσιος
Καθηγητής

Θεσσαλονίκη, 11 Ιουνίου 2026

Τίτλος Δ.Ε.: Πρόβλεψη ζήτησης και βέλτιστη αναπλήρωση αποθεμάτων στην αγορά ανταλλακτικών αυτοκινήτου μέσω Μηχανικής Μάθησης

Κωδικός Δ.Ε.: 25308

Ονοματεπώνυμα φοιτητών: Κωνσταντίνος Αϊτσίδης, Περικλής Βουτσάς

Ονοματεπώνυμο εισηγητή: Περικλής Χατζημίσιος

Ημερομηνία ανάληψης Δ.Ε.: 30/08/2025

Ημερομηνία περάτωσης Δ.Ε.: 11/06/2026

Βεβαιώνω ότι είμαι ο συγγραφέας αυτής της εργασίας και ότι κάθε βοήθεια την οποία είχα για την προετοιμασία της είναι πλήρως αναγνωρισμένη και αναφέρεται στην εργασία. Επίσης, έχω καταγράψει τις όποιες πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, εικόνων και κειμένου, είτε αυτές αναφέρονται ακριβώς είτε παραφρασμένες. Επιπλέον, βεβαιώνω ότι αυτή η εργασία προετοιμάστηκε από εμένα προσωπικά, ειδικά ως διπλωματική εργασία, στο Τμήμα Μηχανικών Πληροφορικής και Ηλεκτρονικών Συστημάτων του ΔΙ.ΠΑ.Ε.

«Αφιερώνω την παρούσα εργασία σε όλη μου την οικογένεια, για την υποστήριξη, την συμπαράσταση, την κατανόηση και την υπομονή τους ώστε να μπορέσω να ανταπεξέλθω σε αυτή την απαιτητική αλλά όμορφη διαδρομή.

Στην Άννα, την Μαρία, και την Χρυσούλα…. α!, και στον Κούπερ….

Κωνσταντίνος Αϊτσίδης»

«Αφιερώνεται στη σύζυγό μου και την κόρη μου για την απεριόριστη υπομονή τους και την υποστήριξη που μου παρείχαν κατά το απαιτητικό ταξίδι της ερευνητικής αναζήτησης.

Στην Αθηνά και τη Νικολέτα…»

Περικλής Βουτσάς

Πρόλογος

Η παρούσα διπλωματική εργασία σηματοδοτεί την ολοκλήρωση ενός απαιτητικού και δημιουργικού κύκλου σπουδών. Η επιλογή του συγκεκριμένου θέματος, το οποίο αφορά την πρόβλεψη της μελλοντικής ζήτησης και τη βελτιστοποίηση αναπλήρωσης αποθεμάτων στη δευτερογενή αγορά αυτοκινήτου, πηγάζει από το έντονο ενδιαφέρον μας για την εφαρμογή προηγμένων τεχνικών Μηχανικής Μάθησης σε σύνθετα επιχειρησιακά προβλήματα. Η αντιμετώπιση της ακραίας σποραδικότητας των δεδομένων και του φαινομένου της λογοκριμένης ζήτησης (*demand censoring*) αποτέλεσε μια σημαντική τεχνική και ερευνητική πρόκληση.

Το μεγαλύτερο όφελος από αυτή τη διαδικασία ήταν η ευκαιρία να γεφυρώσουμε το χάσμα ανάμεσα στην αλγοριθμική θεωρία και την πρακτική εφοδιαστική αλυσίδα. Μέσω της ανάπτυξης ενός ολοκληρωμένου συστήματος πρόβλεψης και αναπλήρωσης, αποκτήσαμε πολύτιμες γνώσεις στη διαχείριση μεγάλων βιομηχανικών δεδομένων, την υλοποίηση εξειδικευμένων μοντέλων Τεχνητής Νοημοσύνης και τον μαθηματικό υπολογισμό αποθεμάτων ασφαλείας. Η εμπειρία αυτή ενίσχυσε καθοριστικά τις αναλυτικές και τεχνικές μας δεξιότητες με σκοπό την προετοιμασία μας για τις προκλήσεις του σύγχρονου ψηφιακού επιχειρηματικού περιβάλλοντος.

Περίληψη

Η διαχείριση αποθεμάτων στη δευτερογενή αγορά ανταλλακτικών αυτοκινήτου αποτελεί ένα από τα πλέον απαιτητικά προβλήματα της εφοδιαστικής αλυσίδας εξαιτίας της δομικά διαλείπουσας και λογοκριμένης (*censored*) ζήτησης. Στην παρούσα διπλωματική εργασία αναπτύσσεται ένα ολοκληρωμένο σύστημα πρόβλεψης ζήτησης απευθείας σε οκτώ χρονικούς ορίζοντες (15 έως 365 ημερών) και υποστήριξης αποφάσεων αναπλήρωσης για ένα σύνολο περίπου 72.000 κωδικών ειδών μιας ελληνικής εταιρείας που δραστηριοποιείται στη δευτερογενή αγορά αυτοκινήτου (*automotine aftermarket*).

Η τελική ροή εργασιών (*pipeline*) πρόβλεψης προέκυψε κατόπιν μιας διαφανούς και επαναληπτικής ερευνητικής διαδρομής, κατά την οποία αξιολογήθηκαν ποικίλες αρχιτεκτονικές και τεχνολογίες. Η ροή αυτή συνδυάζει μηχανική εξαγωγής χαρακτηριστικών βάσει στιγμιοτύπων (*snapshot-based*) που αποτρέπει τη διαρροή μελλοντικής πληροφορίας, αρχιτεκτονική Εμποδίου (Hurdle) δύο σταδίων με χρήση δενδρικών μοντέλων για τη διαχείριση των μηδενικών παρατηρήσεων, συσσωρευτική γενίκευση μέσω μοντέλου μετα-μάθησης (*meta-learner stacking ensemble*), στρωματοποιημένη βαθμονόμηση (*stratified calibration*) για τη θωράκιση των κρίσιμων προϊόντων και ρητή αντιμετώπιση της λογοκριμένης ζήτησης (*censored demand*) μέσω σταθμισμένης εκπαίδευσης.

Το σύστημα επιτυγχάνει σφάλμα WAPE από 1,08 έως 0,54 με καθολικό λόγο μεροληψίας πρόβλεψης (*forecasting bias ratio*) 0,95–1,04 σε όλους τους ορίζοντες πρόβλεψης. Το ανώτατο όριο των διαστημάτων πρόβλεψης τροφοδοτεί τον μη-παραμετρικό υπολογισμό των σημείων αναπαραγγελίας και των αποθεμάτων ασφαλείας. Όσον αφορά την επιχειρησιακή αξία, επιτυγχάνεται αξιόπιστος εντοπισμός των ειδών υψηλού τζίρου με κάλυψη (*revenue capture*) άνω του στην πλειοψηφία των οριζόντων πρόβλεψης. Παράλληλα, παρατηρείται πολύ καλή επίδοση στα είδη της «μακράς ουράς» που οδηγεί στη μείωση της συνολικής αξίας της αποθήκης και, κατ' επέκταση, στην εξασφάλιση μεγαλύτερης ρευστότητας.

Επίσης, αναπτύχθηκε η αυτόνομη διαδικτυακή εφαρμογή «FM_App», η οποία αποτελεί τη μοναδική διεπαφή του συστήματος με τον τελικό χρήστη. Μέσω αυτής γίνεται η παραμετροποίηση και η εκτέλεση της υπολογιστικής ροής πρόβλεψης. Τα αποτελέσματα αποθηκεύονται στην εφαρμογή και μετατρέπονται σε έτοιμες, εξατομικευμένες προτάσεις παραγγελιών γεφυρώνοντας το χάσμα μεταξύ Μηχανικής Μάθησης και καθημερινής εμπορικής πρακτικής.

«Demand Forecasting and Optimal Inventory Replenishment in the Automotive Aftermarket Using Machine Learning»

«Konstantinos Aitsidis, Periklis Voutsas»

Abstract

Inventory management in the automotive aftermarket is one of the most challenging problems in supply chain operations due to the inherently intermittent and censored nature of demand. This thesis develops an end-to-end demand forecasting and replenishment decision support system for approximately 72,000 stock-keeping units (SKUs) belonging to a Greek company operating in the automotive aftermarket. The forecasting pipeline produces direct multi-horizon predictions across eight planning horizons, ranging from 15 to 365 days.

The final forecasting workflow was obtained through a transparent and iterative research process in which multiple architectures and technologies were evaluated. The resulting pipeline combines snapshot-based feature engineering that structurally prevents data leakage, a two-stage Hurdle architecture built on tree ensembles to handle the prevalence of zero observations, stacked generalization via a meta-learning ensemble, stratified calibration to protect commercially critical products and an explicit treatment of censored demand through weighted training.

The proposed system achieves a WAPE ranging from 1.08 to 0.54, with a global forecasting bias ratio between 0.95 and 1.04 across all forecast horizons. The upper bound of the prediction intervals is used as input to a non-parametric calculation of reorder points and safety stock levels. In operational terms, the system provides reliable identification of high-revenue items resulting in an over 93% revenue capture across most forecast horizons. At the same time, it performs strongly on long-tail items, contributing to a reduction in total inventory value and, consequently, improved liquidity.

In addition, an autonomous web-based application, 'FM_App', was developed as the system's sole user interface. Users can configure and execute the forecasting workflow through the application. Results are stored within the application cache and transformed into ready-to-use, customized replenishment recommendations, bridging the gap between Machine Learning and day-to-day commercial practice.

Ευχαριστίες

Θα θέλαμε να ευχαριστήσουμε τον Καθηγητή Περικλή Χατζημίσιο για την εμπιστοσύνη που μας έδειξε με την ανάθεση του συγκεκριμένου θέματος και την καθοδήγηση του κατά την εκπόνηση της παρούσης διπλωματικής εργασίας.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Συντομογραφίες

ΔΕ	Διπλωματική Εργασία
ΔΙΠΑΕ	Διεθνές Πανεπιστήμιο Ελλάδος
ΒΔ	Βάση Δεδομένων
SKU	Stock Keeping Unit
GNN	Graph Neural Network
ROP	Re-Order Point
EOQ	Economic Order Quantity
RF	Random Forest
ET	Extra Trees
RC	Reservoir Computing
SS	Safety Stock
PI	Prediction Interval
UI	User Interface
ERP	Enterprise Resource Planning
WMS	Warehouse Management System

Κεφάλαιο 1ο: Εισαγωγή

1.1 Πλαίσιο και ερευνητική αφορμή

Η διαχείριση αποθεμάτων στη δευτερογενή αγορά ανταλλακτικών αυτοκινήτου συνιστά ένα από τα πλέον απαιτητικά προβλήματα εφοδιαστικής αλυσίδας. Αντίθετα με τα συχνά πωλούμενα προϊόντα της τυπικής καταναλωτικής αγοράς λιανικής, τα οποία διακινούνται σε ρυθμικά και εύκολα προβλέψιμα πρότυπα, η ζήτηση ανταλλακτικών είναι δομικά **ετερογενής**, **αραιή** και **μη προβλέψιμη** σε επίπεδο μεμονωμένου κωδικού είδους. Ο συνδυασμός χιλιάδων κωδικών ειδών (*Stock Keeping Units* – SKU) με εξαιρετικά διαφορετικά προφίλ κίνησης, από τα 20–30 κρίσιμα και ταχέως κινούμενα είδη (*fast movers*) που αντιπροσωπεύουν δυσανάλογα μεγάλο τμήμα του τζίρου, έως τις δεκάδες χιλιάδες προϊόντα που πωλούνται μία έως λίγες φορές το χρόνο (*slow movers*), καθιστά αδύνατη την εφαρμογή ενιαίων κλασικών μεθόδων πρόβλεψης και διαχείρισης αποθεμάτων [1], [2], [3].

Η κατηγορία αυτή, γνωστή στη βιβλιογραφία ως **διαλείπουσα ζήτηση** (*intermittent demand*), χαρακτηρίζεται από μεγάλα χρονικά διαστήματα μεταξύ των συναλλαγών, ακανόνιστο μέγεθος παραγγελίας και εξαιρετικά υψηλή εμφάνιση μηδενικών παρατηρήσεων [4], [5]. Σε ένα τυπικό χαρτοφυλάκιο χονδρεμπορίου ανταλλακτικών, ποσοστό άνω του 90% των κωδικών ειδών εμπίπτει σε αυτή την κατηγορία (βλ. κεφάλαιο 4), το οποίο αποτελεί ένα εύρημα που επιβεβαιώνεται απόλυτα και από τα δεδομένα της παρούσας μελέτης. Η κατάσταση περιπλέκεται περαιτέρω από φαινόμενα **λογοκριμένης ζήτησης** (*censored demand*), η οποία παρατηρείται κατά τις περιόδους ανεπάρκειας αποθέματος (*stock-out*). Οι απολεσθείσες πωλήσεις δεν καταγράφονται στα πληροφοριακά συστήματα, με αποτέλεσμα τα ιστορικά δεδομένα να υποτιμούν συστηματικά την «αληθινή» ζήτηση [6], [7], [8].

Παράλληλα, η επιχειρησιακή πραγματικότητα επιβάλλει έναν διπλό, και εν μέρει ανταγωνιστικό, στόχο. Αφενός, η **υψηλή διαθεσιμότητα** (αποφυγή ελλείψεων προϊόντων) είναι κρίσιμη για τη διατήρηση της πελατείας και την αξιοπιστία της υπηρεσίας. Αφετέρου, το **πλεονάζον απόθεμα** δεσμεύει ρευστότητα, αυξάνει το κόστος αποθήκευσης και εγκυμονεί κινδύνους απαξίωσης (*obsolescence*). Ο συνδυασμός αυτός, συχνά αποκαλούμενος στη βιβλιογραφία ως η "ένταση μεταξύ επιπέδου εξυπηρέτησης και επένδυσης σε αποθέματα" (*tension between service level and inventory investment*), αποτελεί το κεντρικό ζήτημα των ολοκληρωμένων συστημάτων πρόβλεψης και αποθεμάτων (*forecasting–inventory systems*) [9], [10].

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο συνεργασίας με **ελληνική εταιρεία χονδρικής εμπορίας ανταλλακτικών αυτοκινήτων**, η οποία διαχειρίζεται ένα χαρτοφυλάκιο 79.363 κωδικών προϊόντων, εκ των οποίων οι 72.310 παρουσιάζουν τουλάχιστον μία πώληση κατά την περίοδο 2021–2025. Τα δεδομένα αυτά αποτελούν πραγματικές συναλλαγές πωλήσεων, αναπλήρωσης και απογραφής συνιστώντας ένα από τα λίγα σύνολα τέτοιας κλίμακας στον κλάδο των ανταλλακτικών οχημάτων που έχουν αναλυθεί.

1.2 Δήλωση του προβλήματος

Ο **κεντρικός στόχος** της παρούσας εργασίας διατυπώνεται ως εξής:

Σχεδιασμός και υλοποίηση ενός ολοκληρωμένου συστήματος υποστήριξης αποφάσεων, το οποίο: (α) προβλέπει τη μελλοντική ζήτηση ανά κωδικό προϊόντος για πολλαπλούς χρονικούς ορίζοντες, και (β) μετατρέπει αυτές τις προβλέψεις, σε συνδυασμό με στοιχεία χρόνου προμήθειας (lead time) και δυναμικού

αποθέματος ασφαλείας, σε έτοιμες προτάσεις παραγγελιών αναπλήρωσης ανά ορίζοντα, ανά κωδικό και ανά προμηθευτή.

Η επίλυση του παραπάνω προβλήματος εντοπίζεται σε τρία αλληλένδετα επίπεδα:

Επίπεδο 1 – Πρόβλεψη ζήτησης σε μεγάλη κλίμακα και σε πολλαπλούς ορίζοντες: Η παρούσα εργασία απαιτεί προβλέψεις για περίπου 72.000 κωδικούς ειδών, ταυτόχρονα, σε οκτώ διαφορετικούς ορίζοντες (από 15 έως 365 ημέρες) με αποκλειστική χρήση ιστορικών δεδομένων διασφαλίζοντας την απουσία διαρροής μελλοντικής πληροφορίας (*data leakage*). Η διαλείπουσα φύση της ζήτησης, η ακανόνιστη εναλλαγή του προϊοντικού καταλόγου (ετήσια είσοδος περί των 14.000 νέων SKUs και κατάργηση περίπου 12.000) και η λογοκριμένη ζήτηση αποτελούν τις βασικές αλγοριθμικές προκλήσεις.

Επίπεδο 2 – Βαθμονόμηση αβεβαιότητας για τη διαχείριση αποθεμάτων: Για τον υπολογισμό του αποθέματος ασφαλείας και των σημείων αναπαραγγελίας, η σημειακή πρόβλεψη (*point forecast*) δεν επαρκεί. Οπότε, απαιτείται η εκτίμηση ολόκληρης της κατανομής της ζήτησης κατά τον χρόνο προμήθειας (D_L). Αυτό επιτρέπει τη μη-παραμετρική εκτίμηση του σημείου αναπαραγγελίας ως ποσοστημόριο $q_\alpha(D_L)$, το οποίο αποτελεί μια προσέγγιση δραστικά πιο ρεαλιστική για τη διαλείπουσα ζήτηση συγκριτικά με την κλασική Γκαουσιανή παραδοχή περί υπόθεσης κανονικής κατανομής [11].

Επίπεδο 3 – Ολοκλήρωση σε επιχειρησιακό εργαλείο: Το σύστημα πρέπει να είναι επαναλήψιμο, αξιολογήσιμο και άμεσα χρησιμοποιήσιμο από το τμήμα αγορών. Οφείλει να παράγει αναγνώσιμα αρχεία προτάσεων ανά κωδικό είδους και προμηθευτή γεφυρώνοντας το χάσμα μεταξύ της Μηχανικής Μάθησης και της καθημερινής εμπορικής πρακτικής.

1.3 Ερευνητικά ερωτήματα

Από τη δήλωση του προβλήματος απορρέουν τέσσερα κεντρικά ερευνητικά ερωτήματα, στα οποία η παρούσα εργασία επιχειρεί να δώσει τεκμηριωμένες απαντήσεις:

- **Ερώτημα 1:** Ποιες αρχιτεκτονικές Μηχανικής Μάθησης είναι καταλληλότερες για *άμεση* (*direct*) πρόβλεψη σε πολλαπλούς χρονικούς ορίζοντες υπό ενός περιβάλλοντος μεγάλης κλίμακας με ακραία διαλείπουσα και λογοκριμένη ζήτηση;
- **Ερώτημα 2:** Πώς μπορεί να εκτιμηθεί αξιόπιστα η αβεβαιότητα των προβλέψεων μέσω διαστημάτων εμπιστοσύνης, ώστε να υποστηριχθεί ο μη-παραμετρικός υπολογισμός των σημείων αναπαραγγελίας (ROP);
- **Ερώτημα 3:** Πώς μπορεί η βαθμονόμηση (*calibration*) των προβλέψεων να διατηρηθεί συνεπής τόσο στα κρίσιμα επιχειρησιακά προϊόντα (υψηλός τζίρος) όσο και στη «μακρά ουρά» (*long-tail*) της χαμηλής κίνησης αποφεύγοντας φαινόμενα συστηματικής υπο-εκτίμησης;
- **Ερώτημα 4:** Σε ποιον βαθμό η ενσωμάτωση αυτών των προγνωστικών αλγορίθμων σε μια πλήρη κανονιστική πολιτική αποθεμάτων οδηγεί σε μετρήσιμη επιχειρησιακή βελτίωση εξισορροπώντας το επίπεδο εξυπηρέτησης με τη δεσμευμένη ρευστότητα;

1.4 Η δομή του προτεινόμενου συστήματος

Το προτεινόμενο σύστημα διαρθρώνεται σε **δύο βασικά αλληλένδετα υποσυστήματα**, τα οποία αναλύονται στα Κεφάλαια 5 και 6, αντίστοιχα.

1.4.1 Υποσύστημα 1 – Πρόβλεψη ζήτησης (*Forecasting pipeline*)

Το πρώτο υποσύστημα εκτελεί άμεση πρόβλεψη πολλαπλών οριζόντων (*direct multi-horizon forecasting*) με σταθερό σημείο εκκίνησης (01/01/2025). Τα βασικά χαρακτηριστικά της αρχιτεκτονικής του περιλαμβάνουν:

- **Μηχανική χαρακτηριστικών (*Feature engineering*) βάσει στιγμιοτύπων:** Ο υπολογισμός των χαρακτηριστικών εκτελείται σε αυστηρά καθορισμένα, εξαμηνιαία χρονικά σημεία αποκοπής (για την περίοδο 2021–2024) κατά τη φάση της εκπαίδευσης διασφαλίζοντας, έτσι, την απόλυτη αποτροπή διαρροής μελλοντικής πληροφορίας.
- **Δι-σταδιακή αρχιτεκτονική Εμποδίου (Hurdle *model*):** Η μοντελοποίηση διασπάται σε δύο σκέλη, έναν ταξινομητή (πιθανότητα εμφάνισης ζήτησης) και έναν παλινδρομητή (μέγεθος ζήτησης), αντιμετωπίζοντας με φυσικό τρόπο την κυριαρχία των μηδενικών πωλήσεων.
- **Συσσωρευτική γενίκευση (*Ensemble meta-learning*):** Ετερογενή δενδρικά μοντέλα (Random Forest, Extra Trees, HistGradientBoosting, CatBoost) συνδυάζονται βέλτιστα μέσω ενός μοντέλου μετα-μάθησης (*meta-learner)*, το οποίο εκπαιδεύεται αυστηρά σε εκτός-δείγματος (*out-of-sample*) σφάλματα.
- **Πιθανοκρατικές προβλέψεις:** Παράγονται διαστήματα πρόβλεψης μέσω εμπειρικών καταλοίπων. Τα δεδομένα αυτά τροφοδοτούν τον απευθείας υπολογισμό των αποθεμάτων ασφαλείας (*safety stocks*) και των σημείων αναπαραγγελίας (ROP), τα οποία αποθηκεύονται σε εξωτερικά αρχεία.

1.4.2 Υποσύστημα 2 – Παραγωγή Προτάσεων Αναπλήρωσης

Το δεύτερο υποσύστημα (Κεφάλαιο 6) αποτελεί την επιχειρησιακή προέκταση της αλγοριθμικής ροής λειτουργώντας ως μια ανεξάρτητη διαδικτυακή εφαρμογή (*web-based application*). Αυτή η εφαρμογή συνιστά τη μοναδική διεπαφή του τελικού χρήστη με το σύστημα. Αντλώντας τα έτοιμα δεδομένα (προβλέψεις και αποθέματα ασφαλείας) από τα εξωτερικά αρχεία του αλγοριθμικού αγωγού, η εφαρμογή επιτρέπει στους υπεύθυνους αγορών να:

- **Παραμετροποιούν δυναμικά** τις παραγγελίες τους, καθώς επιλέγουν κρίσιμες μεταβλητές, όπως ο χρονικός ορίζοντας, το επιθυμητό επίπεδο εξυπηρέτησης (*service level*), το σημείο αναφοράς της πρόβλεψης, ή συγκεκριμένους προμηθευτές και ομάδες ειδών.
- **Μετατρέπουν** τις προϋπολογισμένες μετρικές σε έτοιμες, εξατομικευμένες προτάσεις παραγγελιών αναπλήρωσης, προσαρμοσμένες στις τρέχουσες εμπορικές ανάγκες της επιχείρησης.

Με τον τρόπο αυτό, η μαθηματική και αλγοριθμική πολυπλοκότητα της Μηχανικής Μάθησης παραμένει στο παρασκήνιο προσφέροντας στον χρήστη ένα φιλικό, ευέλικτο και ταχύτατο εργαλείο λήψης αποφάσεων.

1.5 Μεθοδολογική προσέγγιση και συνεισφορά

1.5.1 Επαναληπτική ανάπτυξη και διαφάνεια σχεδιασμού

Η ανάπτυξη του τελικού συστήματος δεν ακολούθησε μια απλοϊκή, γραμμική πορεία. Προκειμένου να διασφαλιστεί η βέλτιστη αρχιτεκτονική, αναπτύχθηκαν, αξιολογήθηκαν και συγκρίθηκαν πολλαπλές σύγχρονες προσεγγίσεις. Για παράδειγμα, δοκιμάστηκε εκτενώς η χρήση Γραφικών Νευρωνικών Δικτύων (GNNs) για την εκμάθηση τοπολογικών σχέσεων, καθώς, και η ενσωμάτωση Γλωσσικών Μοντέλων (Transformers) για την εξαγωγή χαρακτηριστικών από τα μεταδεδομένα των ειδών. Η

υιοθέτηση της αρχιτεκτονικής Εμποδίου δύο σταδίων (Hurdle) ως τελικής λύσης αποτέλεσε απόρροια αυστηρής συγκριτικής αξιολόγησης. Όπως καταδείχθηκε πειραματικά, οι προηγμένες προσεγγίσεις Βαθιάς Μάθησης παρουσίασαν δομικές αδυναμίες, καθότι είτε εισήγαγαν συστηματικό θόρυβο στο προγνωστικό σήμα, είτε υπολείπονταν σημαντικά έναντι των δενδρικών συνόλων (*tree ensembles*) ως προς την ικανότητα αποτελεσματικής διαχείρισης της ακραίας πληθώρας μηδενικών παρατηρήσεων (*zero-inflation*). Η παρουσίαση αυτής της εξελικτικής πορείας στην εργασία αποτελεί συνειδητή επιλογή ερευνητικής διαφάνειας αποδεικνύοντας ότι οι τελικές σχεδιαστικές αποφάσεις είναι εμπειρικά και επιστημονικά θεμελιωμένες.

1.5.2 Επιστημονική και επιχειρησιακή συνεισφορά

Η παρούσα εργασία συνεισφέρει στη βιβλιογραφία και την επιχειρησιακή πρακτική σε τέσσερις διακριτούς άξονες:

Πρακτική αντιμετώπιση του φαινομένου της λογοκριμένης ζήτησης.

Σε περιβάλλοντα, όπου οι ελλείψεις αποθεμάτων (*stock-outs*) συνιστούν δομικό – και όχι περιστασιακό – φαινόμενο, τα ιστορικά δεδομένα με μηδενικές πωλήσεις υποδηλώνουν ανεπάρκεια διαθεσιμότητας εξίσου συχνά με την πραγματική απουσία ζήτησης. Η παρούσα εργασία ενσωματώνει ρητά δείκτες εξάντλησης αποθέματος (*out-of-stock flags*) ως χαρακτηριστικά εκπαίδευσης και αντιμετωπίζει τη λογοκριμένη ζήτηση (*censored demand*) εφαρμόζοντας μειωμένη στάθμιση στις αντίστοιχες εγγραφές. Με αυτόν τον τρόπο, η απόκλιση μεταξύ των *καταγεγραμμένων πωλήσεων* και της *αληθούς ζήτησης* αναδεικνύεται ως ένα μεθοδολογικό ζήτημα πρώτης τάξεως, το οποίο αποτελεί μια προσέγγιση που παραμένει εξαιρετικά σπάνια στη σύγχρονη βιβλιογραφία [6], [7], [7][8].

Μεθοδολογικά ορθή εφαρμογή άμεσης πρόβλεψης πολλαπλών οριζόντων από σταθερό σημείο αναφοράς.

Παρότι η άμεση πρόβλεψη σε πολλαπλούς ορίζοντες (*direct multi-horizon forecasting*) αποτελεί γνωστή στρατηγική, η συνδυαστική εφαρμογή της με την κατασκευή χαρακτηριστικών βάσει στιγμιοτύπων (*snapshot-based features*) σε πολλαπλά χρονικά σημεία αποκοπής (*training cutoffs*) παραμένει σπάνια στο πλαίσιο της διαλείπουσας ζήτησης. Η συγκεκριμένη αρχιτεκτονική αποτρέπει δομικά οποιαδήποτε διαρροή μελλοντικής πληροφορίας (*data leakage*). Παράλληλα, αποτελεί τη μοναδική ορθή μεθοδολογική επιλογή για τις αποφάσεις αναπλήρωσης. Ο υπεύθυνος προμηθειών απαιτεί την αθροιστική εκτίμηση της ζήτησης *από τη σημερινή ημέρα* για ένα μελλοντικό διάστημα και όχι μια αλυσίδα ημερήσιων αναδρομικών προβλέψεων.

Κάλυψη του βιβλιογραφικού κενού στη δευτερογενή αγορά ανταλλακτικών αυτοκινήτου.

Η πλειονότητα των εφαρμοσμένων μελετών γύρω από τη διαλείπουσα ζήτηση επικεντρώνεται σε αεροπορικά ή βιομηχανικά εξαρτήματα, ενώ οι στοχευμένες έρευνες στον κλάδο των ανταλλακτικών οχημάτων (*automotive aftermarket*) απουσιάζουν αισθητά [12], [13]. Η παρούσα εργασία παρέχει μια πλήρη εφαρμογή και αξιολόγηση σε ένα πραγματικό σύνολο δεδομένων ~72.000 κωδικών ειδών (SKUs) από ελληνική εταιρεία χονδρικής αποτελώντας ένα από τα ελάχιστα τεκμηριωμένα παραδείγματα τόσο μεγάλης κλίμακας στο συγκεκριμένο πεδίο.

Υιοθέτηση ενιαίας αξιολόγησης με μικτά επιχειρησιακά κριτήρια.

Το πλαίσιο αξιολόγησης του συστήματος συνδυάζει κανονικοποιημένες μετρικές προγνωστικής ακρίβειας (WAPE, MASE) με δείκτες απόδοσης προσανατολισμένους στη διαχείριση αποθεμάτων (*inventory-oriented* KPI*s*, όπως σημεία αναπαραγγελίας, επίπεδα εξυπηρέτησης, αποθέματα

ασφαλείας). Επιπροσθέτως, ενσωματώνει ανάλυση κατά επίπεδα (*stratified analysis*) βάσει της βαθμίδας εμπορικής κρισιμότητας (Top-20, Κλάση A, Κλάσεις B/C). Η προσέγγιση αυτή ανταποκρίνεται στη διαπίστωση ότι η επιχειρησιακή αξία των προβλέψεων δεν καθορίζεται απλώς από την καθολική ακρίβεια του αλγορίθμου, αλλά εξαρτάται άμεσα από την εμπορική σπουδαιότητα του εκάστοτε κωδικού [3], [10], [14], [15].

1.6 Δομή της εργασίας

Η εργασία οργανώνεται σε επτά κεφάλαια, ακολουθώντας τη φυσική ροή από τη θεωρητική θεμελίωση έως την επιχειρησιακή εφαρμογή:

- Το **Κεφάλαιο 2 (Θεωρητικό υπόβαθρο)** εισάγει τις βασικές έννοιες πρόβλεψης χρονοσειρών, τις πολιτικές αποθεμάτων, τη θεωρία της Μηχανικής Μάθησης (δενδρικά μοντέλα, μοντέλα συνδυασμού συνόλων – *ensembles*), καθώς, και προηγμένες τεχνικές αντιμετώπισης της διαλείπουσας ζήτησης και της πιθανοκρατικής αβεβαιότητας.
- Το **Κεφάλαιο 3 (Βιβλιογραφική ανασκόπηση)** χαρτογραφεί τις σύγχρονες τάσεις και τις εμπειρικές μελέτες γύρω από τη δευτερογενή αγορά ανταλλακτικών και τη διαχείριση της διαλείπουσας ζήτησης, γενικότερα, εντοπίζοντας τα ερευνητικά κενά που η παρούσα εργασία καλείται να καλύψει.
- Το **Κεφάλαιο 4 (Στατιστική ανάλυση συνόλου δεδομένων)** παρουσιάζει την ανάλυση των δεδομένων της επιχείρησης τεκμηριώνοντας φαινόμενα όπως η «μακρά ουρά» (*long-tail*) και οι ελλείψεις αποθεμάτων (*stock-outs*), τα οποία καθοδηγούν τις μεθοδολογικές επιλογές.
- Το **Κεφάλαιο 5 (Μεθοδολογία πρόβλεψης ζήτησης)** αποτελεί τον αλγοριθμικό πυρήνα. Αναλύει βήμα προς βήμα τη μεθοδολογία, τις αρχιτεκτονικές που δοκιμάστηκαν και εγκαταλείφθηκαν και παρουσιάζει διεξοδικά το τελικό *Sparse True Forecast Pipeline* μαζί με τα αποτελέσματα αξιολόγησής του.
- Το **Κεφάλαιο 6 (Εφαρμογή παραγωγής προτάσεων παραγγελιών)** περιγράφει τη μετάβαση από την αλγοριθμική πρόβλεψη στην επιχειρησιακή απόφαση. Παρουσιάζεται η ανάπτυξη μιας ανεξάρτητης διαδικτυακής (*web-based*) εφαρμογής που λειτουργεί ως η τελική διεπαφή του συστήματος αναλύοντας τον τρόπο, με τον οποίο ο χρήστης παραμετροποιεί τα δεδομένα, ώστε να λάβει τις τελικές προτάσεις αναπλήρωσης.
- Τέλος, το **Κεφάλαιο 7 (Αποτίμηση και μελλοντικές προκλήσεις)** συνοψίζει τα συμπεράσματα, συζητά τους περιορισμούς του συστήματος και προτείνει κατευθύνσεις για μελλοντική έρευνα.

Κεφάλαιο 2ο: Θεωρητικό υπόβαθρο

2.1 Εισαγωγή

Το παρόν κεφάλαιο επιτελεί έναν θεμελιώδη ρόλο στην οικονομία της παρούσας διπλωματικής εργασίας λειτουργώντας ως το αυστηρό επιστημονικό και μεθοδολογικό υπόβαθρο, επί του οποίου εδράζονται όλα τα επόμενα στάδια της έρευνας. Στόχος του κεφαλαίου δεν είναι η κριτική αποτίμηση ή η συγκριτική αξιολόγηση της υπάρχουσας βιβλιογραφίας, αυτό αποτελεί μια διαδικασία που επιφυλάσσεται αποκλειστικά για το Κεφάλαιο 3 (Βιβλιογραφική ανασκόπηση), αλλά η αναλυτική παρουσίαση και η μαθηματική τυποποίηση των εννοιών, των μοντέλων και των αλγοριθμικών δομών που αποτελούν τα δομικά υλικά της μελέτης.

Με τον τρόπο αυτό, το κεφάλαιο αυτό συνιστά ένα αυτόνομο «εγχειρίδιο αναφοράς» για τον αναγνώστη. Οι θεωρητικές έννοιες που αναλύονται εδώ θα μετατραπούν σε συγκεκριμένα εργαλεία στα επόμενα κεφάλαια. Δηλαδή η στατιστική περιγραφή των χρονοσειρών θα εφαρμοσθεί στην πράξη κατά τη στατιστική ανάλυση του συνόλου δεδομένων (Κεφάλαιο 4), οι αλγόριθμοι μηχανικής μάθησης και τα στατιστικά μοντέλα θα συγκροτήσουν τη μεθοδολογία πρόβλεψης (Κεφάλαιο 5), και οι θεωρητικές κατανομές της ζήτησης κατά τον χρόνο προμήθειας θα μετουσιωθούν σε εφαρμοσμένες πολιτικές παραγγελιών αναπλήρωσης (Κεφάλαιο 6).

Η διάρθρωση του κεφαλαίου ακολουθεί μια γραμμική και λογική ροή, η οποία αντικατοπτρίζει τη φυσική αλληλουχία του προβλήματος της διαχείρισης αποθεμάτων. Ξεκινά από τις γενικές και εισαγωγικές έννοιες της επιστήμης των χρονοσειρών και των μετρικών αξιολόγησης (βλ. ενότητα 2.2) και εξειδικεύεται άμεσα στις επιχειρησιακές ιδιαιτερότητες των συστημάτων αποθεμάτων και της διαλείπουσας ζήτησης ανταλλακτικών (βλ. ενότητα 2.3). Στη συνέχεια, εισάγεται το κρίσιμο στατιστικό φαινόμενο της περικοπής (λογοκρισίας) της ζήτησης λόγω ελλείψεων αποθέματος (βλ. ενότητα 2.4) και, έπειτα, παρουσιάζονται οι κλασικές παραμετρικές μέθοδοι επίλυσης (βλ. ενότητα 2.5).

Ακολούθως, το ενδιαφέρον μετατοπίζεται στο σύγχρονο υπολογιστικό παράδειγμα αναλύοντας τις βασικές αρχές της Μηχανικής Μάθησης (βλ. ενότητα 2.6), τα μοντέλα βασισμένα σε δέντρα απόφασης και τη συσσωρευτική γενίκευση (βλ. ενότητα 2.7), καθώς, και τις αρχιτεκτονικές Βαθιάς Μάθησης (βλ. ενότητα 2.8). Το κεφάλαιο ολοκληρώνεται με τη μελέτη των μοντέλων Εμποδίου δύο σταδίων (βλ. ενότητα 2.9), των τεχνικών πιθανοκρατικής πρόβλεψης και βαθμονόμησης (βλ. ενότητα 2.10), και των αναδυόμενων προσεγγίσεων Reservoir Computing και Μεγάλων Γλωσσικών Μοντέλων (βλ. ενότητα 2.11). Κλείνοντας, λαμβάνει χώρα μια σύντομη σύνοψη (βλ. ενότητα 2.12) που γεφυρώσει το θεωρητικό αυτό πλαίσιο με τη βιβλιογραφική ανασκόπηση που έπεται.

Πίνακας 2.1: Πίνακας συμβολισμών

Σύμβολο / Ακρωνύμιο	**Ερμηνεία και μαθηματική φύση**
t	Διακριτός χρονικός δείκτης (π.χ. $t = 1,2,\dots,T$).
T	Συνολικό μήκος ιστορικού.
h	Ορίζοντας πρόβλεψης (σε ημέρες).
N	Πλήθος παρατηρήσεων ή κωδικών ειδών (SKUs) στο δείγμα ελέγχου.

Σύμβολο / Ακρωνύμιο	**Ερμηνεία και μαθηματική φύση**
y_t	Παρατηρούμενη μεταβλητή (πωλήσεις ή καταγεγραμμένη ζήτηση) στον χρόνο t.
$\bar{y}$	Μέση ιστορική τιμή ζήτησης.
y'_t	Λογαριθμικά μετασχηματισμένη ζήτηση (`log1p`) για σταθεροποίηση διακύμανσης
d_t	Λανθάνουσα, πραγματική ζήτηση στον χρόνο t (ενδέχεται $d_t \geq y_t$).
$\hat{y}_t$	Σημειακή πρόβλεψη (ή προσαρμοσμένη τιμή – *fitted value*) της μεταβλητής για τον χρόνο t.
$\hat{y}_{t+h}$	Σημειακή πρόβλεψη της μεταβλητής για ορίζοντα h βημάτων στο μέλλον.
e_t	Σφάλμα πρόβλεψης στον χρόνο t, οριζόμενο ως $e_t = y_t - \hat{y}_t$.
$\mathbf{x} \in \mathbb{R}^p$	Διάνυσμα χαρακτηριστικών εισόδου (*feature vector*) διάστασης p.
$f_\theta(\mathbf{x})$	Συνάρτηση υπόθεσης/μοντέλου με διάνυσμα ρυθμίσιμων παραμέτρων θ.
L	Χρόνος προμήθειας (*lead time*) από τον προμηθευτή έως την αποθήκη.
D_L	Τυχαία μεταβλητή που αντιπροσωπεύει τη συνολική ζήτηση κατά τον χρόνο προμήθειας.
$\mathcal{L}(\theta)$	Συνάρτηση κόστους ή απώλειας (loss function) προς ελαχιστοποίηση.
$\mathbb{P}(y_t > 0)$	Πιθανότητα εμφάνισης θετικής ζήτησης (Πρώτο στάδιο – Hurdle).
$\mathbb{E}[y_t \mid y_t > 0]$	Υπό συνθήκη αναμενόμενη τιμή του μεγέθους της ζήτησης (Δεύτερο στάδιο).
τ	Κατώφλι απόφασης (*decision threshold*) για την ενεργοποίηση της πρόβλεψης.
τ_h	Σταθερά χρονικής απόσβεσης (decay time constant).
$\mathbb{1}_{\{\cdot\}}$	Ενδεικτική συνάρτηση (indicator function). Λαμβάνει τιμή 1 αν η συνθήκη ισχύει, αλλιώς 0.
w_i / w_m	Βάρος i-οστού δείγματος στην εκπαίδευση / Βάρος m-οστού μοντέλου στο ensemble.
q_α	Το α -ποσοστημόριο της προγνωστικής κατανομής ($0 < \alpha < 1$).
$\left[q^{low}, q^{high}\right]$	Κάτω και άνω όριο του διαστήματος πρόβλεψης (*Prediction Interval*).
ADI	Μέσο Μεσοδιάστημα Ζήτησης (*Average Demand Interval*).
CV^2	Τετράγωνο του συντελεστή μεταβλητότητας των μη μηδενικών μεγεθών ζήτησης.
SS	Απόθεμα Ασφαλείας (Safety Stock) για την προστασία από την αβεβαιότητα.

Σύμβολο / Ακρωνύμιο	Ερμηνεία και μαθηματική φύση
(s, S)	Πολιτική ελέγχου αποθεμάτων: s (Σημείο αναπαραγγελίας), S (Μέγιστη στάθμη).
(Q, R) / (T, S)	Πολιτικές αποθεμάτων: (Ποσότητα, Σημείο αναπαραγγελίας) / (Περίοδος, Μέγιστη στάθμη).

2.2 Βασικές έννοιες χρονοσειρών

Η επιστημονική περιοχή της πρόβλεψης της ζήτησης (*demand forecasting*) συνιστά έναν από τους πιο κρίσιμους πυλώνες της σύγχρονης επιχειρησιακής έρευνας και της διοίκησης εφοδιαστικών αλυσίδων. Η ικανότητα μιας επιχείρησης, που δραστηριοποιείται στη δευτερογενή αγορά αυτοκινήτου (*automotive aftermarket*), να εκτιμά με ακρίβεια τις μελλοντικές ανάγκες της σε υλικά, επηρεάζει άμεσα τη χρηματοοικονομική της υγεία, τη λειτουργική της αποτελεσματικότητα και το επίπεδο εξυπηρέτησης των πελατών της [16].

Από στατιστικής άποψης, η διαδικασία αυτή βασίζεται στη συστηματική ανάλυση χρονοσειρών. Ο όρος χρονοσειρά (*time series*) ορίζεται αυστηρά ως μια διατεταγμένη ακολουθία παρατηρήσεων $\{y_1, y_2, \dots, y_T\}$, όπου κάθε στοιχείο αντιστοιχεί στην τιμή μιας τυχαίας μεταβλητής που καταγράφεται σε ένα συγκεκριμένο, διακριτό χρονικό βήμα [17]. Στο πλαίσιο της παρούσας εργασίας, τα χρονικά αυτά βήματα δύνανται να εκφράζουν ημερήσια, εβδομαδιαία ή μηνιαία διαστήματα συγκέντρωσης των πωλήσεων.

2.2.1 Χαρακτηριστικά και αποσύνθεση χρονοσειρών

Η υποκείμενη στοχαστική διαδικασία που γεννά μια χρονοσειρά μπορεί να προσεγγιστεί μαθηματικά μέσω της εννοιολογικής της αποσύνθεσης (*decomposition*). Η κλασική στατιστική θεωρία [16], [17] δέχεται ότι μια σειρά αποτελείται από τη σύνθεση τεσσάρων διακριτών, μη παρατηρήσιμων συνιστωσών:

- **Το Επίπεδο (Level – ℓ_t):** Αντιπροσωπεύει την τοπική μέση τιμή της χρονοσειράς, απογυμνωμένη από βραχυπρόθεσμες διακυμάνσεις και περιοδικά φαινόμενα.
- **Την Τάση (Trend – T_t):** Εκφράζει τη μακροχρόνια, μονότονη (αύξουσα ή φθίνουσα) κίνηση της σειράς σε διευρυμένους χρονικούς ορίζοντες, η οποία οφείλεται σε δομικές αλλαγές της αγοράς, όπως η σταδιακή αύξηση του μεγέθους του στόλου των οχημάτων.
- **Την Εποχικότητα (Seasonality – S_t):** Αφορά σε οικονομικά ή κλιματικά φαινόμενα που επαναλαμβάνονται με σταθερή και γνωστή περιοδικότητα (π.χ. αυξημένη ζήτηση για συγκεκριμένα φίλτρα ή ψυκτικά υγρά κατά τους θερινούς μήνες).
- **Τον Τυχαίο όρο (Irregular / Noise – ε_t):** Ενσωματώνει το μη συστηματικό, στοχαστικό υπόλοιπο της σειράς, το οποίο δεν μπορεί να προβλεφθεί και υποτίθεται ότι ακολουθεί, υπό κανονικές συνθήκες, μια διαδικασία λευκού θορύβου (*white noise*).

Ανάλογα με τον τρόπο με τον οποίο αλληλεπιδρούν οι παραπάνω συνιστώσες, διακρίνονται δύο κύριες μαθηματικές δομές αποσύνθεσης:

- **Προσθετικό μοντέλο (Additive model):**

$$y_t = \ell_t + T_t + S_t + \varepsilon_t \quad (2.1)$$

Το προσθετικό μοντέλο κρίνεται κατάλληλο, όταν το εύρος των εποχικών διακυμάνσεων ή η διακύμανση του τυχαίου όρου παραμένει σταθερή και ανεξάρτητη από την τρέχουσα μέση στάθμη (επίπεδο) της χρονοσειράς.

- **Πολλαπλασιαστικό μοντέλο (Multiplicative model):**

$$\ell_t \cdot T_t \cdot S_t \cdot \varepsilon_t \quad (2.2)$$

Το πολλαπλασιαστικό μοντέλο είναι επιχειρησιακά συχνότερο σε δεδομένα πωλήσεων, καθώς, το εύρος των εποχικών διακυμάνσεων τείνει να είναι ανάλογο του επιπέδου της σειράς (αν οι συνολικές πωλήσεις διπλασιαστούν, διπλασιάζεται και το απόλυτο μέγεθος της εποχικής έξαρσης).

Στην περίπτωση που η πολλαπλασιαστική συμπεριφορά προκαλεί έντονη ετεροσκεδαστικότητα, εφαρμόζεται ο λογαριθμικός μετασχηματισμός, ο οποίος μετατρέπει τη δομή σε προσθετική:

$$log(y_t) = \log(\ell_t) + log(T_t) + log(S_t) + log(\varepsilon_t) \quad (2.3)$$

Γενικότερη μορφή αποτελεί η οικογένεια μετασχηματισμών Box-Cox (Box & Cox, 1964), η οποία ορίζεται ως:

$$w_t = \begin{cases} \dfrac{y_t^{\lambda} - 1}{\lambda}, \lambda \neq 0 \\ \log(y_t), \lambda = 0 \end{cases} \quad (2.3)$$

Η παράμετρος λ σταθεροποιεί τη διακύμανση κατά μήκος της σειράς διευκολύνοντας τη γραμμική μοντελοποίηση. Σε περιβάλλοντα διαλείπουσας ζήτησης απαιτείται ιδιαίτερη προσοχή λόγω της ύπαρξης αυστηρά μηδενικών τιμών ($y_t = 0$), γεγονός που καθιστά αναγκαία την προσθήκη μιας σταθεράς μετατόπισης ($y_t + c$).

Για τη στατιστική διάγνωση των χρονικών εξαρτήσεων εντός της ίδιας της σειράς, χρησιμοποιείται η Συνάρτηση Αυτοσυσχέτισης (*Autocorrelation Function* - ACF). Για μια υστέρηση (lag) k, η ACF ορίζεται ως:

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{\mathbb{E}[(y_t - \mu)(y_{t-k} - \mu)]}{\mathbb{E}[(y_t - \mu)^2]} \quad (2.4)$$

Παράλληλα, η Μερική Συνάρτηση Αυτοσυσχέτισης (*Partial Autocorrelation Function* - PACF) μετρά τη γραμμική συσχέτιση μεταξύ των y_t και y_{t-k}, αφού έχει αφαιρεθεί η επίδραση των ενδιάμεσων χρονικών υστερήσεων $\{y_{t-1}, y_{t-2}, \dots, y_{t-k+1}\}$. Οι συναρτήσεις αυτές αποτελούν τα κύρια εργαλεία αναγνώρισης της τάξης των αυτοπαλίνδρομων διεργασιών και διεργασιών κινητού μέσου.

Η Έννοια της στασιμότητας

Μια χρονοσειρά χαρακτηρίζεται ως ασθενώς στάσιμη (*weakly stationary* / *covariance stationary*), όταν οι στατιστικές της ιδιότητες είναι ανεξάρτητες από τη μετατόπιση του χρόνου. Συγκεκριμένα, πρέπει να ικανοποιούνται ταυτόχρονα τρεις μαθηματικές συνθήκες:

- Η αναμενόμενη τιμή είναι σταθερή για κάθε χρονική στιγμή:

$$\mathbb{E}[y_t] = \mu, \forall t \quad (2.5)$$

- Η διακύμανση είναι πεπερασμένη και σταθερή:

$$\mathrm{Var}(y_t) = \sigma^2 < \infty, \forall t \tag{2.6}$$

- Η συνδιακύμανση μεταξύ δύο παρατηρήσεων εξαρτάται αποκλειστικά από τη χρονική τους απόσταση (υστέρηση) και όχι από τον απόλυτο χρόνο :

$$\mathrm{Cov}(y_t, y_{t-k}), \forall t, k \tag{2.7}$$

Η ύπαρξη μη στάσιμων συνιστωσών (όπως στοχαστικές τάσεις ή «τυχαίοι περίπατοι») ελέγχεται τυπικά μέσω δοκιμών μοναδιαίας ρίζας (unit root tests). Η πλέον διαδεδομένη είναι η *Augmented Dickey-Fuller* (ADF) (Dickey & Fuller, 1979), η οποία εξετάζει την υπόθεση ύπαρξης μοναδιαίας ρίζας μέσω της παλινδρόμησης:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \delta_i \, \Delta y_{t-i} + \varepsilon_t \tag{2.7}$$

Η μηδενική υπόθεση ($H_0: \gamma = 0$) συνεπάγεται μη στασιμότητα (ύπαρξη στοχαστικής τάσης), ενώ η εναλλακτική ($H_1: \gamma < 0$) υποδηλώνει στασιμότητα. Συμπληρωματικά λειτουργεί η δοκιμή KPSS [18], η οποία αναστρέφει τις υποθέσεις, θέτοντας ως μηδενική υπόθεση (H_0) ότι η σειρά είναι στάσιμη γύρω από ένα επίπεδο ή μια αιτιοκρατική τάση.

Σε περιβάλλοντα διαλείπουσας ζήτησης ανταλλακτικών, η έννοια της κλασικής στασιμότητας αναδεικνύει σοβαρές θεωρητικές και πρακτικές προκλήσεις. Λόγω της κυριαρχίας των μηδενικών παρατηρήσεων, οι κατανομές είναι εξαιρετικά ασύμμετρες και η έννοια της σταθερής διακύμανσης καταλύεται. Αν και τα κλασικά γραμμικά μοντέλα (π.χ. ARIMA) απαιτούν διαφοροποίηση για την επίτευξη στασιμότητας, οι εξειδικευμένες μέθοδοι διαλείπουσας ζήτησης (τύπου Croston) παρακάμπτουν αυτή τη δοκιμασία, καθότι δεν βασίζονται σε υποθέσεις γραμμικής στασιμότητας.

Ωστόσο, η σταθερότητα των υποκείμενων δομικών παραμέτρων της σειράς, όπως ο μέσος χρόνος μεταξύ δύο ζητήσεων (ADI) και η αναλογία των μηδενικών εμφανίσεων, παραμένει θεμελιώδης. Αυτό, διότι επιτρέπει τη σταθερή ταξινόμηση των προϊόντων σε διακριτές κατηγορίες, μια διαδικασία που αναλύεται στην ενότητα 2.3.2.

2.2.2 Σημειακή έναντι πιθανοκρατικής πρόβλεψης

Η ιστορική εξέλιξη της βιβλιογραφίας των χρονοσειρών επικεντρώθηκε κατά κύριο λόγο στην παραγωγή **σημειακών προβλέψεων** (*point forecasts*). Μια σημειακή πρόβλεψη εκφράζει μια μοναδική τιμή $\hat{y}_{t+h}$, η οποία αποτελεί την εκτίμηση του μοντέλου για τη μελλοντική τιμή της χρονοσειράς στον ορίζοντα h. Μαθηματικά, η σημειακή πρόβλεψη αντιπροσωπεύει συνήθως την υπό συνθήκη αναμενόμενη τιμή (*conditional expectation*) της τυχαίας μεταβλητής, δεδομένης της πληροφορίας $\mathcal{F}_t$ που είναι διαθέσιμη μέχρι τη χρονική στιγμή t:

$$\hat{y}_{t+h} = \mathbb{E}[y_{t+h} \mid \mathcal{F}_t] \tag{2.7}$$

Αν και οι σημειακές προβλέψεις είναι εύκολα κατανοητές, κρίνονται ως δομικά ανεπαρκείς για τη λήψη βέλτιστων αποφάσεων σε συστήματα ελέγχου αποθεμάτων [19]. Για παράδειγμα, η πληροφορία ότι η αναμενόμενη ζήτηση ενός κωδικού για τον επόμενο μήνα είναι 0,4 μονάδες δεν προσφέρει καμία ουσιαστική καθοδήγηση για το αν πρέπει να γίνει παραγγελία αναπλήρωσης και σε τι μέγεθος. Αν η ζήτηση εμφανίζει υψηλή διακύμανση, η πιθανότητα να ζητηθούν 2 ή 3 τεμάχια είναι υπαρκτή και ένα σύστημα βασισμένο αποκλειστικά στο 0,4 θα οδηγηθεί μαθηματικά σε έλλειψη αποθέματος (stock-out).

Για την υπέρβαση αυτού του περιορισμού, η σύγχρονη έρευνα επιτάσσει τη μετάβαση στο παράδειγμα της **πιθανοκρατικής πρόβλεψης** (*probabilistic forecasting*) [3], [19]. Στόχος της πιθανοκρατικής προσέγγισης είναι η πλήρης ποσοτικοποίηση της μελλοντικής αβεβαιότητας μέσω της εκτίμησης ολόκληρης της υπό συνθήκη συνάρτησης κατανομής της τυχαίας μεταβλητής:

$$F_{t+h}(y) = \mathbb{P}(y_{t+h} \leq y \mid \mathcal{F}_t) \tag{2.8}$$

Η γνώση της προγνωστικής κατανομής (predictive distribution) επιτρέπει τον ακριβή υπολογισμό συγκεκριμένων **ποσοστημορίων** (*quantiles*) q_α, τα οποία ορίζονται ως:

$$q_\alpha = \inf\{y \in \mathbb{R}: F_{t+h}(y) \geq \alpha\},\ \alpha \in (0,1) \tag{2.9}$$

Στη διαχείριση αποθεμάτων, τα υψηλά ποσοστημόρια (π.χ. $\alpha = 0{,}95$ ή $\alpha = 0{,}99$) συνδέονται άμεσα με το επιθυμητό επίπεδο εξυπηρέτησης (*cycle service level*), καθώς ορίζουν τη στάθμη του αποθέματος που επαρκεί για την κάλυψη της ζήτησης με πιθανότητα α.

Στο πλαίσιο της παρούσας διπλωματικής εργασίας, υιοθετείται ρητά η πιθανοκρατική προσέγγιση. Όλα τα αναπτυχθέντα υπολογιστικά μοντέλα (που παρουσιάζονται στο Κεφάλαιο 5) ξεπερνούν τον περιορισμό της σημειακής εκτίμησης και παράγουν ένα πλήρες διάνυσμα ποσοστημορίων, το οποίο τροφοδοτεί απευθείας τους αλγορίθμους βέλτιστης αναπλήρωσης.

2.2.3 Σφάλματα και μετρικές ακρίβειας

Η διαδικασία της αξιολόγησης (*evaluation*) των μοντέλων πρόβλεψης αποτελεί κρίσιμο στάδιο, καθώς καθορίζει ποιο μοντέλο θα επιλεγεί για επιχειρησιακή χρήση. Η αξιολόγηση βασίζεται στη στατιστική ανάλυση του σφάλματος πρόβλεψης, το οποίο για κάθε χρονική στιγμή t ορίζεται ως η διαφορά μεταξύ της πραγματικής (παρατηρημένης) τιμής y_t και της προβλεφθείσας τιμής $\hat{y}_t$:

$$e_t = y_t - \hat{y}_t \tag{2.10}$$

Στην κλασική βιβλιογραφία, οι πλέον διαδεδομένες μετρικές για τη σύνοψη των σφαλμάτων σε ένα δείγμα ελέγχου μεγέθους είναι οι εξής:

Μέσο Απόλυτο Σφάλμα (Mean Absolute Error – MAE):

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|e_t| \tag{2.11}$$

Το MAE μετρά το μέσο μέγεθος των σφαλμάτων χωρίς να λαμβάνει υπόψη την κατεύθυνσή τους. Από στατιστικής άποψης, η ελαχιστοποίηση του MAE οδηγεί το μοντέλο στο να προβλέψει την υπό συνθήκη διάμεσο (*median*) της κατανομής.

Ρίζα Μέσου Τετραγωνικού Σφάλματος (Root Mean Squared Error – RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}e_t^2} \tag{2.12}$$

Λόγω της ύψωσης του σφάλματος στο τετράγωνο, το RMSE επιβάλλει δυσανάλογα μεγαλύτερη ποινή στα μεγάλα σφάλματα. Η ελαχιστοποίησή του οδηγεί στην εκτίμηση της υπό συνθήκη μέσης τιμής (*mean*).

Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (Mean Absolute Percentage Error – MAPE):

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|e_t|}{|y_t|}\cdot 100\% \tag{2.13}$$

Παρά την ευρεία χρήση του MAPE στον επιχειρησιακό κόσμο λόγω της απλής του ερμηνείας ως ποσοστό, η μετρική αυτή, καθώς και η συμμετρική της παραλλαγή (sMAPE), κρίνεται απολύτως ακατάλληλη για τη διαλείπουσα ζήτηση [14], [20]. Όταν η πραγματική τιμή είναι μηδέν ($y_t = 0$), ο παρανομαστής του MAPE μηδενίζεται, καθιστώντας τον τύπο μαθηματικά απροσδιόριστο. Ακόμη και στην περίπτωση που εξαιρούνται οι μηδενικές παρατηρήσεις, το MAPE εισάγει μια συστηματική μεροληψία, καθώς ευνοεί μοντέλα που υποεκτιμούν τη ζήτηση και, ταυτόχρονα, τιμωρεί αυστηρότερα τις υπερεκτιμήσεις σε σχέση με τις υποεκτιμήσεις της ίδιας απόλυτης αξίας.

Κλιμακωμένες μετρικές (*Scaled errors*)

Για την υπέρβαση των μαθηματικών προβλημάτων που προκαλεί ο μηδενισμός των παρατηρήσεων, οι Hyndman & Koehler [20] εισήγαγαν την έννοια των κλιμακωμένων σφαλμάτων (*scaled errors*). Η κεντρική ιδέα συνίσταται στην κανονικοποίηση του σφάλματος του εξεταζόμενου μοντέλου ως προς το μέσο απόλυτο σφάλμα ενός απλού, ιστορικού μοντέλου αναφοράς (*naive baseline model*), υπολογισμένου εντός του δείγματος εκπαίδευσης (*in-sample*) ιστορικού μήκους T.

Η μετρική **Μέσο Απόλυτο Κλιμακούμενο Σφάλμα** (**Mean Absolute Scaled Error – MASE)** για μη εποχικά δεδομένα (όπου το μοντέλο αναφοράς είναι το naive της προηγούμενης περιόδου, ($\hat{y}_t = y_{t-1}$) ορίζεται ως:

$$\text{MASE} = \frac{\frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|}{\frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|} \tag{2.14}$$

Αντιστοίχως, για εποχικά δεδομένα με περίοδο m, ο παρανομαστής τροποποιείται με βάση το εποχιακό «αφελές» μοντέλο ($\hat{y}_t = y_{t-1}$):

$$\text{MASE} = \frac{\frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|}{\frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|} \tag{2.15}$$

Η ερμηνεία της MASE είναι άμεση και μαθηματικά στιβαρή. Δηλαδή, τιμή $MASE = 1$ υποδηλώνει ότι το εξεταζόμενο μοντέλο παράγει, κατά μέσο όρο, σφάλματα ισοδύναμα με εκείνα του απλού μοντέλου αναφοράς. Τιμές μικρότερες της μονάδας ($MASE < 1$) τεκμηριώνουν ότι το προτεινόμενο μοντέλο προσφέρει ανώτερη προγνωστική ακρίβεια. Δεδομένου ότι ο παρανομαστής βασίζεται στο ιστορικό των διαθέσιμων δεδομένων, παραμένει σταθερός και μη μηδενικός, εκτός αν η χρονοσειρά είναι απολύτως επίπεδη σε όλο το ιστορικό της. Οπότε η MASE παραμένει πάντοτε ορισμένη.

Η τετραγωνική παραλλαγή της MASE, η **Root Mean Squared Scaled Error (RMSSE)**, ορίζεται ως:

$$RMSSE = \sqrt{\frac{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\frac{1}{T-1}\sum_{t=2}^{T}(y_t - y_{t-1})^2}} \tag{2.16}$$

Η RMSSE αποτελεί τη μετρική που επιλέχθηκε ως το επίσημο κριτήριο αξιολόγησης στον παγκόσμιο διαγωνισμό πρόβλεψης M5 [21], εξαιτίας της εξαιρετικής της σταθερότητας κατά τη σύγκριση χρονοσειρών με εντελώς διαφορετικές κλίμακες μεγέθους και έντονη παρουσία μηδενικών.

Μια εναλλακτική προσέγγιση, η οποία χρησιμοποιείται ευρέως στην επιχειρησιακή πρακτική και στη διοικητική αναφορά μεγάλων αποθηκών, είναι το **Σταθμισμένο Απόλυτο Ποσοστιαίο Σφάλμα (Weighted Absolute Percentage Error – WAPE)**. Η WAPE παρακάμπτει το πρόβλημα της διαίρεσης με το μηδέν σε επίπεδο μεμονωμένης παρατήρησης, αθροίζοντας πρώτα τα απόλυτα σφάλματα για όλο το δείγμα και κανονικοποιώντας τα στη συνέχεια με τη συνολική πραγματική ζήτηση της ίδιας περιόδου:

$$\text{WAPE} = \frac{\sum_{t=1}^{n}|y_t - \hat{y}_t|}{\sum_{t=1}^{n}|y_t|} \tag{2.16}$$

Η WAPE είναι μαθηματικά ισοδύναμη με τη MAPE, κατά την οποία το σφάλμα κάθε περιόδου σταθμίζεται από το μέγεθος της πραγματικής ζήτησης της συγκεκριμένης περιόδου. Παραμένει σταθερά ορισμένη για κάθε σύνολο κωδικών, υπό την προϋπόθεση ότι η συνολική ζήτηση στο εξεταζόμενο παράθυρο είναι θετική ($\sum|y_t| > 0$). Από επιχειρησιακής άποψης, η WAPE παρουσιάζει μια φυσική σύνδεση με τα χρηματοοικονομικά μεγέθη μιας αποθήκης, καθώς το συνολικό σφάλμα αντιμετωπίζεται ως ποσοστό επί του συνολικού διακινούμενου όγκου [3], [15].

Επιχειρησιακή ευθυγράμμιση

Η επιλογή των μετρικών αξιολόγησης δεν αποτελεί μια ουδέτερη και καθαρά στατιστική απόφαση. Σύμφωνα με τη θεωρία των συναρτήσεων απώλειας (*loss functions*), διαφορετικές μετρικές επιβάλλουν διαφορετική μαθηματική συμπεριφορά και οδηγούν στη δημιουργία διαφορετικών βέλτιστων μοντέλων, ακόμη, και όταν εφαρμόζονται ακριβώς στα ίδια δεδομένα [14], [15].

Σε χρονοσειρές με διαλείπουσα ζήτηση, για παράδειγμα, η απευθείας ελαχιστοποίηση του MAE αναγκάζει συχνά τα μοντέλα να παράγουν μια σταθερά μηδενική πρόβλεψη ($\hat{y}_t = 0$), διότι το μηδέν αποτελεί τη διάμεσο της κατανομής, όταν οι μηδενικές παρατηρήσεις υπερβαίνουν το 50%. Αν και μια τέτοια πρόβλεψη εμφανίζει εξαιρετικό (χαμηλό) MAE, είναι επιχειρησιακά καταστροφική, καθώς θα οδηγούσε σε μηδενικές παραγγελίες και πλήρη παράλυση της εφοδιαστικής αλυσίδας.

Για την αποφυγή αυτών των στρεβλώσεων, στην παρούσα εργασία υιοθετείται μια στρατηγική **επιχειρησιακής ευθυγράμμισης (*operational alignment*)** μέσω της χρήσης ενός διευρυμένου *πακέτου μετρικών (evaluation suite)*. Το πακέτο αυτό περιλαμβάνει:

- Τη MASE και τη WAPE για την αποτίμηση της κεντρικής τάσης των σημειακών προβλέψεων.
- Τον Λόγο συνολικής Προβλεφθείσας προς Πραγματική ζήτηση (*Ratio of Actual/Predicted*) για τον έλεγχο της συστηματικής μεροληψίας (*bias*).
- Τη συνάρτηση «*Pinball Loss*» για τη μέτρηση της ακρίβειας των επιμέρους εκτιμώμενων ποσοστημορίων.
- Τις μετρικές κάλυψης και εύρους των Διαστημάτων Πρόβλεψης (*Prediction Interval coverage and width*) για την αξιολόγηση της ποιότητας της πιθανοκρατικής βαθμονόμησης.

Η πλήρης μαθηματική ανάλυση και η αιτιολόγηση της χρήσης αυτών των μετρικών αναπτύσσονται διεξοδικά στο Κεφάλαιο 5.

2.3 Διαχείριση Αποθεμάτων και Διαλείπουσα Ζήτηση Ανταλλακτικών

Η πρόβλεψη της μελλοντικής ζήτησης δεν αποτελεί έναν αυτοσκοπό, ούτε εξαντλείται στη βελτιστοποίηση μιας στατιστικής μετρικής σφάλματος. Αντιθέτως, εντάσσεται οργανικά σε ένα ευρύτερο και δυναμικό σύστημα αποφάσεων αναπλήρωσης. Ο απώτερος στόχος αυτού του συστήματος είναι η επίτευξη μιας βέλτιστης οικονομικής ισορροπίας μεταξύ της διαθεσιμότητας των προϊόντων, η

οποία μεταφράζεται σε υψηλό επίπεδο εξυπηρέτησης πελατών, και του κόστους διαχείρισης των αποθεμάτων [11], [22].

Στην ειδική περίπτωση των αποθεμάτων ανταλλακτικών της δευτερογενούς αγοράς αυτοκινήτου (*automotive aftermarket*), οι ακραίες ιδιαιτερότητες της ζήτησης, όπως η διαλείπουσα φύση της, η τεράστια ετερογένεια των κωδικών, το ασύμμετρα υψηλό κόστος έλλειψης και ο συνεχής κίνδυνος απαρχαίωσης, καθιστούν την κλασική θεωρία αποθεμάτων ανεπαρκή χωρίς θεμελιώδεις αναπροσαρμογές. Η παρούσα ενότητα αναλύει το σχετικό θεωρητικό πλαίσιο σε τρία διαδοχικά επίπεδα, τη μελέτη των δομικών χαρακτηριστικών των ανταλλακτικών, τη μαθηματική ποσοτικοποίηση της διαλείπουσας ζήτησης και τη συσχέτιση της πρόβλεψης με τις βέλτιστες αποθεματικές πολιτικές.

2.3.1 Ιδιαιτερότητες αποθεμάτων ανταλλακτικών

Η διεθνής βιβλιογραφία που προσεγγίζει τον κλάδο των ανταλλακτικών και της συντήρησης (service parts / spare parts inventory) συγκλίνει στην αναγνώριση τεσσάρων δομικών χαρακτηριστικών, τα οποία διαφοροποιούν ριζικά το πρόβλημα σε σχέση με το τυπικό λιανεμπόριο ταχέως κινούμενων αγαθών (*Fast Moving Consumer Goods* – FMCG) [23], [24], [25]:

- **Διαλείπουσα ή Ακανόνιστη ζήτηση (*Intermittent/Erratic demand*):** Σε αντίθεση με τα καταναλωτικά αγαθά, η ζήτηση για ανταλλακτικά δεν είναι συνεχής. Εμφανίζεται εξαιρετικά σποραδικά και σχεδόν αποκλειστικά στο πλαίσιο προληπτικής συντήρησης ή εμφάνισης βλάβης που απαιτεί άμεση επισκευή. Αυτό δημιουργεί χρονοσειρές με εκτεταμένα χρονικά διαστήματα μηδενικής δραστηριότητας, τα οποία διακόπτονται από απότομες και συχνά απρόβλεπτες κορυφές (*spikes*) στη ζήτηση.
- **Μαζική ποικιλία κωδικών (SKU *proliferation*):** Οι τυπικοί κατάλογοι των εταιρειών του κλάδου περιλαμβάνουν από δεκάδες έως εκατοντάδες χιλιάδες μοναδικούς κωδικούς (SKUs). Βάσει της ανάλυσης Pareto (ή κατάταξης ABC, βλ. ενότητα 4.3.1), η συντριπτική πλειοψηφία αυτών των κωδικών (συχνά πάνω από το 80%) ανήκει στη «μακριά ουρά» (*long-tail*) της χαμηλής κίνησης και δεσμεύει τεράστιο όγκο δεδομένων με ελάχιστες θετικές παρατηρήσεις ανά κωδικό. Αυτό, κατά συνέπεια, δυσχεραίνει την εκπαίδευση μεμονωμένων προγνωστικών μοντέλων.
- **Υψηλός κίνδυνος απαρχαίωσης (*Obsolescence risk*):** Η ραγδαία εξέλιξη της τεχνολογίας οχημάτων και η συνεχής αντικατάσταση των παλαιότερων μοντέλων με νέα καθιστούν ολόκληρες κατηγορίες ανταλλακτικών ξαφνικά απαξιωμένες [2]. Σε συνδυασμό με τη διαλείπουσα ζήτηση, ο κίνδυνος αυτός εξελίσσεται σε ιδιαίτερα κρίσιμο. γιατί ένα ανταλλακτικό που δεν έχει πουληθεί για 12 μήνες μπορεί απλώς να βιώνει τον φυσιολογικό (αραιό) κύκλο ζήτησής του ή να έχει καταστεί οριστικά "νεκρό απόθεμα" (*dead stock*). Η απουσία μοντέλων ικανών να διαχειριστούν ρητά την παύση της ζήτησης οδηγεί σε συσσώρευση άχρηστου κεφαλαίου στις αποθήκες.
- **Στενή σύνδεση με τη συντήρηση (*Maintenance dependency*):** Η ζήτηση σπάνια καθοδηγείται από την αυθόρμητη επιθυμία του τελικού καταναλωτή. Προέρχεται αυστηρά από τεχνικές αστοχίες υλικού (*reliability curves*) ή θεσμοθετημένα χρονοδιαγράμματα συντήρησης. Αυτό σημαίνει ότι τα δεδομένα κρύβουν λανθάνοντα μοτίβα (όπως ο μέσος χρόνος ζωής ενός εξαρτήματος) που, θεωρητικά, θα μπορούσαν να αξιοποιηθούν ως πηγή προγνωστικού σήματος από εξελιγμένα μοντέλα μηχανικής μάθησης.

Σε επιχειρησιακό επίπεδο, η σύγκρουση αυτών των παραγόντων δημιουργεί ένα δίκοπο μαχαίρι. Από τη μία πλευρά, υφίσταται ένα εξαιρετικά **υψηλό κόστος έλλειψης** (*shortage cost*). Δηλαδή, η αδυναμία άμεσης παροχής ενός ανταλλακτικού μεταφράζεται σε ακινητοποίηση του οχήματος (vehicle

downtime), οικονομική ζημία για τον επαγγελματία οδηγό, ακύρωση της εγγύησης ή οριστική απώλεια του πελάτη. Από την άλλη πλευρά, ελλοχεύει ένα υπέρογκο **κόστος διακράτησης** (*holding cost*), καθώς η διατήρηση ασφαλούς αποθέματος για εκατοντάδες χιλιάδες είδη με ελάχιστη κίνηση δεσμεύει καταχρηστικά το κεφάλαιο κίνησης της επιχείρησης.

Όπως διαπιστώνεται εμπειρικά και στα δεδομένα του ελληνικού δικτύου λιανικής που αναλύονται στο Κεφάλαιο 4, η πλειοψηφία του τζίρου παράγεται από ελάχιστα είδη τακτικής κίνησης (Κατηγορία A), ενώ ο κύριος όγκος του προβλήματος της πρόβλεψης εντοπίζεται στα είδη ελάχιστης κίνησης (Κατηγορίες B/C), εκεί όπου τα παραδοσιακά μοντέλα παρουσιάζουν τις μεγαλύτερες αστοχίες.

2.3.2 Διαλείπουσα και ακανόνιστη ζήτηση: ADI, CV² και Ταξινόμηση SBC

Για να αντιμετωπιστεί συστηματικά η ποικιλομορφία των χρονοσειρών, η βιβλιογραφία εισήγαγε μαθηματικά κριτήρια ποσοτικοποίησης της σποραδικότητας (*intermittency/sparsity*) και της μεταβλητότητας (*volatility*). Μια χρονοσειρά ζήτησης θεωρείται *διαλείπουσα*, όταν ένα σημαντικό ποσοστό των περιόδων καταγράφει μηδενικές πωλήσεις. Οι θεμέλιοι λίθοι για την ταξινόμηση αυτή είναι δύο κλασικοί δείκτες [5], [26]:

- **Μέσο Μεσοδιάστημα Ζήτησης (*Average Demand Interval* – ADI):** Εκφράζει τον μέσο αριθμό των χρονικών περιόδων που μεσολαβούν μεταξύ δύο διαδοχικών παρατηρήσεων με μη μηδενική ζήτηση. Υπολογίζεται ως ο λόγος του συνολικού μήκους της χρονοσειράς προς το πλήθος των περιόδων με θετική ζήτηση:

$$ADI = \frac{T}{\#\{t: y_t > 0\}} \tag{2.17}$$

  Μικρό ADI (κοντά στο 1) υποδηλώνει συνεχή (fast-moving) ζήτηση, ενώ μεγάλο ADI χαρακτηρίζει αργά κινούμενους κωδικούς (*slow-movers*).
- **Τετράγωνο Συντελεστή Μεταβλητότητας (*Coefficient of Variation Squared* – CV^2):** Μετρά τη σχετική διακύμανση του *μεγέθους* της ζήτησης. Είναι εξαιρετικά σημαντικό να τονιστεί ότι ο CV^2 υπολογίζεται **αποκλειστικά επί των μη μηδενικών παρατηρήσεων** ($y_t > 0$). Αν περιλαμβάνονταν τα μηδενικά, η διακύμανση θα διογκωνόταν τεχνητά από την ύπαρξη των κενών διαστημάτων συγχέοντας τη συχνότητα με το μέγεθος. Ορίζεται ως:

$$CV^2 = \left(\frac{\sigma_{y>0}}{\mu_{y>0}}\right)^2 \tag{2.18}$$

  όπου $\mu_{y>0}$ και $\sigma_{y>0}$ είναι η μέση τιμή και η τυπική απόκλιση των θετικών μεγεθών ζήτησης, αντίστοιχα.

Ο συνδυασμός αυτών των δύο δεικτών οδήγησε στη δημιουργία του πλέον καθιερωμένου εργαλείου ταξινόμησης στην επιστήμη των αποθεμάτων, τον **Πίνακα Ταξινόμησης Syntetos–Boylan–Croston (SBC *matrix*)** [5]. Ο πίνακας αυτός διαμερίζει το σύνολο των κωδικών μιας αποθήκης σε τέσσερα τεταρτημόρια ορίζοντας θεωρητικά τεκμηριωμένα κατώφλια: $ADI = 1{,}32$ και $CV^2 = 0{,}49$. Οι διακριτές κατηγορίες ταξινόμησης είναι **Ομαλή (*Smooth*)**, **Ακανόνιστη (*Erratic*)**, **Διαλείπουσα (*Intermittent*)** και **Ανομοιόμορφη (*Lumpy*)**.

Τα συγκεκριμένα κατώφλια δεν επελέγησαν αυθαίρετα, αλλά αποτελούν τα μαθηματικά σημεία τομής, στα οποία παύει να είναι αποδοτική η απλή Εκθετική Εξομάλυνση και απαιτείται η χρήση εξειδικευμένων μοντέλων (βλ. ενότητα 2.5). Η αναλυτική παρουσίαση του πίνακα SBC, καθώς, και η

εκτενής εφαρμογή του για την κατηγοριοποίηση των πραγματικών δεδομένων της παρούσας διπλωματικής εργασίας παρατίθενται στην ενότητα 4.6.1.

Σε εκτεταμένες εμπειρικές μελέτες σε δίκτυα ανταλλακτικών, αποδεικνύεται σταθερά ότι η συντριπτική πλειοψηφία των SKUs κατατάσσεται στις κατηγορίες *Διαλείπουσα* και *Ανομοιόμορφη* [24], [25], [27], γεγονός που επιτάσσει την εγκατάλειψη των παραδοσιακών γραμμικών μοντέλων προς όφελος σύγχρονων προσεγγίσεων μηχανικής μάθησης.

2.3.3 Αποθεματικές πολιτικές και σύνδεση με την πρόβλεψη

Σε ένα στοχαστικό επιχειρησιακό περιβάλλον, όπου η μελλοντική ζήτηση δεν είναι ντετερμινιστικά γνωστή, η διαχείριση επιτυγχάνεται μέσω του καθορισμού μιας αποθεματικής πολιτικής (inventory policy). Η πολιτική αυτή συνιστά ένα σύνολο μαθηματικών κανόνων που απαντούν σε δύο ερωτήματα: «Πότε πρέπει να παραγγείλω;» και «Πόσο *πρέπει να παραγγείλω;»*. Οι τρεις επικρατέστερες οικογένειες πολιτικών είναι οι κάτωθι [11], [22]:

- **Συνεχής έλεγχος ($\boldsymbol{Q, R}$):** Το σύστημα παρακολουθεί το απόθεμα διαρκώς. Μόλις το διαθέσιμο απόθεμα μειωθεί στο Σημείο Αναπαραγγελίας (*Reorder Point* - R), εκδίδεται άμεσα μια παραγγελία σταθερής ποσότητας (η οποία συχνά ισούται με την Οικονομική Ποσότητα Παραγγελίας - EOQ).
- **Περιοδικός έλεγχος ($\boldsymbol{T, S}$):** Η επιθεώρηση γίνεται σε αυστηρά, προκαθορισμένα χρονικά διαστήματα T (π.χ. κάθε Παρασκευή). Ανάλογα με το απόθεμα που βρίσκεται εκείνη τη στιγμή στο ράφι, εκδίδεται μια μεταβλητή παραγγελία, τόση ώστε το συνολικό απόθεμα να φτάσει σε ένα επιθυμητό Μέγιστο Επίπεδο S.
- **Πολιτική Ελάχιστου – Μέγιστου ($\boldsymbol{s, S}$):** Πρόκειται για έναν ισχυρό συνδυασμό. Το σύστημα ελέγχεται συνεχώς (ή περιοδικά), αλλά παραγγελία εκδίδεται *μόνο* αν το απόθεμα πέσει κάτω από το κατώφλι s (σημείο αναπαραγγελίας). Η ποσότητα της παραγγελίας είναι μεταβλητή και υπολογίζεται, ώστε το νέο απόθεμα να αγγίξει το άνω όριο S.

Η πολιτική (s, S) έχει αποδειχθεί μαθηματικά ως η πλέον **βέλτιστη επιλογή** (*optimal policy*) υπό ένα ευρύ φάσμα συνθηκών που περιλαμβάνουν πάγια κόστη παραγγελίας, αναλογικά κόστη αγοράς, κόστη διακράτησης και ποινές έλλειψης [28]. Επιπλέον, αποτελεί το απόλυτο βιομηχανικό πρότυπο για είδη με διαλείπουσα ζήτηση, καθώς η απαίτηση να ξεπεραστεί το «εμπόδιο» του s αποτρέπει το σύστημα από το να εκδίδει συνεχείς, μικροκαμωμένες και ακριβές παραγγελίες για κάθε μεμονωμένο τεμάχιο που πωλείται.

Η πρόκληση της ζήτησης κατά τον χρόνο προμήθειας (D_L)

Σε όλες ανεξαιρέτως τις παραπάνω πολιτικές, ο ακρογωνιαίος λίθος της επιτυχίας δεν είναι η πρόβλεψη της αυριανής ζήτησης, αλλά η εκτίμηση της **κατανομής της ζήτησης κατά τον χρόνο προμήθειας** (D_L), όπου L είναι το *lead time* (ο χρόνος από τη στιγμή έκδοσης της παραγγελίας έως την παραλαβή της στην αποθήκη).

Στην κλασική, παραμετρική θεωρία, εάν υποτεθεί ότι η ζήτηση ακολουθεί την κανονική κατανομή, το Απόθεμα Ασφαλείας (*Safety Stock* - SS) και το Σημείο Αναπαραγγελίας (ROP) υπολογίζονται ως εξής:

$$SS = z_\alpha \cdot \sigma_{D_L} \tag{2.19}$$

όπου το σ_{D_L} αντιπροσωπεύει την τυπική απόκλιση της ζήτησης κατά τον χρόνο παράδοσης και $z_\alpha = \Phi^{-1}(\alpha)$ είναι το ποσοστημόριο της τυποποιημένης κανονικής κατανομής που αντιστοιχεί στο επιθυμητό επίπεδο εξυπηρέτησης α (π.χ. για $\alpha = 0{,}95, z = 1{,}645$).

$$ROP = \mu_{D_L} + SS = \mu_{D_L} + z_\alpha \cdot \sigma_{D_L} \quad (2.20)$$

όπου μ_{D_L} είναι η μέση ζήτηση κατά τον χρόνο παράδοσης.

Ωστόσο, αυτή η παραμετρική προσέγγιση καταρρέει πλήρως σε συνθήκες διαλείπουσας ζήτησης. Η κατανομή των πωλήσεων στα ανταλλακτικά δεν είναι κανονική αλλά χαρακτηρίζεται από έντονη δεξιά ασυμμετρία (right-skewed) και ακραία συγκέντρωση τιμών στο μηδέν (*zero-inflation*). Η χρήση του z_α σε τέτοια δεδομένα οδηγεί μαθηματικά σε υπερεκτίμηση των αναγκών και διόγκωση του νεκρού αποθέματος.

Για την υπέρβαση αυτού του κρίσιμου θεωρητικού αδιεξόδου, η σύγχρονη βιομηχανική πρακτική εγκαταλείπει τις παραδοχές της κανονικότητας. Αντ' αυτού, μεταβαίνει στην άμεση εκτίμηση των ποσοστημορίων της ζήτησης (*quantile forecasting*). Μέσω τεχνικών μηχανικής μάθησης (βλ. ενότητα 2.10.2), το μοντέλο μαθαίνει να προβλέπει απευθείας το σημείο αναπαραγγελίας χωρίς να παρεμβάλλεται καμία κατανομή:

$$\text{ROP} = \hat{q}_\alpha(D_L \mid \mathbf{x}) \quad (2.21)$$

όπου το επίπεδο α μπορεί να καθοριστεί δυναμικά μέσω της σχέσης κόστους έλλειψης προς συνολικό κόστος, όπως ακριβώς υπαγορεύει το πλαίσιο του εφημεριδοπώλη (*newsvendor problem*) [3], [29].

Αυτή ακριβώς η βασισμένη στα δεδομένα (*data-driven*) φιλοσοφία ακολουθείται στην παρούσα ΔΕ. Στο Κεφάλαιο 5 εξάγονται μη-παραμετρικές πιθανοκρατικές προβλέψεις και τα υψηλά ποσοστημόριά τους (π.χ. P90, P95) μεταφράζονται απευθείας στα κατώφλια (s, S) του Κεφαλαίου 6.

Συστηματικός κλιμακωτός σχεδιασμός σε εκτενείς καταλόγους ειδών

Παρά τη θεωρητική υπεροχή των πολιτικών (s, S), ο ακριβής υπολογισμός των παραμέτρων s και S για κάθε κωδικό ξεχωριστά συνιστά ένα NP-*Hard* πρόβλημα. Αυτό καθιστά την πλήρη εφαρμογή του απαγορευτική για καταλόγους με δεκάδες χιλιάδες είδη. Στην πράξη, το πρόβλημα αυτό αντιμετωπίζεται με τη χρήση ισχυρών ευρετικών (*heuristics*) μεθόδων διαστασιολόγησης, όπως η προσέγγιση των Babai, Syntetos και Teunter [30], οι οποίες αξιοποιούν τις προβλεπτικές εξόδους των μοντέλων μηχανικής μάθησης για τον ταχύ και αυτοματοποιημένο καθορισμό των παραμέτρων αναπλήρωσης. Η πλήρης περιγραφή της επιχειρησιακής εφαρμογής της μεθόδου παρουσιάζεται αναλυτικά στο Κεφάλαιο 6.

2.4 Λογοκριμένη (περικομμένη) ζήτηση και ελλείψεις αποθέματος

Μια θεμελιώδης δυσκολία στην εκτίμηση των κατανομών ζήτησης από ιστορικά δεδομένα έγκειται στο γεγονός ότι οι καταγεγραμμένες πωλήσεις δεν αποτελούν πάντοτε πιστή αναπαράσταση της πραγματικής πρόθεσης αγοράς. Όταν ένα προϊόν βρίσκεται σε κατάσταση έλλειψης αποθέματος (stock-out), οι πελάτες που θα επιθυμούσαν να αγοράσουν περαιτέρω μονάδες δεν εμφανίζονται στα δεδομένα των συναλλαγών και, άρα, η ζήτησή τους είτε χάνεται οριστικά είτε ικανοποιείται από ανταγωνιστές.

Από στατιστικής σκοπιάς, το φαινόμενο αυτό συνιστά **δεξιά λογοκρισία (*right-censoring*)** ή περικοπή της πραγματικής τυχαίας μεταβλητής της ζήτησης [31], [32]. Το πρόβλημα καθίσταται ιδιαίτερα οξύ στη διαχείριση αποθεμάτων ανταλλακτικών αυτοκινήτου, όπου οι ελλείψεις είναι συχνές λόγω της διαλείπουσας φύσης και οι τελικοί πελάτες (συνεργεία ή οδηγοί) σπάνια μπαίνουν στη διαδικασία να καταγράψουν την ανικανοποίητη ζήτησή τους. Η παρούσα ενότητα τυποποιεί μαθηματικά το πρόβλημα, αναλύει τις σοβαρές στατιστικές του συνέπειες και επισκοπεί τις κύριες μεθόδους της βιβλιογραφίας για την αντιμετώπισή του.

2.4.1 Παρατηρούμενες πωλήσεις έναντι λανθάνουσας ζήτησης

Για την αυστηρή περιγραφή του προβλήματος, έστω, $d_t \geq 0$, η *λανθάνουσα* (πραγματική) ζήτηση κατά την περίοδο t, $I_t \geq 0$, το διαθέσιμο φυσικό απόθεμα στην αρχή της περιόδου και, $y_t \geq 0$, οι παρατηρούμενες πωλήσεις που καταγράφονται στο πληροφοριακό σύστημα. Υπό τη σύμβαση της οριστικής απώλειας πωλήσεων (*lost sales*), δηλαδή υποθέτοντας ότι δεν υπάρχει δυνατότητα άμεσης έκτακτης αναπλήρωσης εντός της ίδιας περιόδου, ισχύει η εξής θεμελιώδης μη γραμμική σχέση:

$$y_t = \min(d_t, I_t) \tag{2.22}$$

Για τον διαχωρισμό των παρατηρήσεων, ορίζεται η κάτωθι δυαδική ενδεικτική συνάρτηση λογοκρισίας:

$$c_t = \mathbb{1}_{(y_t = I_t)} = \begin{cases} 1, \textit{ η περίοδος t είναι λογοκριμένη (stock-out)} \\ 0, \textit{ η περίοδος t είναι μη λογοκριμένη} \end{cases} \tag{2.23}$$

Η κρίσιμη παρατήρηση σε αυτό το πλαίσιο είναι η εξής, όταν $c_t = 0$ (επαρκές απόθεμα), η μεταβλητή των πωλήσεων ταυτίζεται απόλυτα με την d_t και αποκαλύπτεται η πραγματική ζήτηση. Όταν, όμως, $c_t = 1$ (εξάντληση αποθέματος), η μόνη διαθέσιμη πληροφορία είναι η ανισότητα $d_t \geq I_t$. Η ακριβής τιμή της d_t παραμένει άγνωστη. Επομένως, για κάθε λογοκριμένη παρατήρηση δημιουργείται ένα «τυφλό σημείο» στην εμπειρική κατανομή, το οποίο εντοπίζεται συστηματικά στη δεξιά (άνω) ουρά της κατανομής της ζήτησης [31], [33].

Σε επιχειρησιακό επίπεδο, η διάκριση μεταξύ απώλειας πωλήσεων (*lost sales*) και εκκρεμών παραγγελιών (*backordering*) είναι καθοριστική. Στην πρώτη περίπτωση, η οποία αποτελεί τη ρεαλιστική συνθήκη για τις πωλήσεις λιανικής της δευτερογενούς αγοράς, η ανικανοποίητη ζήτηση χάνεται οριστικά. Αντιθέτως, κατά την υποβολή εκκρεμών παραγγελιών, η ζήτηση διατηρείται στο σύστημα και ικανοποιείται ετεροχρονισμένα εμφανιζόμενη ως πώληση σε μελλοντική περίοδο (προκαλώντας, ωστόσο, στρεβλώσεις στον χρονισμό της ζήτησης). Στην παρούσα διπλωματική εργασία υιοθετείται η αυστηρή σύμβαση των χαμένων πωλήσεων. Δηλαδή, τα δεδομένα συναλλαγών αντανακλούν την τελική πράξη της πώλησης και όχι την αρχική πρόθεση του πελάτη.

2.4.2 Μηχανισμοί λογοκρισίας

Στα πραγματικά βιομηχανικά δεδομένα, η λογοκρισία εκδηλώνεται μέσω διαφόρων μηχανισμών, πέραν της απλής συνθήκης $y_t = I_t$ ως ακολούθως:

- **Πλήρης έλλειψη αποθέματος ($I_t = 0$):** Αποτελεί την πιο σαφή μορφή. Καμία πώληση δεν είναι τεχνικά εφικτή, με αποτέλεσμα $y_t = 0$. Το φαινόμενο αυτό συχνά συγχέεται από τα στατιστικά μοντέλα με τη γνήσια απουσία ζήτησης (*genuine zero*) στο πλαίσιο της διαλείπουσας ζήτησης.
- **Μερική λογοκρισία ($0 < I_t < d_t$):** Η πώληση πραγματοποιείται μέχρι την εξάντληση του διαθέσιμου ορίου και η υπερβάλλουσα ποσότητα που θα επιθυμούσε ο πελάτης χάνεται σιωπηλά.
- **Λογοκρισία ροής εργασιών (Workflow censoring):** Σε ορισμένα περιβάλλοντα λογιστικών εφαρμογών (ERP), οι παραγγελίες πελατών καταγράφονται ως πωλήσεις μόνο κατά τη στιγμή της τιμολόγησης ή της αποστολής. Αν μια παραγγελία ακυρωθεί λόγω έλλειψης πριν τιμολογηθεί, δεν αφήνει κανένα ίχνος στο σύστημα.

Ένα μείζον πρακτικό εμπόδιο είναι ότι τα ιστορικά αρχεία διατηρούν τυπικά μόνο τη ροή των πωλήσεων (y_t) και όχι τη χρονοσειρά του φυσικού αποθέματος (I_t). Συνεπώς, η ένδειξη δεν είναι άμεσα παρατηρήσιμη. Σε τέτοιες περιπτώσεις επιστρατεύονται τεχνικές ανίχνευσης (*proxies*), όπως η έμμεση

ανακατασκευή του υπολοίπου ή η ανίχνευση λογοκρισίας από τα ίδια τα μοτίβα πωλήσεων [34]. Οι Jain, Rudi και Wang [7] απέδειξαν ότι η ακριβής χρονική στιγμή που συμβαίνει η έλλειψη αποθέματος (*stock-out*) αποτελεί το κλειδί για την κατανόηση της πραγματικής ζήτησης. Πρακτικά, η παρατήρηση, αν το απόθεμα εξαντληθεί στην αρχή ή στο τέλος της εξεταζόμενης περιόδου, αποκαλύπτει την πραγματική δυναμική της αγοράς. Βασιζόμενοι σε αυτό, πρότειναν μεθόδους εκτίμησης που αξιοποιούν απλώς την ταχύτητα (ή τον ρυθμό), με την οποία αδειάζει το ράφι.

2.4.3 Στατιστικές συνέπειες της λογοκρισίας

Η απουσία διόρθωσης της λογοκρισίας και η απευθείας εκπαίδευση προγνωστικών μοντέλων πάνω στη μεταβλητή οδηγεί σε μαθηματικά προβλέψιμες και εξαιρετικά δυσμενείς αλλοιώσεις μεροληψίας (*bias*). Δεδομένου ότι η σχέση $y_t \leq d_t$ ισχύει πάντοτε, και $y_t < d_t$ ισχύει αυστηρά όταν $c_t = 1$, προκύπτουν οι εξής ανισότητες για τις ροπές των κατανομών [31], [32]:

$$\mathbb{E}[y_t] \leq \mathbb{E}[d_t] \text{ και } Var(y_t) \leq Var(d_t) \tag{2.24}$$

Δηλαδή, η μέση τιμή των πωλήσεων υποεκτιμά συστηματικά την πραγματική μέση ζήτηση, και η εμπειρική διακύμανση των πωλήσεων είναι τεχνητά συρρικνωμένη λόγω της αποκοπής της άνω ουράς. Και τα δύο αυτά φαινόμενα είναι καταστροφικά για τον έλεγχο αποθεμάτων:

- Οι χαμηλότερες κεντρικές εκτιμήσεις οδηγούν σε συστηματικές υπο-παραγγελίες.
- Η συρρικνωμένη διακύμανση οδηγεί σε επικίνδυνη υπο-εκτίμηση του μεγέθους του αποθέματος ασφαλείας (*safety stock*) αφήνοντας το σύστημα εκτεθειμένο.

Η αλληλεπίδραση των δύο αυτών σφαλμάτων πυροδοτεί έναν επιχειρησιακό φαύλο κύκλο (*spiral of death*):

έλλειψη αποθέματος → λογοκριμένα δεδομένα → υποεκτίμηση προβλέψεων → μικρότερες παραγγελίες → νέες, συχνότερες ελλείψεις

Το φαινόμενο αυτό έχει ποσοτικοποιηθεί αναλυτικά από τους Lau και Lau [33]. Επιπροσθέτως, η λογοκρισία αλλοιώνει ριζικά την εκτίμηση των ποσοστημορίων. Δηλαδή, όλα τα εμπειρικά ποσοστημόρια q_a για επίπεδα $a > 1 - \bar{c}$, όπου το $\bar{c}$ ποσοστό των λογοκριμένων παρατηρήσεων, καθίστανται μη ταυτοποιήσιμα χωρίς την προσθήκη παραμετρικών υποθέσεων. Αυτό είναι κρίσιμο για τη διαχείριση ανταλλακτικών, καθώς τα υψηλά ποσοστημόρια που απαιτούνται για επίπεδα εξυπηρέτησης 90%–99% είναι ακριβώς εκείνα που υφίστανται την εντονότερη παραμόρφωση.

2.4.4 Μεθοδολογικές προσεγγίσεις αντιμετώπισης της λογοκρισίας ζήτησης (*Demand censoring*)

Ένα από τα πλέον δυσεπίλυτα προβλήματα στη διαχείριση εφοδιαστικής αλυσίδας είναι η διάκριση μεταξύ των καταγεγραμμένων πωλήσεων (*sales*) και της πραγματικής, υποκείμενης ζήτησης (*demand*). Σε περιόδους, κατά τις οποίες ένα προϊόν παρουσιάζει έλλειψη αποθέματος, οι πωλήσεις αναγκαστικά μηδενίζονται (ή περιορίζονται στο διαθέσιμο υπόλοιπο), παρότι η πραγματική ζήτηση της αγοράς παραμένει θετική. Αυτό το φαινόμενο παράγει «λογοκριμένα» (*censored*) δεδομένα. Η εκπαίδευση αλγορίθμων απευθείας σε αυτά, χωρίς καμία διόρθωση, εισάγει μια ισχυρή μεροληψία υπο-εκτίμησης (*under-forecasting bias*), καθώς το μοντέλο εκλαμβάνει την αδυναμία εξυπηρέτησης ως έλλειψη αγοραστικού ενδιαφέροντος.

Για την αποκατάσταση των περιορισμών της ζήτησης (*unconstraining demand*), η επιστήμη των δεδομένων και η επιχειρησιακή έρευνα έχουν αναπτύξει τέσσερις βασικές σχολές σκέψης:

Παραγοντοποιημένη πιθανοφάνεια (censored likelihood) και αλγόριθμος EM

Η πρώτη, αμιγώς στατιστική προσέγγιση, στοχεύει στην ανακάλυψη της πραγματικής μαθηματικής κατανομής της ζήτησης. Βασίζεται στον αλγόριθμο **EM (*Expectation-Maximization*)**, ο οποίος λειτουργεί επαναληπτικά σε δύο βήματα:

- **Βήμα Ε (Προσδοκία):** Εκτιμά (μαντεύει) την πραγματική ζήτηση κατά τις ημέρες έλλειψης βασιζόμενος στο υπόλοιπο ιστορικό. (π.χ. «Εφόσον το απόθεμα των 10 τεμαχίων εξαντλήθηκε, η πραγματική ζήτηση πιθανώς να ήταν 13»).
- **Βήμα Μ (Μεγιστοποίηση):** Προσαρμόζει τις παραμέτρους της θεωρητικής κατανομής της ζήτησης ενσωματώνοντας αυτές τις εκτιμώμενες τιμές. Η διαδικασία επαναλαμβάνεται μέχρι να υπάρξει σύγκλιση. Παρότι η μέθοδος είναι μαθηματικά άρτια, προϋποθέτει ότι η μορφή της κατανομής (π.χ. Κανονική, Poisson) είναι γνωστή εκ των προτέρων. Αυτή, όμως, είναι μια υπόθεση που καταρρίπτεται συστηματικά στα περιβάλλοντα αραιής και διαλείπουσας ζήτησης.

Επαναξιολόγηση και καταλογισμός δεδομένων (*data imputation*)

Πρόκειται για μια πρακτική προσέγγιση καθαρισμού δεδομένων (*data engineering*). Αντί να τροποποιηθεί το μοντέλο πρόβλεψης, παρεμβαίνουμε στο ίδιο το σύνολο δεδομένων. Οι ημέρες που καταγράφηκε έλλειψη αποθέματος εντοπίζονται, οι καταγεγραμμένες πωλήσεις διαγράφονται και στη θέση τους «καταλογίζεται» (*imputed*) μια νέα τιμή. Η τιμή αυτή υπολογίζεται συνήθως ως η αναμενόμενη τιμή της ζήτησης υπό τη συνθήκη ότι αυτή υπερβαίνει το (μηδενικό) διαθέσιμο απόθεμα. Αν και υπολογιστικά γρήγορη, η μέθοδος αυτή εισάγει τεχνητά δεδομένα, αλλοιώνοντας συχνά τη φυσική διακύμανση της χρονοσειράς.

Δεσμευμένα μοντέλα με εξωγενή χαρακτηριστικά (*indicator features*)

Σε αυτή τη σύγχρονη προσέγγιση Μηχανικής Μάθησης (Machine Learning), τα πρωτογενή δεδομένα πωλήσεων παραμένουν αναλλοίωτα, αποφεύγοντας τον κίνδυνο τεχνητής παραποίησης. Αντίθετα, εισάγεται στο διάνυσμα χαρακτηριστικών (*feature vector*) μια νέα δυαδική μεταβλητή (OOS *flag*). Όπως αναφέρθηκε νωρίτερα, ισχύει $c_t = 1$, όταν παρατηρείται έλλειψη αποθέματος. Αλγόριθμοι αιχμής, όπως τα Δέντρα Διαβαθμισμένης Ενίσχυσης (*Gradient Boosting Trees*), έχουν την ικανότητα να εντοπίσουν αυτόματα το μοτίβο. Δηλαδή, μαθαίνουν ότι, όταν ισχύει $c_t = 1$, η τιμή-στόχος είναι τεχνητά συμπιεσμένη προς τα κάτω και προσαρμόζουν αυτόνομα τις μελλοντικές τους προβλέψεις προς τα πάνω. Αυτή η μέθοδος προσφέρει εξαιρετική στιβαρότητα (*robustness*) σε μη γραμμικά περιβάλλοντα.

Αναλυτική βάσει ποσοστημορίων (censored newsvendor)

Πρόκειται για μια κομψή εναλλακτική που προέρχεται από την προδιαγραφική αναλυτική (*prescriptive analytics*). Η μέθοδος βασίζεται στο γεγονός ότι, για τον καθορισμό του αποθέματος ασφαλείας, μας ενδιαφέρει πρωτίστως το άνω άκρο της κατανομής της ζήτησης (π.χ. το 95ο ποσοστημόριο για 95% επίπεδο εξυπηρέτησης). Εάν, ιστορικά, οι ημέρες με έλλειψη αποθέματος αντιπροσωπεύουν ένα μικρό ποσοστό (π.χ. 3%), τότε η «λογοκρισία» των δεδομένων αφορά μόνο το ανώτατο 3% της κατανομής. Δεδομένου ότι το 95ο ποσοστημόριο βρίσκεται χαμηλότερα από αυτό το όριο (στο «υγιές» τμήμα των δεδομένων), η μέθοδος προτείνει την πλήρη αφαίρεση των ημερών έλλειψης από το σύνολο εκπαίδευσης. Έτσι, το μοντέλο μαθαίνει τα σωστά ποσοστημόρια χωρίς πολύπλοκους καταλογισμούς ή στατιστικές παραδοχές.

2.4.5 Λογοκρισία ζήτησης στο πλαίσιο της παρούσας εργασίας

Η αποτελεσματική διαχείριση της λογοκριμένης ζήτησης παραμένει ένα ανοιχτό ερευνητικό ζήτημα, ειδικά σε βιομηχανικά περιβάλλοντα μεγάλης κλίμακας (όπως η δευτερογενής αγορά αυτοκινήτου), όπου η πλήρης ανακατασκευή της καμπύλης ζήτησης (Μέθοδος 1) είναι πρακτικά ανέφικτη. Στο πλαίσιο της παρούσας διπλωματικής, το πρόβλημα δεν αντιμετωπίζεται απλώς ως ένα τεχνικό ζήτημα καθαρισμού δεδομένων, αλλά ως μια δομική παράμετρος που διαμορφώνει ολόκληρη την αρχιτεκτονική του συστήματος.

Στην τελική αρχιτεκτονική του προτεινόμενου συστήματος (`Sparse True Forecast Pipeline`), υιοθετείται μια υβριδική στρατηγική που αντλεί στοιχεία από τις σύγχρονες προσεγγίσεις της Μηχανικής Μάθησης και διατρέχει όλα τα στάδια της ανάλυσης:

- **Στο σύνολο δεδομένων (Κεφάλαιο 4):** Λόγω της πρόσβασης τόσο στις ιστορικές πωλήσεις, όσο και στα ημερήσια υπόλοιπα αποθέματος, υπολογίζεται με ακρίβεια η ενδεικτική μεταβλητή λογοκρισίας (*Out-of-Stock indicator* – `OOS`) για κάθε συνδυασμό είδους και ημέρας.
- **Στη μεθοδολογία πρόβλεψης (Κεφάλαιο 5):** Αντί για την πλήρη διαγραφή των δεδομένων (Μέθοδος 4) ή τον αυθαίρετο καταλογισμό τους (Μέθοδος 2), εφαρμόζονται δύο συνδυαστικές τεχνικές. Πρώτον, η μεταβλητή `OOS` αξιοποιείται μέσω ρητής εισαγωγής της εντός του διανύσματος χαρακτηριστικών (Μέθοδος 3) επιτρέποντας, έτσι, στα μοντέλα Εμποδίου δύο σταδίων (*two-stage* Hurdle *models*) να λαμβάνουν τη δομική πληροφορία της διαθεσιμότητας. Δεύτερον, εφαρμόζεται μια προηγμένη τεχνική «στάθμισης περικοπής» (*demand censoring weighting*). Στα δείγματα που τεκμηριωμένα παρουσιάζουν έλλειψη αποθέματος, αποδίδεται μειωμένος συντελεστής βαρύτητας κατά την εκπαίδευση ($w_{oos} = 0.80$).
- **Στην αξιολόγηση και εφαρμογή (Κεφάλαιο 6):** Κατά την προσομοίωση των πολιτικών αναπλήρωσης, οι ημέρες εξάντλησης αποθέματος απομονώνονται από τον υπολογισμό συγκεκριμένων εμπειρικών μετρικών (π.χ. *Fill Rate*), ώστε να αποτραπεί η παραπλανητική επιβράβευση υπο-διαστασιολογημένων συστημάτων.

Αυτή η συνδυαστική προσέγγιση ενισχύει κατακόρυφα τη στιβαρότητα του συστήματος, γιατί, αφενός διατηρεί το πολύτιμο στατιστικό σήμα ότι πρόκειται για κωδικούς χαμηλής κυκλοφορίας, και αφετέρου, μειώνει δραστικά την επιζήμια μεροληψία υπο-εκτίμησης που προκαλείται από τα μηδενικά των ελλείψεων.

2.5 Κλασικές Μέθοδοι Πρόβλεψης Διαλείπουσας Ζήτησης

Πριν από την ευρεία υιοθέτηση των μοντέλων Μηχανικής Μάθησης, στο πεδίο της πρόβλεψης διαλείπουσας ζήτησης δέσποζε, για περισσότερες από τρεις δεκαετίες, η μέθοδος Croston [4] και οι, μετέπειτα, διορθωτικές παραλλαγές της. Ακόμη και σήμερα, αυτές οι παραδοσιακές στατιστικές μέθοδοι είναι ενσωματωμένες στις περισσότερες εμπορικές πλατφόρμες ERP και αποτελούν το απόλυτο μοντέλο αναφοράς (*standard baseline*) έναντι του οποίου αξιολογείται κάθε σύγχρονη υπολογιστική προσέγγιση [1], [21], [35].

2.5.1 Η μέθοδος Croston και η στατιστική της μεροληψία

Η κεντρική καινοτομία του Croston [4] ήταν η διαπίστωση ότι η απλή εκθετική εξομάλυνση αποτυγχάνει στα διαλείποντα δεδομένα, επειδή «σέρνεται» στο μηδέν μετά από ανενεργές περιόδους. Αντ' αυτού, πρότεινε την αποσύνθεση της ζήτησης σε δύο ξεχωριστές συνιστώσες, στο απόλυτο μέγεθος της ζήτησης (z_t) και στο μεσοδιάστημα μεταξύ διαδοχικών πωλήσεων (τ_t).

Οι εκτιμήσεις αυτών των δύο μεγεθών ($\overline{z_t}$ και $\overline{\tau_t}$) ανανεώνονται μέσω εκθετικής εξομάλυνσης με παράμετρο , αποκλειστικά και μόνο όταν καταγράφεται πώληση ($y_t \geq 0$). Τις ημέρες με μηδενική ζήτηση, οι εκτιμήσεις παραμένουν αμετάβλητες. Η τελική σημειακή πρόβλεψη προκύπτει από το πηλίκο τους:

$$\hat{y}_{t+1} = \frac{\tilde{z}_t}{\tilde{\tau}_t} \tag{2.25}$$

Παρά την εμπορική της επιτυχία, η μέθοδος πάσχει από ένα κρίσιμο μαθηματικό σφάλμα, το οποίο αναδείχθηκε δεκαετίες αργότερα από τους Shenstone & Hyndman [36]. Λόγω της Ανισότητας Jensen, η αναμενόμενη τιμή ενός πηλίκου είναι σταθερά μεγαλύτερη από το πηλίκο των αναμενόμενων τιμών.

$$\mathbb{E}\left[Z/_T\right] \leq \frac{\mathbb{E}[Z]}{\mathbb{E}[T]} \tag{2.26}$$

Επομένως, ο αλγόριθμος Croston είναι μαθηματικά μεροληπτικός προς τα άνω (*upward biased*) οδηγώντας τις αποθήκες σε συστηματική υπερεκτίμηση των αναγκών και, κατ' επέκταση, σε υπερ-παραγγελίες.

2.5.2 Η διόρθωση Syntetos–Boylan Approximation (SBA)

Για την εξάλειψη της μεροληψίας του Croston, οι Syntetos & Boylan [5] βασίστηκαν σε ανάπτυγμα σειράς Taylor δευτέρας τάξης αποδεικνύοντας ότι το σφάλμα υπερεκτίμησης σχετίζεται άμεσα με τη διακύμανση του μεσοδιαστήματος. Για να αντισταθμίσουν το σφάλμα, εισήγαγαν έναν αναλυτικό διορθωτικό συντελεστή $(1 - a/2)$, διαμορφώνοντας τον εκτιμητή **SBA**:

$$\hat{y}_{t+1}^{\mathrm{SBA}} = \left(1 - \frac{\alpha}{2}\right)\frac{\tilde{z}_t}{\tilde{\tau}_t} \tag{2.27}$$

Καθώς ο συντελεστής είναι μικρότερος της μονάδας, η μέθοδος SBA παράγει μετριοπαθέστερες και ακριβέστερες προβλέψεις. Σήμερα, η SBA έχει αντικαταστήσει τον απλό Croston ως η de facto βιομηχανική πρακτική για ανταλλακτικά με αραιή ζήτηση.

2.5.3 Η μέθοδος TSB και η θεωρητική σύνδεση με τα Μοντέλα Εμποδίου

Το σημαντικότερο επιχειρησιακό μειονέκτημα των Croston και SBA εμφανίζεται στο τέλος του κύκλου ζωής ενός ανταλλακτικού (απαρχαίωση / *obsolescence*). Επειδή οι παράμετροι ανανεώνονται μόνο, όταν υπάρχει πώληση, εάν ένα είδος σταματήσει οριστικά να πωλείται, η πρόβλεψη «παγώνει» εσαεί σε μια θετική τιμή συμβουλεύοντας τη διατήρηση νεκρού αποθέματος.

Τη λύση έδωσε η μέθοδος TSB (Teunter, Syntetos, Babai) [2], η οποία αντικατέστησε το μεσοδιάστημα (τ_t) με την πιθανότητα εμφάνισης θετικής ζήτησης (p_t).

$$p_t \;=\; \mathbb{1}_{(y_t>0)} \in \{0,1\} \tag{2.28}$$

Η πιθανότητα εξομαλύνεται με μια παράμετρο β σε κάθε χρονική περίοδο, ανεξάρτητα από το αν σημειώθηκε πώληση:

$$\tilde{p}_t \;=\; \beta\, p_t + (1-\beta)\,\tilde{p}_{t-1}, \quad \forall\, t \tag{2.29}$$

Η σημειακή πρόβλεψη μετατρέπεται πλέον στο γινόμενο:

$$\hat{y}_{t+1}^{\mathrm{TSB}} = \tilde{p}_t \cdot \tilde{z}_t \tag{2.30}$$

Εάν η ζήτηση σταματήσει, η $\tilde{p}_t$ φθίνει εκθετικά στο μηδέν, εξαλείφοντας τον κίνδυνο της απαρχαίωσης. Από μεθοδολογική σκοπιά, η προσέγγιση TSB είναι εξαιρετικά κρίσιμη για την παρούσα διπλωματική

εργασία. Ο διαχωρισμός της πρόβλεψης σε Πιθανότητα (*Probability*) και Υπό-Συνθήκη Μέγεθος (*Conditional Size*) καθιστά την TSB τον άμεσο, παραμετρικό πρόγονο των μοντέλων Εμποδίου δύο σταδίων (*two-stage* Hurdle *models*), τα οποία αναλύονται στην ενότητα 2.9 και αποτελούν τον πυρήνα της προτεινόμενης αρχιτεκτονικής Μηχανικής Μάθησης στο Κεφάλαιο 5.

2.5.4 Δομικοί περιορισμοί και η μετάβαση στη Μηχανική Μάθηση

Παρά τη βελτιωτική τους εξέλιξη, οι παραπάνω κλασικές μέθοδοι φέρουν εγγενείς περιορισμούς που υπαγόρευσαν τη μετάβαση στη σύγχρονη Επιστήμη Δεδομένων:

- **Μονομερής και τοπική εκπαίδευση:** Κάθε κωδικός είδους (SKU) μοντελοποιείται ανεξάρτητα καθιστώντας αδύνατη τη μάθηση κοινών μοτίβων (*cross-learning*) σε πολυμελείς καταλόγους χιλιάδων ειδών.
- **Μονομεταβλητή φύση (u*nivariate*):** Αδυνατούν να ενσωματώσουν πολύτιμα εξωγενή χαρακτηριστικά (*features*), όπως οι δείκτες περικοπής ζήτησης (OOS *flags*) που εξετάστηκαν στην ενότητα 2.4.
- **Σημειακή αδυναμία:** Παράγουν αποκλειστικά μια σημειακή κεντρική εκτίμηση, αδυνατώντας να εκτιμήσουν απευθείας τα ακραία ποσοστημόρια (*quantiles*) που απαιτούνται για τον βέλτιστο υπολογισμό των σημείων αναπαραγγελίας (s, S).

Αυτοί οι τρεις περιορισμοί αποτελούν το εφαλτήριο για την εισαγωγή των αλγορίθμων Μηχανικής Μάθησης που ακολουθούν στις ενότητες 2.6–2.9. Παρά τις αδυναμίες τους, τα κλασικά μοντέλα διατηρούνται στην παρούσα εργασία (Κεφάλαια 5 και 6) ως η απαραίτητη βάση αναφοράς, προκειμένου να τεκμηριωθεί με αυστηρότητα η πραγματική προστιθέμενη αξία των σύγχρονων προσεγγίσεων στο επιχειρησιακό περιβάλλον.

2.6 Βασικές έννοιες Μηχανικής Μάθησης και Πρόβλεψη Ζήτησης

Η Μηχανική Μάθηση (Machine Learning – ML) αποτελεί τον κλάδο της τεχνητής νοημοσύνης που μελετά αλγόριθμους ικανούς να «εκπαιδεύονται» από ιστορικά δεδομένα και να βελτιώνουν την προγνωστική τους ικανότητα χωρίς την ανάγκη ρητής, χειροκίνητης κωδικοποίησης πολύπλοκων κανόνων [37], [38]. Στο πεδίο της πρόβλεψης ζήτησης, η Μηχανική Μάθηση αξιοποιείται για να προσεγγίσει μια υποκείμενη, συνήθως εξαιρετικά πολύπλοκη και μη γραμμική, συνάρτηση που χαρτογραφεί τα ιστορικά χαρακτηριστικά της ζήτησης (μαζί με εξωγενείς πληροφορίες) στις μελλοντικές της τιμές.

Η τελευταία δεκαετία υπήρξε καθοριστική για την εξέλιξη του πεδίου, γνωρίζοντας εκρηκτική ανάπτυξη. Ορόσημο αποτέλεσε ο παγκόσμιος διαγωνισμός πρόβλεψης χρονοσειρών του διαγωνισμού M5 [21], ο οποίος επισφράγισε τη συστηματική υπεροχή των σύγχρονων μοντέλων Μηχανικής Μάθησης έναντι των κλασικών στατιστικών προσεγγίσεων (όπως αυτές που παρουσιάστηκαν στην ενότητα 2.4). Ωστόσο, η εφαρμογή αυτών των αλγορίθμων σε χρονοσειρές συνοδεύεται από κρίσιμες μεθοδολογικές παγίδες [39], [40] , οι οποίες πρέπει να αποφευχθούν κατά τον σχεδιασμό των πειραμάτων. Η παρούσα ενότητα θεμελιώνει το αλγοριθμικό πλαίσιο και τις θεωρητικές αρχές που διέπουν την αρχιτεκτονική του Κεφαλαίου 5.

2.6.1 Επιβλεπόμενη μάθηση, συναρτήσεις κόστους και χρονικά συνεπής διαχωρισμός

Η αρχιτεκτονική της παρούσας διπλωματικής εντάσσεται αυστηρά στο παράδειγμα της **Επιβλεπόμενης Μάθησης (Supervised Learning)**. Σε αυτό το πλαίσιο, τα διαθέσιμα ιστορικά δεδομένα αναπαρίστανται ως μια συλλογή ζευγών (x_i, y_i). Κάθε δείγμα i αποτελείται από ένα διάνυσμα

χαρακτηριστικών $\mathbf{x} \in \mathbb{R}^p$ (που περιλαμβάνει π.χ. ιστορικές υστερήσεις, κυλιόμενους μέσους, ημερολογιακούς δείκτες και τον δείκτη `OOS` της ενότητας 2.4) και μία μεταβλητή-στόχο (η μελλοντική ποσότητα της ζήτησης ή μια δυαδική ένδειξη ύπαρξης ζήτησης).

Ο απώτερος σκοπός είναι η ανακάλυψη μιας συνάρτησης υπόθεσης $f_\theta(\mathbf{x})$, με ρυθμίσιμες παραμέτρους θ, η οποία να μπορεί να γενικεύει με επιτυχία σε νέα, αθέατα δεδομένα. Η βελτιστοποίηση των παραμέτρων επιτυγχάνεται μέσω της ελαχιστοποίησης μιας αντικειμενικής συνάρτησης (συνάρτησης απώλειας – *loss function*), η οποία συχνά ενσωματώνει και έναν όρο κανονικοποίησης [38]:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(y_i,\, f_\theta(\mathbf{x}_i)\right) + \lambda \Omega(\theta) \tag{2.31}$$

όπου $\mathcal{L}(\cdot)$ είναι το σφάλμα μεταξύ της παρατηρούμενης τιμής και της πρόβλεψης, $\Omega(\theta)$ είναι ο όρος που αποτρέπει την υπερβολική πολυπλοκότητα του μοντέλου, και λ η υπερπαράμετρος που ελέγχει την ένταση της κανονικοποίησης.

Η πρόκληση του διαχωρισμού των δεδομένων

Στις παραδοσιακές εφαρμογές, τα δεδομένα διαχωρίζονται (συχνά τυχαία) σε τρία ανεξάρτητα σύνολα: Εκπαίδευσης (*Train*) για την εκτίμηση του θ, Επικύρωσης (*Validation*) για την επιλογή υπερπαραμέτρων (*early stopping*) και Ελέγχου ή Αξιολόγησης (*Test*) για την τελική αποτίμηση του συστήματος. Σε δεδομένα χρονοσειρών, ωστόσο, η τυχαιοποιημένη ανάμειξη είναι καταστροφική. Είναι προφανές ότι ένας τυχαίος διαχωρισμός, που δεν τηρεί τη χρονική ακολουθία, προκαλεί **διαρροή πληροφορίας από το μέλλον** (*data leakage*) επιτρέποντας στο μοντέλο να "δει" μελλοντικές παρατηρήσεις κατά τη διάρκεια της εκπαίδευσης. Αυτό, κατά συνέπεια, οδηγεί σε πλασματικά άριστα αποτελέσματα, τα οποία καταρρέουν κατά τη γενίκευση σε νέα δεδομένα [39], [41].

Για την αποτροπή αυτού του φαινομένου, επιβάλλονται **χρονικά συνεπείς (*time-consistent*) στρατηγικές διαχωρισμού**:

- **Απλός χρονικός διαχωρισμός (*Holdout split*):** Επιλέγεται ένα αυστηρό χρονικό σημείο T_{split}. Κάθε παρατήρηση πριν το T_{split} ανήκει στο σύνολο εκπαίδευσης και κάθε μεταγενέστερη στο σύνολο ελέγχου.
- **Αναδρομικός έλεγχος σε κυλιόμενο παράθυρο (*Rolling-origin* / *Walk-forward validation*):** Πρόκειται για τον πλέον ενδεδειγμένο τρόπο αξιολόγησης [40]. Το μοντέλο εκπαιδεύεται επαναληπτικά σε ένα αυξανόμενο (ή κυλιόμενο) ιστορικό παράθυρο και δοκιμάζεται στο αμέσως επόμενο βήμα. Αυτό προσομοιώνει ακριβώς τη βιομηχανική πρακτική, κατά την οποία τα μοντέλα εκπαιδεύονται εκ νέου κάθε μήνα με τα νέα δεδομένα και παράγει εξαιρετικά στιβαρές στατιστικές εκτιμήσεις.

Στα περιβάλλοντα διαλείπουσας ζήτησης (όπως η παρούσα διπλωματική εργασία), οι «αόρατες» παγίδες διαρροής είναι συχνές. Περιλαμβάνουν χαρακτηριστικά (π.χ. συνολικός μέσος όρος του προϊόντος) που υπολογίζονται επί όλου του ιστορικού, προτού γίνει ο διαχωρισμός, ή την κωδικοποίηση κατηγορικών μεταβλητών (*target encoding*) χωρίς κατάλληλη εκτός πτυχής (*out-of-fold*) στρατηγική. Στον σχεδιασμό του Κεφαλαίου 5 έχουν ληφθεί αυστηρά προγραμματιστικά μέτρα για την πλήρη στεγανοποίηση των χρονικών παραθύρων.

2.6.2 Ισορροπία μεροληψίας – διακύμανσης και κανονικοποίηση

Η θεμελιώδης θεωρητική πρόκληση σε κάθε μοντέλο πρόβλεψης είναι η αναζήτηση της βέλτιστης ισορροπίας (*trade-off*) μεταξύ υποπροσαρμογής (*underfitting* / έντονη μεροληψία – *bias*) και

(2.32)

υπερπροσαρμογής (*overfitting* / υψηλή διακύμανση – *variance*). Μαθηματικά, το αναμενόμενο τετραγωνικό σφάλμα οποιουδήποτε μοντέλου αναλύεται μέσω της κλασικής αποσύνθεσης [38]:

$$\mathbb{E}\left[\left(y-\hat{f}(\mathbf{x})\right)^2\right] = \underbrace{\left(\mathbb{E}[\hat{f}(\mathbf{x})]-f(\mathbf{x})\right)^2}_{\text{bias}^2} + \underbrace{\text{Var}\left(\hat{f}(\mathbf{x})\right)}_{\text{variance}} + \underbrace{\sigma_\varepsilon^2}_{\text{irreducible noise}} \quad (2.32)$$

Απλά (ή γραμμικά) μοντέλα έχουν υψηλή μεροληψία, καθώς αδυνατούν να συλλάβουν σύνθετα μοτίβα (υποπροσαρμογή). Από την άλλη, εξαιρετικά πολύπλοκα μοντέλα με χιλιάδες παραμέτρους τείνουν να απομνημονεύουν τον στατιστικό θόρυβο των δεδομένων οδηγώντας σε υψηλή διακύμανση και πλήρη κατάρρευση στα νέα δεδομένα (υπερπροσαρμογή).

Η διαδικασία που «χαλιναγωγεί» την πολυπλοκότητα των αλγορίθμων ονομάζεται **Κανονικοποίηση (*Regularization*)**. Ανάλογα με την οικογένεια των μοντέλων, χρησιμοποιούνται διαφορετικοί μηχανισμοί:

- Σε γραμμικά μοντέλα, προστίθενται ποινές επί των συντελεστών: L1 (Lasso) για να μηδενιστούν άχρηστες μεταβλητές, ή L2 (Ridge) για να αποτραπούν τεράστιες τιμές στα βάρη [42].
- Στα δέντρα απόφασης (όπως ο XGBoost που αναλύεται εκτενώς στην ενότητα 2.7), επιβάλλονται δομικοί περιορισμοί (μέγιστο βάθος δέντρου, ελάχιστος αριθμός δειγμάτων ανά φύλλο) ή εφαρμόζεται στοχαστική δειγματοληψία (*subsampling*).
- Στα Νευρωνικά Δίκτυα, κυριαρχεί η πρόωρη διακοπή (*early stopping*) και η παράλειψη νευρώνων (*dropout)* [43].

Στην προτεινόμενη αρχιτεκτονική (Κεφ. 5), όλοι αυτοί οι μηχανισμοί δεν ορίζονται αυθαίρετα, αλλά ρυθμίζονται δυναμικά μέσω εκτεταμένης διασταυρούμενης επικύρωσης (*cross-validation*) επί του αντίστοιχου συνόλου.

2.6.3 Κατηγορίες μοντέλων για πρόβλεψη ζήτησης

Η σύγχρονη έρευνα κατατάσσει τα μοντέλα Μηχανικής Μάθησης που εφαρμόζονται στις χρονοσειρές σε τέσσερις κυρίαρχες οικογένειες [44], [45]:

- **Γραμμικά μοντέλα παλινδρόμησης (*Linear/Penalized models*):** Πρόκειται για επεκτάσεις της κλασικής παλινδρόμησης (Ridge, Lasso, Elastic Net). Παρέχουν την υψηλότερη δυνατή ερμηνευσιμότητα και λειτουργούν εξαιρετικά ως γραμμές βάσης (*baselines*), αλλά υστερούν δραματικά όταν υπάρχουν σύνθετες μη γραμμικές εξαρτήσεις.
- **Δέντρα Απόφασης και Μέθοδοι Συνόλου (*Tree-based ensembles*):** Περιλαμβάνουν τους αλγορίθμους «Τυχαίων Δασών» (Random Forests) και «Εξαιρετικά Τυχαίων Δέντρων» (Extra Trees) και, κυρίως, τους αλγορίθμους Διαβαθμισμένης Ενίσχυσης (*Gradient Boosting*, όπως XGBoost, LightGBM, CatBoost). Όπως αποδείχθηκε από τον διαγωνισμό M5, αποτελούν την απόλυτη αιχμή του δόρατος (*state-of-the-art*) σε πίνακες δεδομένων (*tabular data*) και προβλήματα πολλαπλών, ταυτόχρονων σειρών αποτελώντας τον βασικό κορμό (*backbone*) των αλγορίθμων του Κεφαλαίου 5.
- **Νευρωνικά Δίκτυα και Βαθιά Μάθηση (Deep Learning):** Αναδρομικά δίκτυα (RNN/LSTM/GRU), Συνελικτικά δίκτυα (CNNs) και, προσφάτως, αρχιτεκτονικές *Attention/Transformers*. Παρέχουν τεράστια ευελιξία στην αναγνώριση μακροπρόθεσμων εξαρτήσεων, αλλά απαιτούν θηριώδεις όγκους δεδομένων για να αποφύγουν την υπερπροσαρμογή.

- **Πιθανοκρατικές αρχιτεκτονικές (*Probabilistic/Foundation models*):** Σύγχρονες υλοποιήσεις (π.χ. DeepAR, TFT) που έχουν σχεδιαστεί ρητά για να παράγουν ολόκληρες κατανομές αντί για σημειακές προβλέψεις ικανοποιώντας τις ανάγκες των αποθηκών που συζητήθηκαν στην ενότητα 2.3.3.

Σε απαιτητικά βιομηχανικά συστήματα (όπως αυτό της ΔΕ), η επιλογή δεν περιορίζεται σε έναν μόνο αλγόριθμο. Εφαρμόζονται προηγμένες τεχνικές συγχώνευσης (όπως η Συσσωρευτική Γενίκευση – *Stacking*, που αναλύεται στην ενότητα 2.7.5) αξιοποιώντας την αρχή ότι διαφορετικά μοντέλα κάνουν συχνά διαφορετικά, μη συσχετισμένα λάθη.

2.6.4 Καθολικά (*Global*) έναντι Τοπικών (*Local*) μοντέλων: Η κρίσιμη αλλαγή παραδείγματος

Ίσως η πιο ριζοσπαστική καινοτομία που έφερε η Μηχανική Μάθηση στην πρόβλεψη αποθεμάτων είναι η μετάβαση από τα μοντέλα τοπικής εμβέλειας (*local models*) στα **παγκόσμια (*global models*)** [44], [46].

- Στα **Τοπικά μοντέλα (*Local*):** Κάθε χρονοσειρά (κάθε κωδικός είδους/SKU) απομονώνεται. Ο αλγόριθμος (π.χ. ARIMA ή Croston) εκπαιδεύεται χρησιμοποιώντας αποκλειστικά και μόνο το ιστορικό αυτού του συγκεκριμένου κωδικού. Εάν ο κωδικός είναι νέος (*cold-start*) ή έχει πωληθεί μόνο 3 φορές (αραιή ζήτηση), το μοντέλο αποτυγχάνει παταγωδώς.
- Στα **Παγκόσμια/Καθολικά μοντέλα (*Global*):** Ένας **ενιαίος** εκτιμητής εκπαιδεύεται χρησιμοποιώντας τα δεδομένα από *όλους* τους κωδικούς του καταλόγου, ταυτόχρονα, σε μορφή ενιαίου πίνακα (*panel data*). Το μοντέλο μαθαίνει κοινούς υποκείμενους παράγοντες και δομές συσχέτισης (*cross-learning*).

Το πλεονέκτημα των καθολικών μοντέλων είναι ανεκτίμητο στον τομέα των ανταλλακτικών. Για παράδειγμα, η αυτόνομη χρονοσειρά ενός συγκεκριμένου ιμάντα διαθέτει ελάχιστο «σήμα» λόγω των υπερβολικών μηδενικών. Όταν, όμως, το μοντέλο εκπαιδευτεί ταυτόχρονα σε χιλιάδες ιμάντες (χρησιμοποιώντας εξωγενή χαρακτηριστικά όπως κατηγορία, μάρκα οχήματος, τιμή), συγκεντρώνει έναν τεράστιο όγκο πληροφορίας. Η νίκη των καθολικών δέντρων απόφασης (*Global* GBTs) στο σύνολο σχεδόν των κατηγοριών του διαγωνισμού M5 [21] επισφράγισε αυτή την αλλαγή παραδείγματος.

Στην παρούσα διπλωματική εργασία, εγκαταλείπεται πλήρως το παραδοσιακό παράδειγμα της ανεξάρτητης εκπαίδευσης ανά κωδικό («*per-SKU*») παράδειγμα της ενότητας 2.4. Όλοι οι προτεινόμενοι εκτιμητές του Κεφαλαίου 5 (τόσο ο Ταξινομητής όσο και ο Παλινδρομητής στα μοντέλα Hurdle) έχουν αναπτυχθεί ρητά ως καθολικά μοντέλα. Μέσω εκτεταμένης μηχανικής χαρακτηριστικών (*feature engineering*), το μοντέλο τροφοδοτείται με τη συλλογική γνώση ολόκληρου του καταλόγου ανταλλακτικών και πετυχαίνει πρωτοφανή σταθερότητα και ικανότητα γενίκευσης, ακόμη, και στα είδη με τη χαμηλότερη δυνατή κυκλοφορία (*sparse* SKUs).

2.7 Δέντρα Διαβαθμισμένης Ενίσχυσης και συσσωρευτική γενίκευση (*stacking ensembles*)

Από τη συντριπτική κυριαρχία τους στον παγκόσμιο διαγωνισμό M5 [21] μέχρι τη συστηματική τους χρήση στις πλέον σύγχρονες βιομηχανικές εφαρμογές [47], τα Δέντρα Διαβαθμισμένης Ενίσχυσης (*Gradient Boosted Trees* - GBT) έχουν αναδειχθεί ως το προεπιλεγμένο μοντέλο για μη γραμμικά προβλήματα παλινδρόμησης σε δομημένα δεδομένα (*tabular data*). Η σαρωτική επιτυχία τους

οφείλεται σε έναν συνδυασμό ιδιοτήτων που ευθυγραμμίζονται απόλυτα με το πρόβλημα της διαλείπουσας ζήτησης. Καταρχήν, είναι εγγενώς ανθεκτικά σε μη γραμμικές και πολύπλοκες αλληλεπιδράσεις, διαχειρίζονται αυτόματα μικτά (αριθμητικά και κατηγορικά) χαρακτηριστικά, κλιμακώνονται υπολογιστικά σε εκατοντάδες χιλιάδες χρονοσειρές ταυτόχρονα και υποστηρίζουν άμεσα την παραμετροποίηση ασύμμετρων συναρτήσεων κόστους (Pinball, Tweedie, Huber).

Η παρούσα ενότητα θεμελιώνει αλγοριθμικά τις μεθόδους που συνιστούν τον κορμό της αρχιτεκτονικής του Κεφαλαίου 5. Από τα βασικά δέντρα αποφάσεων και τις προσεγγίσεις *bagging* έως τη θεωρία του *gradient boosting*, τις σύγχρονες υλοποιήσεις (XGBoost, LightGBM, CatBoost) και τις προηγμένες τεχνικές συγχώνευσης πολλαπλών μοντέλων (*stacking*).

2.7.1 Δέντρα Αποφάσεων (CART)

Η αρχιτεκτονική των Δέντρων Ταξινόμησης και Παλινδρόμησης (CART – *Classification and Regression Trees*) των Breiman, Friedman, Olshen και Stone [48] αποτελεί τον δομικό λίθο όλων των δενδρικών μοντέλων. Ένα δέντρο παλινδρόμησης διαμερίζει αναδρομικά (*recursive partitioning*) τον πολυδιάστατο χώρο των χαρακτηριστικών εισόδου σε ορθογώνιες διαμερίσεις (κόμβους-φύλλα) $\{R_1, R_2, \ldots, R_J\}$. Σε κάθε διαμέριση R_j αντιστοιχίζεται μια σταθερή πρόβλεψη w_j, η οποία τυπικά ισούται με τη μέση τιμή των παρατηρήσεων που καταλήγουν στο συγκεκριμένο φύλλο. Η συνολική συνάρτηση πρόβλεψης του μοντέλου εκφράζεται ως:

$$T(\mathbf{x}) = \sum_{j=1}^{J} w_j \cdot \mathbb{1}_{(\mathbf{x} \in R_j)} \tag{2.33}$$

Η κατασκευή του δέντρου προχωρά μέσω άπληστης διαμέρισης (*greedy splitting*). Δηλαδή, σε κάθε κόμβο (με δεδομένα D_t), ο αλγόριθμος αναζητά το βέλτιστο ζεύγος χαρακτηριστικού και κατωφλίου(j, s), το οποίο ελαχιστοποιεί το συνολικό άθροισμα των τετραγωνικών σφαλμάτων των δύο νέων κόμβων-παιδιών:

$$\min_{j,s} \left[\sum_{i \in L(j,s)} (y_i - \bar{y}_L)^2 + \sum_{i \in R(j,s)} (y_i - \bar{y}_R)^2 \right] \tag{2.34}$$

όπου $L(j, s) = \{i : x_{i,j} \leq s\}$ και $R(j, s) = \{i : x_{i,j} > s\}$ είναι το αριστερό και δεξί υποσύνολο αντίστοιχα, ενώ $\bar{y}_L, \bar{y}_R$ οι τοπικές μέσες τιμές τους. Η αναδρομή τερματίζεται, όταν ικανοποιηθεί κάποιο προκαθορισμένο κριτήριο (π.χ. μέγιστο βάθος δέντρου ή ελάχιστος αριθμός δειγμάτων ανά φύλλο).

Τα απλά δέντρα διαθέτουν θεμελιώδη πλεονεκτήματα, μοντελοποιούν άμεσα τις μη γραμμικές αλληλεπιδράσεις των μεταβλητών και δεν απαιτούν κανονικοποίηση δεδομένων (*data normalization*). Το συντριπτικό τους μειονέκτημα, ωστόσο, είναι η υψηλή διακύμανση (*variance*). Ένα πλήρως αναπτυγμένο δέντρο υπερπροσαρμόζεται εύκολα, και μικρές διακυμάνσεις στα δεδομένα εκπαίδευσης οδηγούν σε ριζικά διαφορετικές δομές. Η αντιμετώπιση αυτού του προβλήματος οδήγησε ιστορικά στην ανάπτυξη των μεθόδων συνόλου (*ensemble methods*).

2.7.2 Μέθοδοι *Bagging* και Τυχαία Δάση (Random Forests)

Η τεχνική του Συγκεντρωτικού Ανασυνδυασμού (*Bootstrap Aggregating* ή *Bagging*) που εισήγαγε ο Breiman [49] στοχεύει στη δραστική μείωση της διακύμανσης ενός μοντέλου υπολογίζοντας τον μέσο όρο πολλαπλών, ανεξάρτητων εκδοχών του , καθεμία εκπαιδευμένη σε διαφορετικό δείγμα. Για ένα

(2.35)

πρόβλημα παλινδρόμησης, εάν T_m είναι το m -οστό δέντρο, εκπαιδευμένο σε ένα τυχαίο δείγμα επανατοποθέτησης (*bootstrap sample*) $D^{(m)}$, η τελική πρόβλεψη είναι:

$$\hat{f}_{\text{bag}}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} T_m(\mathbf{x}) \tag{2.35}$$

Εάν υποτεθεί ότι όλα τα δέντρα T_m έχουν παρόμοια διακύμανση σ^2 και μια μέση μεταξύ τους συσχέτιση ρ, τότε η διακύμανση του συνολικού (*bagged*) μοντέλου δίνεται από τη σχέση:

$$\text{Var}(\hat{f}_{\text{bag}}) = \rho\sigma^2 + \frac{1-\rho}{M}\sigma^2 \tag{2.36}$$

Αυτή η εξίσωση αποκαλύπτει ότι το μαθηματικό όφελος της μεθόδου εξαρτάται κρίσιμα από την από-συσχέτιση (*decorrelation*) των επιμέρους δέντρων. Όσο πιο ανεξάρτητα είναι, τόσο πιο γρήγορα μειώνεται η συνολική διακύμανση καθώς αυξάνεται το πλήθος M.

Για να επιτευχθεί αυτό, ο Breiman [50] σχεδίασε τα **Τυχαία Δάση (Random Forests – RF)**, τα οποία προσθέτουν μια δεύτερη πηγή τυχαιότητας. Σε κάθε διάσπαση κόμβου (*node split*), ο αλγόριθμος δεν εξετάζει όλες τις μεταβλητές, αλλά επιλέγει ένα τυχαίο υποσύνολο m χαρακτηριστικών από τα συνολικά p (συνήθως $m \approx \sqrt{p}$ για ταξινόμηση και $m \approx p/3$ για παλινδρόμηση). Αυτό εξαναγκάζει τα δέντρα να εξερευνούν διαφορετικές διαστάσεις του χώρου χαρακτηριστικών μειώνοντας τη μεταξύ τους συσχέτιση ρ και βελτιώνοντας την ικανότητα γενίκευσης. Πρόσθετο πρακτικό πλεονέκτημα είναι η εκτίμηση του «εκτός-σάκου» σφάλματος (*out-of-bag error* ή OOB *error*). Δηλαδή, τα δείγματα που δεν συμμετείχαν στον ανασυνδυασμό (*bootstrap*) εκάστου δέντρου αποτελούν ένα φυσικό σύνολο αξιολόγησης (*holdout*) και συμβάλλουν στην εκτίμηση της γενικευτικής ικανότητας χωρίς την ανάγκη ξεχωριστού συνόλου επικύρωσης (*validation set*). Στο υπολογιστικό *pipeline* της παρούσας εργασίας, τόσο ο `RandomForestClassifier` (για το στάδιο της εμφάνισης ζήτησης – *occurrence*), όσο και ο `RandomForestRegressor` (για το στάδιο του μεγέθους θετικής ζήτησης – *positive size*) χρησιμοποιούνται ως εναλλακτική βάση του Hurdle μοντέλου (`Hurdle_RF`) διατηρώντας, παράλληλα, τη δυνατότητα εξαγωγής ποσοστημορίων μέσω των εκατοστημορίων επιπέδου δέντρου (*tree-level percentiles*).

Μια ακόμη πιο επιθετική παραλλαγή είναι τα **Εξαιρετικά Τυχαία Δέντρα (Extra Trees – ET)** των Geurts, Ernst & Wehenkel [51]. Η ονομασία προκύπτει από τη σύμπτυξη της φράσης «*Extremely Randomised Trees*». Σε αντίθεση με τον Random Forests, όπου για κάθε κόμβο επιλέγεται το βέλτιστο κατώφλι για κάθε ένα από τα τυχαία χαρακτηριστικά, τα Extra Trees επιλέγουν το κατώφλι s_j **τυχαία** από μια ομοιόμορφη κατανομή στο εύρος τιμών της μεταβλητής x_{ij} στον εκάστοτε κόμβο:

$$s_j \sim \mathcal{U}\left[\min_{i\in\mathcal{D}_t} x_{ij},\ \max_{i\in\mathcal{D}_t} x_{ij}\right], \qquad j = 1,\dots,m \tag{2.37}$$

και, στη συνέχεια, επιλέγεται το ζεύγος (j^*, s_{j^*}) που προσφέρει το μεγαλύτερο κέρδος πληροφορίας από τα m τυχαία ζεύγη. Η πλήρης τυχαιοποίηση του κατωφλίου διαγράφει τον χρονοβόρο υπολογισμό της βέλτιστης διάσπασης καθιστώντας την εκπαίδευση σημαντικά ταχύτερη. Μάλιστα, μειώνεται η πολυπλοκότητα σε $\mathcal{O}(K \cdot m \cdot N)$ αντί για $\mathcal{O}(K \cdot m \cdot N\log N)$ και ελαττώνεται περαιτέρω η μεταξύ-δενδρών συσχέτιση ρ. Το αποτέλεσμα είναι ένα μοντέλο με χαμηλότερη διακύμανση (*variance*) και υψηλότερη μεροληψία (*bias*) ανά μεμονωμένο δέντρο έναντι του Random Forest. Στο σύνολο, ωστόσο, η μειωμένη συσχέτιση αντισταθμίζει συχνά το αυξημένο *bias*, ιδίως σε δεδομένα με έντονο θόρυβο ή αραιή δομή. Στο προτεινόμενο *pipeline*, τα Extra Trees (`Hurdle_ET`) αποτελούν το τρίτο προεπιλεγμένο Hurdle μοντέλο και αξιοποιούνται κυρίως ως ένα ισχυρά κανονικοποιημένο (*regularized*) βασικό

μοντέλο και προσφέρουν σημαντικό όφελος στο επίπεδο του μετα-μοντέλου (*meta-learner / stacking*), ακριβώς λόγω της χαμηλής τους συσχέτισης με τα υπόλοιπα μοντέλα.

Παρά τη θεωρητική τους ελκυστικότητα και την ικανότητά τους να αποσβένουν τον θόρυβο, τα μοντέλα συνόλου που βασίζονται στο *bagging* (RF, ET) τείνουν σε πρόσφατες εμπειρικές συγκρίσεις να υστερούν συστηματικά σε ακρίβεια έναντι των μεθόδων διαβαθμισμένης ενίσχυσης (*gradient boosting*) σε προβλήματα πρόβλεψης ζήτησης [21]. Αυτό συμβαίνει, γιατί βελτιστοποιούν μεν τη διακύμανση, αλλά αδυνατούν να μειώσουν ενεργά τη μεροληψία (*bias*), στοιχείο που καθιστά αναγκαία τη μετάβαση στους αλγορίθμους συνδαυσμού συνόλων.

2.7.3 Διαβαθμισμένη Ενίσχυση (*Gradient Boosting*)

Σε αντίθεση με το *bagging*, κατά το οποίο τα δέντρα εκπαιδεύονται παράλληλα και ανεξάρτητα, η Ενίσχυση (*boosting*) τα εκπαιδεύει **διαδοχικά** (*sequentially*). Κάθε νέο δέντρο προστίθεται στο μοντέλο με αποκλειστικό σκοπό να διορθώσει τα σφάλματα που άφησαν τα προηγούμενα. Η σύγχρονη μαθηματική θεμελίωση αυτής της ιδέας ως μέθοδος καθόδου κλίσης σε χώρο συναρτήσεων (*functional gradient descent*) οφείλεται στους Friedman [52] και Mason *et al.* [53].

Έστω $\mathcal{L}(y, F)$ μια διαφορίσιμη συνάρτηση κόστους (loss function) και $F: X \to \mathbb{R}$ η προς υπολογισμό συνάρτηση πρόβλεψης. Η διαδικασία (*Gradient Boosting*) εξελίσσεται ως εξής: Αρχικά, το μοντέλο αρχικοποιείται με μια σταθερή τιμή που ελαχιστοποιεί τη συνάρτηση κόστους στο σύνολο εκπαίδευσης.

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{N} \mathcal{L}\left(y_i, \gamma\right) \tag{2.38}$$

Στη συνέχεια, για κάθε βήμα (δέντρο) $m = 1, \ldots, M$:

- Υπολογίζονται τα ψευδο-κατάλοιπα (*pseudo-residuals*), τα οποία αντιστοιχούν στις αρνητικές κλίσεις (*gradients*) της συνάρτησης κόστους:

$$r_{im} = -\left[\frac{\partial \mathcal{L}\big(y_i, F(\mathbf{x}_i)\big)}{\partial F(\mathbf{x}_i)}\right]_{F=F_{m-1}}, \quad i = 1, \ldots, N \tag{2.39}$$

- Ένα ρηχό δέντρο απόφασης (base learner) εκπαιδεύεται πάνω σε αυτά τα ψευδο-κατάλοιπα (χρησιμοποιώντας τον κλασικό αλγόριθμο ελαχιστοποίησης του MSE).
- Για κάθε φύλλο j του νέου δέντρου, υπολογίζεται το βέλτιστο βάρος γ_{jm} που ελαχιστοποιεί απευθείας την αρχική συνάρτηση κόστους $\mathcal{L}$:

$$\gamma_{jm} = \arg\min_{\gamma} \sum_{\mathbf{x}_i \in R_{jm}} \mathcal{L}\left(y_i,\, F_{m-1}(\mathbf{x}_i) + \gamma\right) \tag{2.40}$$

- Το συνολικό μοντέλο ανανεώνεται, πολλαπλασιάζοντας την προβλεπόμενη τιμή του νέου δέντρου με έναν ρυθμό μάθησης (*learning rate / shrinkage*) $\nu \in (0,1]$**:**

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \sum_{j} \gamma_{jm}\, \mathbb{1}_{(\mathbf{x} \in R_{jm})} \tag{2.41}$$

Αυτή η μαθηματική διατύπωση προσφέρει απόλυτη ευελιξία και επιτρέπει στο μοντέλο να αλλάζει τη "φύση" της πρόβλεψής του απλώς αλλάζοντας τη συνάρτηση $\mathcal{L}$. Για παράδειγμα, η χρήση της Pinball Loss κατευθύνει το μοντέλο στην πρόβλεψη ποσοστημορίων, ενώ η συνάρτηση απώλειας Tweedie

χρησιμοποιείται για δεδομένα με ακραία υπερδιασπορά (όπως η διαλείπουσα ζήτηση). Παράλληλα, ο ρυθμός μάθησης ν αποτελεί τον ισχυρότερο μηχανισμό κανονικοποίησης. Μικρές τιμές (π.χ. 0,05) αναγκάζουν το μοντέλο να μαθαίνει «συντηρητικά» αποφεύγοντας την υπερπροσαρμογή.

2.7.4 Σύγχρονες βιομηχανικές υλοποιήσεις: XGBoost, LightGBM, CatBoost και HistGradientBoosting

Η θεωρητική σύλληψη του *Gradient Boosting* κατέστησε δυνατή την κυριαρχία των GBTs, ωστόσο η βιομηχανική τους υιοθέτηση κατέστη δυνατή μόνο μετά τη δημιουργία εξαιρετικά βελτιστοποιημένων υλοποιήσεων την τελευταία δεκαετία, οι οποίες εισάγουν αλγοριθμικές καινοτομίες πέρα από την αρχική διατύπωση.

XGBoost [54].

Αποτελεί την πιο διαδεδομένη υλοποίηση εισάγοντας δύο κεντρικές καινοτομίες. Πρώτον, ενσωματώνει ρητά όρους ποινής κανονικοποίησης L1/L2 στην ίδια την αντικειμενική συνάρτηση του δέντρου (ελέγχοντας το πλήθος και το βάρος των φύλλων). Δεύτερον, χρησιμοποιεί προσέγγιση σειράς Taylor δεύτερης τάξης επιταχύνοντας εντυπωσιακά τη σύγκλιση:

$$\tilde{\mathcal{L}}^{(t)} \approx \sum_{i=1}^{N} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t(\mathbf{x}_i)^2 \right] + \Omega(f_t) \tag{2.42}$$

όπου g_i και h_i είναι η πρώτη (*gradient*) και η δεύτερη (*hessian*) παράγωγος της συνάρτησης κόστους, αντίστοιχα. Αυτό επιτρέπει τον αναλυτικό υπολογισμό του κέρδους (*gain*) κάθε υποψήφιας διάσπασης. Επιπλέον, το XGBoost εισάγει την εύρεση διάσπασης με επίγνωση αραιότητας (*sparsity-aware split finding*) για την αυτόματη διαχείριση ελλιπών τιμών καθιστώντας το σημαντικά ταχύτερο από τις κλασικές υλοποιήσεις.

LightGBM [55].

Η συγκεκριμένη υλοποίηση επικεντρώνεται στην αλγοριθμική επιτάχυνση (υπολογιστική αποδοτικότητα) μέσω δύο τεχνικών:

- **Μονόπλευρη Δειγματοληψία Βάσει Κλίσης** (*Gradient-based One-Side Sampling* - GOSS): Διατηρεί όλες τις παρατηρήσεις με μεγάλες κλίσεις (*gradients*), οι οποίες συνεισφέρουν περισσότερο στη μάθηση και απορρίπτει τυχαία ορισμένες από εκείνες που παρουσιάζουν μικρές κλίσεις (υπο-δειγματοληψία). Έτσι, μειώνεται ο υπολογιστικός φόρτος χωρίς σημαντική απώλεια ακρίβειας.
- **Αποκλειστική Ομαδοποίηση Χαρακτηριστικών** (*Exclusive Feature Bundling* - EFB): Ομαδοποιεί αμοιβαία αποκλειόμενες αραιές μεταβλητές (που σπάνια λαμβάνουν τιμή ταυτόχρονα) σε ένα ενιαίο χαρακτηριστικό μειώνοντας τη διάσταση του προβλήματος.

Επίσης, το LightGBM χρησιμοποιεί στρατηγική ανάπτυξης κατά φύλλο (*leaf-wise growth*) σε αντίθεση με την ισοβαρή ανάπτυξη κατά επίπεδο (*level-wise growth*) του XGBoost παράγοντας βαθύτερα και ασύμμετρα δέντρα. Το πιθανό μειονέκτημα αυτού, όμως, είναι ο μεγαλύτερος κίνδυνος υπερπροσαρμογής (*overfitting*).

CatBoost [56].

Η νεότερη από τις τρεις αρχιτεκτονικές, σχεδιασμένη ειδικά για την απρόσκοπτη διαχείριση κατηγορικών μεταβλητών (*categorical features*). Αντί να απαιτεί προ-επεξεργασία, όπως το *one-hot encoding*, χρησιμοποιεί διατεταγμένα στατιστικά στόχου (*ordered target statistics*). Το μεγαλύτερο

πλεονέκτημά της είναι ο μηχανισμός διατεταγμένης ενίσχυσης (*ordered boosting*). Βάσει αυτού, κατά την εκπαίδευση υπολογίζεται η κλίση (*gradient*) μιας παρατήρησης από ένα μοντέλο που έχει εκπαιδευτεί μόνο σε προηγούμενες χρονικά παρατηρήσεις εξαλείφοντας πλήρως τον κίνδυνο διαρροής στόχου (*target leakage*) που εμφανίζεται στους παραδοσιακούς αλγορίθμους.

HistGradientBoosting (`scikit-learn`) [57]. Η εγγενής υλοποίηση GBT της βιβλιοθήκης `scikit-learn`, εμπνευσμένη από το LightGBM, αξιοποιεί την ομαδοποίηση ιστογράμματος (*histogram binning*). Κάθε συνεχές χαρακτηριστικό ομαδοποιείται σε K ισο-πληθή διαστήματα (*bins*, προεπιλογή $K = 255$) και η αναζήτηση της βέλτιστης διάσπασης γίνεται μόνο στα ακέραια όριά τους. Υπάρχουν τρία πρακτικά πλεονεκτήματα που το διαφοροποιούν:

- **Εγγενείς ελλιπείς τιμές** (*Native missing values*): Οι απουσιάζουσες τιμές (`NaN`) αντιμετωπίζονται εγγενώς ως μια ξεχωριστή κατηγορία, χαρακτηριστικό ιδιαίτερα σημαντικό για αραιούς πίνακες διαλείπουσας ζήτησης.
- **Ενσωματωμένος πρόωρος τερματισμός** (*Early stopping*): Η παράμετρος `n_iter_no_change` παρακολουθεί το σφάλμα επικύρωσης και τερματίζει τη διαδικασία αυτόματα παρακάμπτοντας την ανάγκη ορισμού σταθερού πλήθους δέντρων (`n_estimators`).
- **Μηδενικές εξωτερικές εξαρτήσεις**: Εκτελείται απρόσκοπτα μέσω της τυπικής εγκατάστασης της βιβλιοθήκης `scikit-learn`.

Στην υπολογιστική ροή της παρούσας εργασίας, το `HistGradientBoosting` αποτελεί το **κύριο προεπιλεγμένο μοντέλο εμποδίου** (`Hurdle_HGB`). Τόσο η κλάση `HistGradientBoostingClassifier` (για την εκτίμηση της πιθανότητας p_t) όσο και η `HistGradientBoostingRegressor` (για την εκτίμηση του μεγέθους ζήτησης $\mu_t{}^+$) εκπαιδεύονται ανά ορίζοντα πρόβλεψης.

Όλες οι παραπάνω βιβλιοθήκες υποστηρίζουν απευθείας τις προσαρμοσμένες συναρτήσεις κόστους (*custom loss functions*) που απαιτούνται στο Κεφάλαιο 5, όπως οι `Pinball Loss`, `Tweedie deviance` και `Huber`. Στον Πίνακα 2.2 συγκρίνονται τα βασικά χαρακτηριστικά των τεσσάρων αυτών αρχιτεκτονικών.

Πίνακας 2.2: Σύγκριση αρχιτεκτονικών XGBoost, LightGBM και CatBoost

Χαρακτηριστικό	**XGBoost**	**LightGBM**	**CatBoost**	**HistGradientBoosting**
Στρατηγική Ανάπτυξης	Κατά επίπεδο (*level-wise*)	Κατά φύλλο (*leaf-wise*)	Συμμετρική (*oblivious*)	Κατά φύλλο (*leaf-wise*)
Διαχείριση Κενών (`NaN`)	Εγγενής (*sparsity-aware*)	Εγγενής	Εγγενής	Εγγενής (ως ξεχωριστό bin)
Κατηγορικές Μεταβλητές	Απαιτεί εξωτερική κωδικοποίηση	Ενσωματωμένη βέλτιστη διάσπαση	Διατεταγμένα στατιστικά στόχου (*ordered target statistics*)	Εγγενής υποστήριξη
Διαρροή Πληροφορίας	Δυνητικά παρούσα	Δυνητικά παρούσα	Αποτρέπεται ρητά (*ordered boosting*)	Δυνητικά παρούσα
Κύριο Πλεονέκτημα (για το Pipeline)	Στιβαρότητα μέσω αυστηρής κανονικοποίησης	Υψηλή ταχύτητα σε μεγάλα σύνολα δεδομένων (GOSS, EFB)	Εγγενής βελτιστοποιημένη διαχείριση κατηγορικών μεταβλητών και GPU	Υψηλή ταχύτητα εκτέλεσης, ενσωματωμένο early stopping και απουσία εξαρτήσεων

2.7.5 Συσσωρευτική Γενίκευση (*Stacking Ensembles*)

Παρά την εξαιρετική απόδοση των μεμονωμένων GBTs, σε βιομηχανικά περιβάλλοντα (και σε διαγωνισμούς προβλέψεων) χρησιμοποιούνται σχεδόν καθολικά **αρχιτεκτονικές συνόλου πολλαπλών μοντέλων (*multi-model ensembles*)**.

Η απλούστερη προσέγγιση συνίσταται στον σταθμισμένο γραμμικό μέσο όρο των προβλέψεων διαφορετικών μοντέλων:

$$\hat{y}(\mathbf{x}) = \sum_{m=1}^{M} w_m \cdot \hat{y}_m(\mathbf{x}), \quad \sum_{m=1}^{M} w_m = 1, \quad w_m \geq 0 \tag{2.43}$$

Η πλέον προηγμένη και θεωρητικά ισχυρή προσέγγιση, ωστόσο, είναι η **Συσσωρευτική Γενίκευση (*Stacked Generalization / Stacking*)** που θεμελίωσε ο Wolpert [58]. Αντί να ορίζονται στατικά βάρη w_m, εκπαιδεύεται ένα νέο, ανώτερου επιπέδου μοντέλο μετα-μάθησης (το *meta-learner*), για να μάθει τον βέλτιστο τρόπο συνδυασμού των βασικών (*base*) μοντέλων. Η διαδικασία απαιτεί αυστηρή διαχείριση για την αποτροπή διαρροών (*leakage*):

- Τα δεδομένα διαχωρίζονται σε K τμήματα (folds).
- Κάθε βασικό μοντέλο εκπαιδεύεται στα $K-1$ τμήματα και παράγει προβλέψεις αποκλειστικά για το αθέατο τμήμα. Η διαδικασία αυτή (**εκτός-τμήματος προβλέψεις / *out-of-fold – OOF–predictions***) εξασφαλίζει ότι το *meta-learner* εκπαιδεύεται στην πραγματική, γενικευμένη συμπεριφορά των βασικών μοντέλων και, όχι, στην υπερπροσαρμοσμένη απόδοσή τους.
- Στο τελικό στάδιο, οι OOF προβλέψεις λειτουργούν ως τα νέα χαρακτηριστικά εισόδου ($\mathbf{x}$) για το *meta-learner*, το οποίο αναλαμβάνει να εκδώσει την τελική πρόβλεψη.

Οι Van der Laan et al. [59] απέδειξαν μαθηματικά ότι ένα καλοσχεδιασμένο *stacking ensemble* (Super Learner) συγκλίνει ασυμπτωτικά στην απόδοση του καλύτερου δυνατού βασικού μοντέλου για το εκάστοτε πρόβλημα.

2.7.6 Σύνοψη και σύνδεση με τη μεθοδολογία της διπλωματικής εργασίας

Τα Δέντρα Διαβαθμισμένης Ενίσχυσης (*Gradient Boosted Trees* - GBTs) και τα σύνολα που βασίζονται στην τεχνική του συγκεντρωτικού ανασυνδυασμού (*bagging ensembles*) δεν αποτελούν απλώς ένα ακόμη εργαλείο Μηχανικής Μάθησης, αλλά τον απόλυτο αλγοριθμικό κορμό (*backbone*) της σύγχρονης βιομηχανικής πρακτικής στην πρόβλεψη δομημένων δεδομένων [47]. Στο υπολογιστικό πλαίσιο (*pipeline*) της παρούσας εργασίας (Κεφάλαιο 5), αξιοποιούνται με τρεις διακριτούς τρόπους:

- Ως **Ταξινομητές** (*Classifiers*) στο Πρώτο Στάδιο του μοντέλου Εμποδίου (Hurdle *model* - βλ. ενότητα 2.9.2) χρησιμοποιώντας τη συνάρτηση λογαριθμικής απώλειας (*log-loss*) για την εκτίμηση της πιθανότητας εκδήλωσης θετικής ζήτησης.
- Ως **Παλινδρομητές** (*Regressors*) στο Δεύτερο Στάδιο χρησιμοποιώντας τη συνάρτηση απώλειας τύπου *pinball* (Pinball Loss) για την απευθείας πρόβλεψη της κατανομής των ποσοστημορίων της ζήτησης ή την απόκλιση Tweedie (Tweedie deviance) για σύγκριση μονού σταδίου.
- Ως **Βασικά Μοντέλα** (*Base Learners*) σε ένα σύνολο Συσσωρευτικής Γενίκευσης (*stacking ensemble*, βλ. ενότητα 2.7.5).

Συγκεκριμένα, τα τέσσερα προεπιλεγμένα (*default*) μοντέλα εμποδίου του pipeline (`Hurdle_HGB`, `Hurdle_RF`, `Hurdle_ET`, `Hurdle_Cat`) αντιστοιχούν στις οικογένειες μοντέλων που αναλύθηκαν στην παρούσα ενότητα, HistGradientBoosting (ενότητα 2.7.4), Random Forests (ενότητα 2.7.2) και Extra

Trees (ενότητα 2.7.2) και CatBoost (ενότητα 2.7.4). Η παράλληλη χρήση τους δεν συνιστά αρχιτεκτονική επικάλυψη ή πλεονασμό, αλλά συνειδητή στρατηγική επιλογή διαφοροποίησης (*diversification*). Τα μοντέλα GBT (όπως τα HistGradientBoosting, XGBoost, LightGBM και CatBoost) παράγουν ασυσχέτιστα σφάλματα μεταξύ τους λόγω των αλγοριθμικών τους διαφορών που περιγράφηκαν νωρίτερα (π.χ. ομαδοποίηση ιστογράμματος, ανάπτυξη κατά επίπεδο έναντι φύλλου, διατεταγμένη ενίσχυση). Αντίστοιχα, τα μοντέλα που βασίζονται στο *bagging* (Random Forests και Extra Trees) εισάγουν διαφορετικού τύπου τυχαιότητα μέσω του *bootstrap* και της τυχαίας επιλογής χαρακτηριστικών ή κατωφλίων, η οποία είναι ορθογώνια ως προς τα σφάλματα των GBTs. Αυτή η ποικιλομορφία μεγιστοποιεί την αποδοτικότητα της τελικής Συσσωρευτικής Γενίκευσης από τα μοντέλα μετα-μάθησης (*meta-learner ensembles*).

Η σύνδεση με την Πιθανοκρατική Πρόβλεψη (που αναλύεται στην ενότητα 2.10) είναι επίσης άμεση. Τα μοντέλα GBT (και ειδικά το HistGradientBoosting) μπορούν να εκπαιδευτούν απευθείας με τη συνάρτηση *pinball loss* για τα επιθυμητά ψηλά ποσοστημόρια, ενώ τα Random Forest και Extra Trees παράγουν ποσοστημόρια έμμεσα, μέσω των εκατοστημορίων στο επίπεδο των μεμονωμένων δέντρων (*tree-level percentiles*).

2.8 Βαθιά Μάθηση (Deep Learning) και χρονοσειρές ζήτησης

Η Βαθιά Μάθηση (Deep Learning – DL) αξιοποιεί νευρωνικά δίκτυα πολλαπλών κρυφών επιπέδων για την εκμάθηση εξαιρετικά πολύπλοκων, μη γραμμικών σχέσεων απευθείας από τα ακατέργαστα δεδομένα ελαχιστοποιώντας την ανάγκη για χειροκίνητη μηχανική χαρακτηριστικών (*feature engineering*) [60]. Στο πεδίο των χρονοσειρών, οι αρχιτεκτονικές αυτές επιδιώκουν να συλλάβουν ταυτόχρονα βραχυπρόθεσμες και μακροπρόθεσμες χρονικές εξαρτήσεις, καθώς, και διασταυρούμενες συσχετίσεις μεταξύ χιλιάδων κωδικών [61].

Αν και η προτεινόμενη αρχιτεκτονική της παρούσας διπλωματικής (Κεφάλαιο 5) βασίζεται στα Δέντρα Ενίσχυσης (GBTs), η συνοπτική παρουσίαση των νευρωνικών δικτύων κρίνεται απαραίτητη για τη θεωρητική πληρότητα της μελέτης και την τεκμηρίωση των τελικών σχεδιαστικών επιλογών.

2.8.1 Αναδρομικά Νευρωνικά Δίκτυα (RNN, LSTM, GRU)

Τα Αναδρομικά Νευρωνικά Δίκτυα (RNN - *Recurrent Neural Networks*) διαφοροποιούνται από τα κλασικά δίκτυα εμπρόσθιας τροφοδότησης (*feed-forward*) μέσω της εισαγωγής μιας εσωτερικής (κρυφής) κατάστασης $\mathbf{h}_t$. Η κατάσταση αυτή λειτουργεί ως μια μορφή "μνήμης" και μεταφέρει πληροφορία από τα προηγούμενα χρονικά βήματα [62]:

$$\mathbf{h}_t = \tanh(\mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{W}_{xh}\mathbf{x}_t + \mathbf{b}_h) \tag{2.44}$$

Παρά τη θεωρητική τους ικανότητα, τα απλά RNNs υποφέρουν πρακτικά από το πρόβλημα της εξαφάνισης/έκρηξης της κλίσης (*vanishing/exploding gradients*) κατά την εκπαίδευση καθιστώντας αδύνατη τη μάθηση μοτίβων με μεγάλη χρονική υστέρηση.

Για την επίλυση αυτού του προβλήματος, αναπτύχθηκαν προηγμένες αρχιτεκτονικές, με κυριότερες τα LSTM (*Long Short-Term Memory*) [63] και τα GRU (*Gated Recurrent Units*) [64]. Αυτά τα δίκτυα ενσωματώνουν έναν εσωτερικό μηχανισμό "πυλών" (*gates*), οι οποίες ελέγχουν δυναμικά ποια πληροφορία πρέπει να διατηρηθεί στη μνήμη, ποια να διαγραφεί και ποια να ανανεωθεί από τα νέα δεδομένα.

Στο πλαίσιο της πρόβλεψης ζήτησης, η πλέον επιτυχημένη υλοποίηση αυτής της φιλοσοφίας είναι το DeepAR [46], ένα παγκόσμιο (*global*) LSTM μοντέλο που έχει σχεδιαστεί ρητά για να εξάγει

πιθανοκρατικές κατανομές αντί για σημειακές προβλέψεις εφαρμόζοντας διασταυρούμενη μάθηση (*cross-learning*) σε όλο τον κατάλογο των προϊόντων.

2.8.2 Συνελικτικά Δίκτυα (CNN) και Υβριδικά Σχήματα

Πέρα από τα αναδρομικά δίκτυα, τα Συνελικτικά Νευρωνικά Δίκτυα (CNN), τα οποία είναι ευρέως γνωστά από την επεξεργασία εικόνας, έχουν προσαρμοστεί επιτυχώς στις χρονοσειρές μέσω της χρήσης μονοδιάστατης συνέλιξης (*1D convolution*).

Η τομή σε αυτόν τον τομέα επήλθε με αρχιτεκτονικές όπως το WaveNet [65] και τα Χρονικά Συνελικτικά Δίκτυα (TCN - *Temporal Convolutional Networks*) [66]. Χρησιμοποιώντας αραιωμένες συνελίξεις (*dilated convolutions*), τα TCN μπορούν να παρατηρούν εκτεταμένα ιστορικά παράθυρα χωρίς δραματική αύξηση του υπολογιστικού κόστους ξεπερνώντας συχνά τα LSTM σε ταχύτητα εκπαίδευσης και ακρίβεια.

Τα πλέον σύγχρονα βιομηχανικά συστήματα συχνά καταφεύγουν σε υβριδικά σχήματα. Χαρακτηριστικό παράδειγμα αποτελεί ο Temporal Fusion Transformer (TFT) [67], ο οποίος συνδυάζει κωδικοποιητές LSTM για την ακολουθιακή δομή, με μηχανισμούς προσοχής (*attention mechanisms*), και εξειδικευμένες πύλες για τη διαχείριση στατικών χαρακτηριστικών (όπως η μάρκα ή η κατηγορία ενός είδους).

2.8.3 Σύγκριση με τα Δέντρα Ενίσχυσης και Επιλογή Μεθοδολογίας

Θεωρητικά, τα μοντέλα Βαθιάς Μάθησης προσφέρουν εντυπωσιακά πλεονεκτήματα, όπως αυτόματη εξαγωγή χαρακτηριστικών, φυσική διαχείριση ακολουθιών μεταβλητού μήκους και εγγενή ικανότητα εξαγωγής πιθανοκρατικών κατανομών (όπως απαιτείται για τις αποθήκες).

Ωστόσο, η παρούσα διπλωματική εργασία υιοθετεί ρητά την αρχιτεκτονική των Δέντρων Διαβαθμισμένης Ενίσχυσης (GBTs της ενότητας 2.7) ως τον κεντρικό πυρήνα του προγνωστικού συστήματος. Η απόφαση αυτή εδράζεται στα συμπεράσματα του παγκόσμιου διαγωνισμού M5 [21] και της σύγχρονης βιομηχανικής πρακτικής, όπου τα GBTs αποδείχθηκαν συστηματικά ανώτερα των Νευρωνικών Δικτύων σε δεδομένα λιανικής και διαλείπουσας ζήτησης για τους εξής λόγους:

- **Απαιτήσεις Δεδομένων:** Τα βαθιά δίκτυα απαιτούν τεράστιους όγκους δεδομένων, ώστε να αποφύγουν την υπερπροσαρμογή. Στη διαλείπουσα ζήτηση ανταλλακτικών, ο κύριος όγκος του καταλόγου αποτελείται από αραιές (*sparse*) χρονοσειρές, συνθήκη στην οποία τα Νευρωνικά Δίκτυα συχνά καταρρέουν.
- **Ανθεκτικότητα:** Σε περιβάλλοντα δομημένων δεδομένων (*tabular data*) με έντονο στατιστικό θόρυβο, ισχυρή ετεροσκεδαστικότητα και πληθώρα στατικών κατηγορικών μεταβλητών, τα GBTs (όταν συνδυάζονται με στιβαρό *feature engineering*) επιδεικνύουν σαφώς μεγαλύτερη σταθερότητα [61].
- **Υπολογιστικό Κόστος και Ρύθμιση:** Τα GBTs συγκλίνουν ταχύτερα, απαιτούν κλασματικό χρόνο εκπαίδευσης σε σχέση με τα DL μοντέλα και είναι λιγότερο ευαίσθητα στη μικρο-ρύθμιση υπερπαραμέτρων.

Συμπερασματικά, στο υπολογιστικό πλαίσιο (*pipeline*) του Κεφαλαίου 5, οι αρχιτεκτονικές Βαθιάς Μάθησης θεωρούνται πολύτιμες κυρίως ως δυνητικά μέλη ενός ευρύτερου ensemble ή ως εργαλεία παραγωγής διανυσματικών αναπαραστάσεων (embeddings), ενώ το βάρος της πρόβλεψης και λήψης αποφάσεων ανατίθεται στα στατιστικά ισχυρότερα (για το συγκεκριμένο πρόβλημα) δενδρικά μοντέλα.

2.9 Δι-σταδιακά μοντέλα Εμποδίου (Hurdle *models*) και μοντέλα Διογκωμένου Μηδενός (*Zero-Inflated models*)

Στη διαλείπουσα ζήτηση ανταλλακτικών, η συντριπτική πλειονότητα των ημερήσιων παρατηρήσεων ανά κωδικό είδους (SKU) είναι μηδενική. Στο σύνολο δεδομένων της παρούσας εργασίας, για παράδειγμα, πάνω από το 95% των συνδυασμών (SKU, t) παρουσιάζει $y_t = 0$ (βλ. ενότητα 4.3). Σε τέτοιες συνθήκες, η εφαρμογή τυπικών εκτιμητών παλινδρόμησης (π.χ. Ελάχιστα Τετράγωνα ή κλασικό Γενικευμένο Γραμμικό Μοντέλο Poisson – GLM) οδηγεί συστηματικά σε υπερβολικά εκτιμημένη μηδενική μάζα και συμπιεσμένες, υποεκτιμημένες θετικές προβλέψεις [68]. Η αντιμετώπιση αυτού του φαινομένου, γνωστού ως διόγκωση μηδενικών (*zero-inflation*) ή πλεονάζοντα μηδενικά (*excess zeros*), απαιτεί μια ειδική κλάση μοντέλων που διαχειρίζονται *ξεχωριστά* την εμφάνιση της θετικής ζήτησης από το μέγεθός της. Η παρούσα ενότητα παρουσιάζει τις τρεις βασικές οικογένειες τέτοιων μοντέλων που χρησιμοποιούνται στην υπολογιστική ροή (*pipeline*) του Κεφαλαίου 5. Αυτές είναι τα δι-σταδιακά μοντέλα Εμποδίου (Hurdle *models*), τα μοντέλα διογκωμένου μηδενός (*zero-inflated models*) και η παλινδρόμηση Tweedie (*Tweedie regression*), καθώς, και η κλασική τεχνική του λογαριθμικού μετασχηματισμού (*log-transformation*) με τη διόρθωση Duan.

2.9.1 Το πρόβλημα της διόγκωσης μηδενικών (*Zero-inflation*)

Έστω μια χρονοσειρά ζήτησης $\{y_t\}_{t=1}^{T}$ ενός κωδικού (SKU), με $y_t \in \mathbb{Z}_{\geq 0}$. Ορίζουμε τη συνάρτηση δείκτη εμφάνισης $\mathbb{1}_{y_t>0}$ και τη δεσμευμένη υπό-σειρά των θετικών τιμών $\{y_t : y_t > 0\}$. Η εμπειρική κατανομή της y_t σε συνθήκες διαλείπουσας ζήτησης αναλύεται φυσικά σε δύο συστατικά:

$$f_Y(y) = \pi \cdot \delta_0(y) + (1 - \pi) \cdot f_{Y^+}(y), \qquad y \in \mathbb{Z}_{\geq 0}, \tag{2.45}$$

όπου $\pi = \mathbb{P}(y = 0)$, $\delta_0(\cdot)$ η συνάρτηση μάζας στο μηδέν και f_{Y^+} μια κατανομή ορισμένη στους θετικούς ακέραιους [68]. Η βασική παρατήρηση είναι ότι μια κλασική κατανομή μέτρησης, όπως η Poisson, με μέσο λ, η οποία εκτιμάται από όλο το δείγμα, αναγκάζεται να αναπαράγει ταυτόχρονα και την υπερβολική μηδενική μάζα και το θετικό σήμα μέσω της ίδιας παραμέτρου λ. Αυτό οδηγεί σε υποδιασπορά (*underdispersion*) στις θετικές τιμές και σε υπερδιασπορά (*overdispersion*) στο σύνολο [69], [70]. Ισοδύναμα, η υπόθεση ισότητας διακύμανσης-μέσης τιμής της Poisson παραβιάζεται σαφώς σε δεδομένα διαλείπουσας ζήτησης ($\mathrm{Var}(y) \gg \mathbb{E}[y]$), γεγονός που αναγνωρίζεται από τους τυπικούς ελέγχους καλής προσαρμογής (*goodness-of-fit tests*) [68].

Η αντιμετώπιση του φαινομένου αυτού κινείται κυρίως σε δύο κατευθύνσεις: τα δι-σταδιακά μοντέλα Εμποδίου (Hurdle *models*), τα οποία αντιμετωπίζουν τα μηδενικά και τα θετικά μεγέθη ως *διαφορετικές διαδικασίες*, και τα μοντέλα διογκωμένου μηδενός (*zero-inflated models*), τα οποία τα αντιμετωπίζουν ως μίγμα (*mixture*) δύο διαφορετικών διαδικασιών παραγωγής μηδενικών [69], [70].

2.9.2 Δι-σταδιακά μοντέλα Εμποδίου (Hurdle *models*)

Η ιδέα του μοντέλου Εμποδίου (Hurdle *model*), η οποία εισήχθη από τον Cragg [71] για συνεχείς μη αρνητικές μεταβλητές και επεκτάθηκε από τον Mullahy [70] σε δεδομένα μέτρησης, βασίζεται στη διάσπαση της ζήτησης σε δύο ανεξάρτητες αποφάσεις: (α) εμφανίζεται ή όχι θετική ζήτηση στην περίοδο (*occurrence*), και (β) δεδομένου ότι εμφανίζεται, ποιο είναι το μέγεθός της (*size*). Η αναμενόμενη ζήτηση γράφεται ως:

$$\mathbb{E}[y_t \mid \mathbf{x}_t] = \underbrace{\mathbb{P}(y_t > 0 \mid \mathbf{x}_t)}_{\text{Stage 1: occurrence}} \cdot \underbrace{\mathbb{E}[y_t \mid y_t > 0, \mathbf{x}_t]}_{\text{Stage 2: positive size}} \tag{2.46}$$

Η πιθανότητα στο πρώτο στάδιο μοντελοποιείται συνήθως μέσω λογιστικής παλινδρόμησης (*logistic regression*) ή οποιουδήποτε ταξινομητή παράγει βαθμονομημένες πιθανότητες:

$$p_t = \mathbb{P}(y_t > 0 \mid \mathbf{x}_t) = \sigma\Big(g_\phi(\mathbf{x}_t)\Big), \qquad \sigma(u) = \frac{1}{1 + e^{-u}}, \tag{2.47}$$

όπου g_ϕ μπορεί να είναι ένα γραμμικό μοντέλο, ένα δέντρο ενίσχυσης ή ένα νευρωνικό δίκτυο με κατάλληλη βαθμονόμηση (βλ. ενότητα 2.10). Η αναμενόμενη ποσότητα υπό συνθήκη θετικής ζήτησης μοντελοποιείται από έναν παλινδρομητή m_ψ εκπαιδευμένο *μόνο στις θετικές παρατηρήσεις*:

$$\mu_t^+ = \mathbb{E}[y_t \mid y_t > 0, \mathbf{x}_t] = m_\psi(\mathbf{x}_t) \tag{2.48}$$

Η συνολική πρόβλεψη συντίθεται ως

$$\hat{y}_t = p_t \cdot \mu_t^+ \tag{2.49}$$

Η κρίσιμη ιδιότητα της διατύπωσης εμποδίου είναι ότι η συνάρτηση πιθανοφάνειας παραγοντοποιείται πλήρως (*factorizes completely*):

$$\mathcal{L}(\phi, \psi) \;=\; \prod_{t:\, y_t = 0} (1 - p_t) \;\cdot\; \prod_{t:\, y_t > 0} p_t \cdot f_{Y^+}(y_t \mid \mathbf{x}_t;\, \psi) \tag{2.50}$$

Επομένως, οι παράμετροι ϕ (Στάδιο 1) και ψ (Στάδιο 2) μπορούν να εκτιμηθούν *ξεχωριστά* και *χωρίς απώλεια αποδοτικότητας*, υπό την υπόθεση ότι τα δύο στάδια είναι παραμετρικά ανεξάρτητα [68], [70]. Αυτό αποτελεί το θεμελιώδες πρακτικό πλεονέκτημα του Hurdle έναντι των μοντέλων τύπου *zero-inflated*. Δηλαδή, επιτρέπει την επιλογή εντελώς διαφορετικών οικογενειών μοντέλων για τις δύο διαδικασίες (π.χ. λογιστική παλινδρόμηση συνδυαστικά με *gradient boosting* στο μέγεθος), όπως η προσέγγιση που υιοθετείται στη μεθοδολογία της παρούσας διπλωματικής (βλ. ενότητα 5.5).

Από στατιστικής σκοπιάς, το μοντέλο Hurdle ερμηνεύει όλα τα μηδενικά ως προερχόμενα από μία και μόνη πηγή (τη διαδικασία μη-εμφάνισης, *non-occurrence*) και θεωρεί ότι, μόλις «ξεπεραστεί» το εμπόδιο (*hurdle*), η θετική ζήτηση διέπεται από μια αποκομμένη στο μηδέν κατανομή (*zero-truncated distribution*) [70]. Συνήθεις επιλογές για το δεύτερο στάδιο περιλαμβάνουν την αποκομμένη στο μηδέν Poisson (*zero-truncated Poisson*), την αποκομμένη Αρνητική Διωνυμική (*zero-truncated Negative Binomial*) και, σε προβλήματα συνεχών τιμών (*continuous settings*), κατανομές *log-normal* ή *Gamma*. Η σχέση των μοντέλων εμποδίου με τις παραδοσιακές μεθόδους Croston/TSB είναι άμεση. Η προσέγγιση Teunter–Syntetos–Babai (βλ. ενότητα 2.4.4) μπορεί να θεωρηθεί ως μια ειδική, παραμετρική περίπτωση *hurdle*, κατά την οποία το στάδιο εμφάνισης ανανεώνεται με εκθετική εξομάλυνση και το στάδιο μεγέθους αποτελεί μια σταθερή εκτίμηση επιπέδου [2], **Error! Reference source not found.**.

2.9.3 Μοντέλα Διογκωμένου Μηδενός (Zero-Inflated Models – ZIP, ZINB)

Σε αντίθεση με τη διατύπωση εμποδίου, τα μοντέλα διογκωμένου μηδενός (*zero-inflated models*), τα οποία εισήχθησαν από τη Lambert [69] για την ανίχνευση κατασκευαστικών ελαττωμάτων, προσεγγίζουν το πρόβλημα εννοιολογικά ως ένα μίγμα (*mixture*) δύο διαφορετικών διαδικασιών.

Για να γίνει κατανοητή η δομική τους διαφορά, στο μοντέλο Poisson Διογκωμένου Μηδενός (*Zero-Inflated Poisson* – ZIP) υποτίθεται ότι ένα μηδενικό αποτέλεσμα πωλήσεων μπορεί να προκύψει από δύο ξεχωριστές πηγές:

- **Δομικό μηδέν** (*Structural zero*): Με μια πιθανότητα π, η ζήτηση είναι εκ των προτέρων αδύνατη (π.χ. ένας κωδικός έχει αποσυρθεί προσωρινά, υπάρχει έλλειψη αποθέματος στην αγορά ή ο πελάτης δεν έχει καμία απολύτως ανάγκη).
- **Δειγματοληπτικό μηδέν** (*Sampling zero*): Με την υπόλοιπη πιθανότητα $1 - \pi$, το σύστημα συμπεριφέρεται «κανονικά» ακολουθώντας την κατανομή Poisson, η οποία, όμως καθαρά λόγω πιθανοτήτων, μπορεί να παραγάγει, επίσης, την τιμή μηδέν.

Αντί, λοιπόν, να καταφύγουμε σε πολύπλοκες συναρτήσεις πυκνότητας, η ουσία του μοντέλου συνοψίζεται στο ότι η συνολική πιθανότητα να παρατηρήσουμε μηδενική ζήτηση είναι απλώς το άθροισμα αυτών των δύο ενδεχομένων:

$$\mathbb{P}(Y = 0) = \pi + (1 - \pi) \cdot e^{-\lambda} \tag{2.51}$$

όπου λ είναι η παράμετρος ρυθμού της κατανομής Poisson [69]. Αντίστοιχα, το μοντέλο Αρνητικής Διωνυμικής Διογκωμένου Μηδενός (*Zero-Inflated Negative Binomial* – ZINB) ακολουθεί ακριβώς την ίδια λογική αντικαθιστώντας απλώς την κατανομή Poisson με την Αρνητική Διωνυμική, ώστε να διαχειριστεί αποτελεσματικότερα την υπερδιασπορά των θετικών παρατηρήσεων [68].

Η θεμελιώδης ποιοτική διαφορά αυτών των μοντέλων από τα μοντέλα Εμποδίου (Hurdle *models*) έγκειται στην ακόλουθη παρατήρηση. Τα ZIP/ZINB διατηρούν θεωρητικά τη δυνατότητα παραγωγής μηδενικών και από τη «θετική» (δεύτερη) διαδικασία. Σε συγκεκριμένα σενάρια, όπως η ύπαρξη ενός λανθάνοντος μηχανισμού που αποκλείει τη ζήτηση, αυτή η ερμηνεία μπορεί να έχει θεωρητικό πλεονέκτημα. Ωστόσο, στην πράξη η εκτίμησή τους είναι ιδιαίτερα πολύπλοκη. Επειδή η συνάρτηση πιθανοφάνειας περιέχει αυτό το μαθηματικό «μίγμα», δεν μπορεί να παραγοντοποιηθεί και να διαχωριστεί σε δύο ανεξάρτητα προβλήματα. Απαιτείται κοινή βελτιστοποίηση παραμέτρων (συχνά μέσω του αλγορίθμου ΕΜ), γεγονός που καθιστά την εφαρμογή τους εξαιρετικά δύσκολη, όταν επιδιώκουμε να τα συνδυάσουμε με προηγμένους αλγορίθμους δέντρων ενίσχυσης (*gradient boosted trees*).

Για τον λόγο αυτό, στην πρόβλεψη ζήτησης μεγάλης κλίμακας με χρήση Μηχανικής Μάθησης, η διατύπωση Hurdle είναι σαφώς προτιμητέα, γιατί επιτρέπει την πλήρη ανεξαρτησία στην εκπαίδευση των δύο σταδίων, είναι αλγοριθμικά ταχύτερη και η ερμηνεία της –*«πρώτα αν, μετά πόσο»*– ευθυγραμμίζεται απολύτως με τη λογική της λήψης αποφάσεων αναπλήρωσης [72]. Αυτή αποτελεί και τη στρατηγική επιλογή της παρούσας εργασίας.

2.9.4 Παλινδρόμηση Tweedie (*Tweedie Regression*)

Μια εναλλακτική, ενιαία προσέγγιση που αντιμετωπίζει τη διόγκωση των μηδενικών χωρίς ρητή διάσπαση σε δύο στάδια είναι η παλινδρόμηση Tweedie (Tweedie *regression*). Η κατανομή Tweedie αποτελεί μια ευέλικτη μαθηματική οικογένεια, η οποία, αναλόγως της τιμής μιας παραμέτρου p, μπορεί να συμπεριφερθεί ως Κανονική, Poisson ή Gamma κατανομή [73], [74].

Για την περίπτωση της διαλείπουσας ζήτησης, η πλέον ενδιαφέρουσα περιοχή προκύπτει, όταν $p \in (1,2)$. Σε αυτό το διάστημα, η κατανομή λειτουργεί πρακτικά ως μια σύνθετη κατανομή Poisson-Gamma (*compound Poisson-Gamma distribution*). Διαισθητικά, αυτό σημαίνει ότι το μοντέλο συνενώνει ταυτόχρονα δύο διαδικασίες, τη συχνότητα άφιξης των πελατών (που ακολουθεί κατανομή Poisson) και το μέγεθος της παραγγελίας του κάθε πελάτη (που ακολουθεί κατανομή Gamma).

Το τεράστιο θεωρητικό της πλεονέκτημα είναι ότι έχει θετική μάζα πιθανότητας στο μηδέν (επιτρέποντας δηλαδή προβλέψεις για περιόδους χωρίς απολύτως καμία πώληση) ενώ, ταυτοχρόνως, μπορεί να παράγει συνεχείς θετικές τιμές στον θετικό ημιάξονα [75]. Αυτό προσφέρει μια *ενιαία*

περιγραφή ολόκληρης της ζήτησης καθιστώντας την εξαιρετικά ελκυστική για συστήματα Μηχανικής Μάθησης που την υποστηρίζουν εγγενώς. Αλγόριθμοι όπως το `XGBoost`, το `LightGBM` και το `HistGradientBoosting` της βιβλιοθήκης `scikit-learn` διαθέτουν ενσωματωμένη συνάρτηση κόστους `Tweedie loss` επιτρέποντας την απευθείας πρόβλεψη της διαλείπουσας ζήτησης χωρίς να απαιτείται η δημιουργία πολύπλοκης δι-σταδιακής αρχιτεκτονικής **Error! Reference source not found.**.

Σε σύγκριση με τα μοντέλα Εμποδίου (Hurdle *models*), η παλινδρόμηση Tweedie είναι εννοιολογικά παράλληλη και αλγοριθμικά απλούστερη (λόγω του ενός σταδίου). Ωστόσο, είναι παραμετρικά πιο περιοριστική, καθώς συνδέει άρρηκτα την πιθανότητα εμφάνισης και το τελικό μέγεθος της ζήτησης μέσω της ίδιας κοινής μέσης τιμής μ.

2.9.5 Λογαριθμικός μετασχηματισμός και η διόρθωση Duan

Μια καθιερωμένη πρακτική για την αντιμετώπιση δεδομένων με έντονη δεξιά ασυμμετρία και βαριές ουρές (*heavy-tailed counts*) είναι ο λογαριθμικός μετασχηματισμός (*log-transformation*). Επειδή η διαλείπουσα ζήτηση περιέχει μηδενικά, εφαρμόζεται συνήθως η συνάρτηση `log1p` [$log(1+y)$]. Η προσέγγιση αυτή επιτρέπει στους αλγορίθμους, όπως οι παλινδρομητές του δεύτερου σταδίου στα μοντέλα Hurdle, να εκπαιδευτούν σε ένα σαφώς πιο «ομαλό» μαθηματικό περιβάλλον ικανοποιώντας προσεγγιστικά τις κλασικές υποθέσεις της ομοσκεδαστικότητας.

Το θεωρητικό πρόβλημα ανακύπτει κατά την επιστροφή των προβλέψεων στον πραγματικό χώρο της ζήτησης μέσω του αντίστροφου μετασχηματισμού της εκθετικής συνάρτησης (`expm1`). Εξαιτίας της αυστηρά κυρτής μορφής της εκθετικής καμπύλης (ένα φαινόμενο που περιγράφεται μαθηματικά από την Ανισότητα Jensen), η απλή εκθετική μετατροπή της πρόβλεψης είναι συστηματικά μεροληπτική (*biased*). Πρακτικά, αυτό σημαίνει ότι η αντίστροφη πρόβλεψη υπολείπεται σταθερά της πραγματικής αναμενόμενης μέσης ζήτησης.

Για την αποκατάσταση αυτής της μεροληψίας, η βιβλιογραφία προτείνει ειδικούς συντελεστές εξομάλυνσης (*smearing factors*), με πλέον γνωστό τον μη παραμετρικό εκτιμητή του Duan [77]. Ωστόσο, μια υπολογιστικά αποδοτική και ευρέως διαδεδομένη πρακτική, η οποία υιοθετείται και στην παρούσα εργασία, είναι η παραμετρική λογαριθμοκανονική (*log-normal*) προσέγγιση. Υποθέτοντας στατιστική ισοδυναμία υπό την παραδοχή της κανονικότητας των σφαλμάτων στον λογαριθμικό χώρο, ο διορθωτικός πολλαπλασιαστικός παράγοντας μπορεί να υπολογιστεί απευθείας από τη διακύμανση (*variance*) των καταλοίπων (*residuals*) που άφησε το μοντέλο κατά την εκπαίδευσή του.

Πολλαπλασιάζοντας την αρχική εκθετική πρόβλεψη με αυτόν τον παράγοντα, το μοντέλο ανακτά την πραγματική κλίμακα της μέσης τιμής. Στα δεδομένα ανταλλακτικών, ο διορθωτικός αυτός παράγοντας κυμαίνεται συνήθως μεταξύ 1,05 και 1,30. Η παράλειψή του θα οδηγούσε σε μια επικίνδυνη, συστηματική υπο-εκτίμηση του απαιτούμενου μεγέθους των αποθεμάτων ασφαλείας. Στην υπολογιστική ροή της παρούσας εργασίας, η διόρθωση εξομάλυνσης ενσωματώνεται ρητά για όλα τα μοντέλα που εκπαιδεύονται σε λογαριθμικό χώρο (βλ. ενότητα 5.6.6.1).

2.9.6 Σύνοψη και σύνδεση με τη μεθοδολογία

Συνοψίζοντας, οι κύριες θεωρητικές προσεγγίσεις για τη διαχείριση της διόγκωσης μηδενικών σε δεδομένα διαλείπουσας ζήτησης καταγράφονται στον Πίνακα 2.3.

Πίνακας 2.3: Σύγκριση προσεγγίσεων διαχείρισης του προβλήματος Διογκωμένου Μηδενός (*Zero-Inflation*)

Προσέγγιση	**Παραγωγή μηδενικών**	**Παραγοντοποίηση**	**Ευελιξία ανά στάδιο**
Δι-σταδιακά μοντέλα Εμποδίου (Hurdle)	Αποκλειστικά από τη μη-εμφάνιση (*non-occurrence*)	Ναι	Πλήρης (πλήρης ανεξαρτησία)
Μοντέλα Διογκωμένου Μηδενός (ZIP/ZINB)	Από δύο πηγές (*structural* + *sampling*)	Όχι (απαιτεί κοινή εκτίμηση MLE)	Περιορισμένη
Tweedie	Έμμεσα, μέσω σύνθετης *Poisson-Gamma*	Δεν εφαρμόζεται	Καμία
Λογαριθμικός μετασχηματισμός + Διόρθωση εξομάλυνσης	Έμμεσα, μέσω αντίστροφου μετασχηματισμού	Δεν εφαρμόζεται	Καμία

Στη μεθοδολογία του Κεφαλαίου 5 αξιοποιούνται και οι τέσσερις προσεγγίσεις. Ωστόσο, ως *κεντρική* αρχιτεκτονική επιλέγεται ρητά η διάσπαση εμποδίου (*hurdle*). Με άλλα λόγια, χρησιμοποιείται ένας ταξινομητής στο πρώτο στάδιο για την πιθανότητα εμφάνισης ζήτησης και ένας ανεξάρτητος παλινδρομητής στο δεύτερο στάδιο για τον καθορισμό του μεγέθους, δεδομένης της εμφάνισης. Η παλινδρόμηση Tweedie και ο λογαριθμικός μετασχηματισμός με παραμετρική διόρθωση εξομάλυνσης (*smearing correction*) χρησιμοποιούνται κυρίως ως αυτόνομα μοντέλα αναφοράς ενός σταδίου (*baseline single-stage models*). Αντιθέτως, οι αρχιτεκτονικές ZIP/ZINB παραμένουν εκτός του πειραματικού πλαισίου εξαιτίας της εξαιρετικής δυσκολίας που παρουσιάζει η κοινή τους εκτίμηση, όταν εφαρμόζονται σε δομές με Δέντρα Διαβαθμισμένης Ενίσχυσης (*Gradient Boosted Trees* – GBTs). Η αναλυτική περιγραφή της εφαρμογής της αρχιτεκτονικής *hurdle* εντός του συστήματος `Sparse True Forecast Pipeline`, καθώς και οι ακριβείς επιλογές βαθμονόμησης και σύνθεσης των δύο σταδίων, τεκμηριώνονται αναλυτικά στην ενότητα 5.6.

2.10 Πιθανοκρατική πρόβλεψη και διαστήματα εμπιστοσύνης

Στις περισσότερες πραγματικές αποφάσεις αναπλήρωσης αποθεμάτων, η βέλτιστη ποσότητα παραγγελίας δεν εξαρτάται από τη μέση τιμή της ζήτησης, αλλά από συγκεκριμένα ποσοστημόρια (*quantiles*), π.χ. το 90ο ποσοστημόριο της ζήτησης κατά τον χρόνο προμήθειας καθορίζει το απόθεμα ασφαλείας για επιθυμητό επίπεδο εξυπηρέτησης 90% [11]. Επομένως, τα μοντέλα που παράγουν μόνο σημειακές προβλέψεις (*point forecasts*) δεν επαρκούν επιχειρησιακά καθιστώντας αναγκαία την παραγωγή διαστημάτων πρόβλεψης (*prediction intervals*), τα οποία ποσοτικοποιούν εμπειρικά την αβεβαιότητα της μελλοντικής ζήτησης.

2.10.1 Διαστήματα Πρόβλεψης μέσω Εμπειρικών Καταλοίπων

Μολονότι η θεωρία της Μηχανικής Μάθησης προτείνει συχνά την εκπαίδευση ανεξάρτητων μοντέλων παλινδρόμησης ποσοστημορίων (*quantile regressors*), η προσέγγιση αυτή είναι εξαιρετικά κοστοβόρα υπολογιστικά για καταλόγους με δεκάδες χιλιάδες κωδικούς ειδών, καθώς απαιτεί την εκπαίδευση ξεχωριστού μοντέλου για κάθε ζητούμενο ποσοστημόριο.

Μια πιο ευέλικτη και υπολογιστικά ελαφριά εναλλακτική, η οποία αντλεί τη θεωρητική της βάση από τις αρχές της Σύμμορφης Πρόβλεψης Διαχωρισμού (*Split Conformal Prediction*) [78], βασίζεται στα εμπειρικά κατάλοιπα εκτός-δείγματος (*out-of-sample residuals*). Σύμφωνα με αυτή τη μέθοδο, καταγράφονται τα ιστορικά σφάλματα $r_i = y_i - \hat{y}_i$ που έκανε το μοντέλο σε ένα σύνολο επικύρωσης (ή κατά τη διάρκεια του αναδρομικού ελέγχου / *backtesting*).

Στη συνέχεια, υπολογίζονται τα αντίστοιχα εμπειρικά ποσοστημόρια (π.χ. $Q_{0.10}$ και $Q_{0.90}$) αυτών των καταλοίπων. Το τελικό διάστημα πρόβλεψης $[\hat{y}_{low}, \hat{y}_{high}]$ δομείται προσθέτοντας αυτά τα ιστορικά ποσοστημόρια σφάλματος στη νέα, μελλοντική σημειακή πρόβλεψη. Όταν η διαδικασία αυτή εκτελείται διακριτά ανά στατιστική κατηγορία προϊόντων (π.χ. βάσει του πλαισίου Syntetos-Boylan), διασφαλίζεται ότι το εύρος του διαστήματος προσαρμόζεται δυναμικά στον πραγματικό βαθμό αβεβαιότητας και σποραδικότητας του εκάστοτε κωδικού (όπως υλοποιείται στην ενότητα 5.6.12 του προτεινόμενου συστήματος).

2.10.2 Ισοτονική παλινδρόμηση και ο αλγόριθμος PAVA

Όταν τα συστήματα Μηχανικής Μάθησης εκπαιδεύονται για να παράγουν προβλέψεις σε πολλαπλούς, αθροιστικούς χρονικούς ορίζοντες (π.χ. 15, 30 και 60 ημέρες), κάθε ορίζοντας μοντελοποιείται τυπικά ως ένα ανεξάρτητο πρόβλημα παλινδρόμησης. Αυτό δημιουργεί τον κίνδυνο παραβίασης της χρονικής μονοτονίας. Εξ' ορισμού, η αθροιστική πρόβλεψη των 30 ημερών είναι μαθηματικά αδύνατον να είναι μικρότερη από την πρόβλεψη των 15 ημερών, παρότι οι ανεξάρτητοι αλγόριθμοι ενδέχεται να παραγάγουν ένα τέτοιο αποτέλεσμα εξαιτίας στατιστικού θορύβου.

Η καθιερωμένη θεωρητική λύση για την εκ των υστέρων (*post-hoc*) αποκατάσταση της μονοτονίας είναι η ισοτονική παλινδρόμηση (*isotonic regression*). Η πλέον αποδοτική αλγοριθμική της υλοποίηση είναι ο αλγόριθμος PAVA (*Pool Adjacent Violators Algorithm*) [79], [80]. Ο αλγόριθμος PAVA σαρώνει διαδοχικά τις τιμές και, μόλις εντοπίσει παραβίαση της μονοτονίας, αντικαθιστά τις προβληματικές τιμές με τον σταθμισμένο μέσο όρο τους. Το μαθηματικό του πλεονέκτημα είναι ότι επιλύει το πρόβλημα σε γραμμικό χρόνο και προσφέρει τη βέλτιστη λύση ελαχίστων τετραγώνων (L2-*optimal*) εξασφαλίζοντας ότι η τελική σειρά των προβλέψεων είναι αυστηρά μη-φθίνουσα. Παράλληλα, ελαχιστοποιεί την απόκλιση από τις αρχικές εκτιμήσεις του μοντέλου. Στο προτεινόμενο υπολογιστικό σύστημα (ενότητα 5.6.6.5), ο αλγόριθμος εφαρμόζεται σε όλα τα μοντέλα πριν από την τελική αποθήκευσή τους διασφαλίζοντας την απόλυτη χρονική τους συνέπεια.

2.10.3 Μετρικές Αξιολόγησης Διαστημάτων Πρόβλεψης

Η αξιολόγηση της ποιότητας των διαστημάτων πρόβλεψης δεν υπολογίζεται με τις κλασικές μετρικές σημειακού σφάλματος (όπως η MASE ή η WAPE), αλλά βασίζεται σε δύο εξειδικευμένους δείκτες:

- **Εμπειρική Κάλυψη Διαστήματος** (*Prediction Interval Coverage* - `PI_Coverage`): Εκφράζει το ποσοστό των περιπτώσεων, κατά τις οποίες η πραγματική μελλοντική ζήτηση «έπεσε», πράγματι μέσα στα όρια του εκτιμώμενου διαστήματος. Ο επιχειρησιακός στόχος είναι η εμπειρική κάλυψη να προσεγγίζει όσο το δυνατόν περισσότερο το ονομαστικό επίπεδο εμπιστοσύνης (π.χ. 80%).
- **Κανονικοποιημένο Εύρος Διαστήματος** (*Prediction Interval Width* - `PI_Width`): Μετρά τη σχετική απόσταση μεταξύ του άνω και του κάτω ορίου του διαστήματος.

Ο ιδανικός συνδυασμός σε ένα πιθανοκρατικό σύστημα πρόβλεψης είναι η επίτευξη της επιθυμητής κάλυψης (αξιοπιστία) με το μικρότερο δυνατό εύρος (οξύτητα/ακρίβεια), καθώς ένα υπερβολικά ευρύ

διάστημα, παρότι τυπικά σωστό, δεν προσφέρει καμία ουσιαστική υποστήριξη στον προγραμματισμό των αγορών της επιχείρησης.

2.11 Υπολογιστική δεξαμενής και Διανυσματικές Αναπαραστάσεις (Embeddings)

Κατά την ανάπτυξη σύγχρονων συστημάτων πρόβλεψης, η έρευνα συχνά στρέφεται σε προηγμένες αρχιτεκτονικές Νευρωνικών Δικτύων για την άντληση πολύπλοκων μοτίβων ή την εκμετάλλευση μη δομημένων δεδομένων. Στο πλαίσιο της παρούσας εργασίας, εξετάστηκαν δύο τέτοιες σύγχρονες προσεγγίσεις πριν από την οριστικοποίηση της τελικής μεθοδολογίας. το Reservoir Computing για τη δυναμική μοντελοποίηση της χρονοσειράς και τα μοντέλα Transformers για την εξαγωγή σημασιολογικών χαρακτηριστικών από μεταδεδομένα ειδών.

2.11.1 Υπολογιστική Δεξαμενής (Reservoir Computing) και Δίκτυα Κατάστασης Ηχούς (ESN)

Η Υπολογιστική Δεξαμενής (Reservoir Computing – RC) αποτελεί μια υποκατηγορία των Αναδρομικών Νευρωνικών Δικτύων (RNN), η οποία εισήχθη από τον Jaeger [81] με την αρχιτεκτονική των Δικτύων Κατάστασης Ηχούς (Echo State Networks – ESN). Η κεντρική καινοτομία των δικτύων αυτών ανατρέπει την κλασική, υπολογιστικά βαρύτατη εκπαίδευση των RNNs (μέσω του αλγορίθμου οπισθοδιάδοσης στον χρόνο – BPTT). Αντί να εκπαιδεύονται όλα τα βάρη, το δίκτυο χρησιμοποιεί μια τυχαία, σταθερή και αραιή «δεξαμενή» κρυφών νευρώνων (το *reservoir*), η οποία προβάλει τα δεδομένα εισόδου σε έναν μη-γραμμικό χώρο υψηλής διάστασης.

Το μοναδικό τμήμα του δικτύου που εκπαιδεύεται είναι το τελικό, γραμμικό επίπεδο εξόδου (*readout layer*), συχνά με μια απλή και ταχύτατη λύση κλειστής μορφής (όπως η παλινδρόμηση κορυφογραμμής – *ridge regression*). Για να λειτουργήσει το δίκτυο, πρέπει να ικανοποιείται η «ιδιότητα της ηχούς» (*echo state property*), η οποία διασφαλίζει ότι η τρέχουσα κατάσταση της δεξαμενής εξαρτάται μοναδικά από το πρόσφατο ιστορικό των εισόδων και σταδιακά «ξεχνά» τις αρχικές συνθήκες [81].

Στη βιβλιογραφία της πρόβλεψης ζήτησης, τα ESNs έχουν συχνά δοκιμαστεί για την ικανότητά τους να μοντελοποιούν μη-γραμμικές δυναμικές αποφεύγοντας, ταυτόχρονα, τα προβλήματα εξαφανιζόμενης κλίσης (*vanishing gradients*) που μαστίζουν τα βαθιά δίκτυα [82]. Μια σύγχρονη ερευνητική κατεύθυνση δεν αντιμετωπίζει τα ESNs ως αυτόνομους εκτιμητές, αλλά τα χρησιμοποιεί ως δυναμικούς εξορυκτές χαρακτηριστικών (*feature extractors*) τροφοδοτώντας την έξοδο της δεξαμενής ως επιπλέον είσοδο (*feature*) σε ισχυρότερα αλγοριθμικά σχήματα, όπως τα δέντρα ενίσχυσης. Η πρακτική εφαρμογή και αξιολόγηση αυτής ακριβώς της αρχιτεκτονικής εντός του υπολογιστικού συστήματος της παρούσας εργασίας εξετάζεται αναλυτικά στο Κεφάλαιο 5.

2.11.2 Γλωσσικά Μοντέλα και Εξαγωγή Χαρακτηριστικών (Transformers)

Μια δεύτερη, εξαιρετικά διαδεδομένη σύγχρονη πρακτική είναι η αξιοποίηση Μεγάλων Γλωσσικών Μοντέλων (*Large Language Models*) βασισμένων στην αρχιτεκτονική Transformer [83]. Ο πυρήνας των Transformers είναι ο μηχανισμός αυτό-προσοχής (*self-attention*), ο οποίος επιτρέπει στο μοντέλο να αντιλαμβάνεται τη σημασιολογική εξάρτηση μεταξύ λέξεων ή συμβόλων (*tokens*), ανεξάρτητα από τη χρονική τους απόσταση.

Στον τομέα της διαχείρισης αποθεμάτων μεγάλης κλίμακας, τα συστήματα συχνά διαθέτουν πλούσια αλλά αδόμητα μεταδεδομένα (όπως περιγραφές προϊόντων, κωδικούς εναλλακτικότητας, εγχειρίδια). Τα προ-εκπαιδευμένα μοντέλα γλώσσας (π.χ. αρχιτεκτονικές τύπου `BERT` ή `MiniLM`) μπορούν να αναλύσουν αυτά τα κείμενα και να τα μετατρέψουν σε πυκνά αριθμητικά διανύσματα, γνωστά ως

διανυσματικές αναπαραστάσεις (embeddings). Τα διανύσματα αυτά κωδικοποιούν τη "σημασιολογική συγγένεια" των προϊόντων. Με άλλα λόγια, κωδικοί ειδών με παρόμοια περιγραφή θα αποκτήσουν embeddings που βρίσκονται κοντά στον διανυσματικό χώρο.

Η προσέγγιση αυτή επιτρέπει την ενσωμάτωση «σημασιολογικής γνώσης» στην πρόβλεψη της ζήτησης μετατρέποντας το αδόμητο κείμενο σε δομημένα χαρακτηριστικά (*features*), τα οποία μπορούν να εισαχθούν σε οποιονδήποτε παραδοσιακό αλγόριθμο Μηχανικής Μάθησης. Ο τρόπος εξαγωγής και αξιολόγησης αυτών των διανυσματικών αναπαραστάσεων στα δεδομένα της παρούσας εργασίας παρουσιάζεται εκτενώς στο Κεφάλαιο 5.

2.12 Νευρωνικά Δίκτυα Γράφων (Graph Neural Networks – GNN) και Μάθηση Αναπαραστάσεων

Η παραδοσιακή πρόβλεψη χρονοσειρών αντιμετωπίζει ιστορικά το κάθε προϊόν (SKU) ως μια απομονωμένη οντότητα. Ωστόσο, στο περιβάλλον της εφοδιαστικής αλυσίδας, τα είδη σπάνια είναι ανεξάρτητα, αλλά συνδέονται μέσω ενός πολύπλοκου πλέγματος σχέσεων, όπως η συμπληρωματικότητα (είδη που χρησιμοποιούνται μαζί) ή η υποκατάσταση (εναλλακτικοί κωδικοί). Η μαθηματική μοντελοποίηση τέτοιων αλληλεξαρτήσεων απαιτεί τη μετάβαση από τον Ευκλείδειο χώρο των παραδοσιακών αλγορίθμων στη θεωρία γράφων [84].

2.12.1 Δομή γράφων και διέλευση μηνυμάτων (*message passing*)

Τα Γραφικά Νευρωνικά Δίκτυα (*Graph Neural Networks* – GNNs) αποτελούν την επέκταση της Βαθιάς Μάθησης σε δεδομένα μη-Ευκλείδειας, γραφικής δομής [85]. Ο μαθηματικός πυρήνας των σύγχρονων αρχιτεκτονικών είναι ο μηχανισμός διέλευσης μηνυμάτων (*message passing*). Σε κάθε επίπεδο του δικτύου, κάθε κόμβος ενημερώνει την εσωτερική του κατάσταση (το διάνυσμα των χαρακτηριστικών του) συγκεντρώνοντας πληροφορίες από τους γειτονικούς του κόμβους:

$$\mathbf{h}_v^{(k)} = \sigma\left(\mathbf{W}^{(k)} \cdot \text{AGGREGATE}(\{\mathbf{h}_u^{(k-1)} : u \in \mathcal{N}(v)\}) + \mathbf{B}^{(k)} \cdot \mathbf{h}_v^{(k-1)}\right) \qquad (2.52)$$

όπου $\mathbf{h}_v^{(k)}$ είναι η διανυσματική αναπαράσταση του κόμβου v στο k-οστό επίπεδο, $\mathcal{N}(v)$ το σύνολο των γειτόνων του, $\mathbf{W}$ και $\mathbf{B}$ οι εκπαιδεύσιμοι πίνακες βαρών, και σ μια μη-γραμμική συνάρτηση ενεργοποίησης [86].

2.12.2 Αντιθετική και Πολυ-Εργασιακή Μάθηση

Η εξαγωγή υψηλής ποιότητας αναπαραστάσεων σε γράφους βελτιώνεται δραστικά μέσω προηγμένων τεχνικών εκπαίδευσης, πέραν της απλής παλινδρόμησης ενός μοναδικού στόχου:

- **Αντιθετική μάθηση (*Contrastive learning*):** Αποτελεί θεμέλιο της αυτο-επιβλεπόμενης μάθησης (*self-supervised learning*) σε γράφους. Ο αλγόριθμος εκπαιδεύεται μέσω δειγματοληψίας θετικών και αρνητικών ζευγών (*positive/negative sampling*). Μέσω της ελαχιστοποίησης μιας αντιθετικής συνάρτησης απώλειας (π.χ. `InfoNCE`), το δίκτυο εξαναγκάζεται να τοποθετεί τους κόμβους με ισχυρή τοπολογική ή σημασιολογική συγγένεια κοντά στον λανθάνοντα διανυσματικό χώρο, ενώ «απωθεί» τους ασυσχέτιστους κόμβους σε μεγαλύτερες αποστάσεις [87].
- **Πολυ-εργασιακή μάθηση (*Multi-task learning*):** Σε αρχιτεκτονικές αυτού του τύπου, το δίκτυο δεν βελτιστοποιείται για ένα μόνο πρόβλημα, αλλά εμπλουτίζεται με πολλαπλές, συγγενείς κεφαλές εξόδου (*auxiliary heads*). Στη θεωρία της Μηχανικής Μάθησης, η κοινή χρήση των κρυφών επιπέδων για διαφορετικές εργασίες λειτουργεί ως ισχυρός μηχανισμός κανονικοποίησης (*regularization*) και εισάγει ωφέλιμη επαγωγική μεροληψία (*inductive bias*).

Έτσι, αποτρέπεται η υπερπροσαρμογή (*overfitting*) και το δίκτυο αναγκάζεται να μάθει γενικεύσιμα χαρακτηριστικά [88]**Error! Reference source not found.**.

2.12.3 Μάθηση Αναπαραστάσεων και Υβριδικά Συστήματα

Στη σύγχρονη βιβλιογραφία πρόβλεψης, τα GNNs, συχνά, δεν χρησιμοποιούνται με λογική «από-άκρο-σε-άκρο» (*end-to-end*), ιδιαιτέρως σε προβλήματα με έντονη ασυμμετρία δεδομένων (όπως η διαλείπουσα ζήτηση), όπου τα νευρωνικά δίκτυα παραδοσιακά υστερούν. Αντίθετα, εφαρμόζεται το παράδειγμα της **Μάθησης Αναπαραστάσεων** (*Representation Learning*).

Σύμφωνα με αυτή τη θεωρητική προσέγγιση, το GNN λειτουργεί αποκλειστικά ως εξορυκτής τοπολογικών χαρακτηριστικών (*feature extractor*). Τα πυκνά διανύσματα εξόδου (*graph* embeddings) αποκόπτονται από τον γράφο και ενσωματώνονται ως στατικά χαρακτηριστικά εισόδου σε παραδοσιακούς αλγορίθμους δέντρων (όπως τα GBTs). Έτσι, συνδυάζεται η ικανότητα των γράφων να ανακαλύπτουν κρυφές σχέσεις στον χώρο με την αδιαμφισβήτητη υπεροχή των δέντρων στην επίλυση προβλημάτων παλινδρόμησης σε πίνακα [89]. Η πρακτική αξιοποίηση αυτού του υβριδικού θεωρητικού πλαισίου στο πλαίσιο της παρούσας εργασίας εξετάζεται αναλυτικά στο Κεφάλαιο 5.

2.13 Σύνοψη και γέφυρα προς τη βιβλιογραφική ανασκόπηση

Το παρόν κεφάλαιο εισήγαγε τις θεωρητικές έννοιες, τις μαθηματικές διατυπώσεις και τις αλγοριθμικές αρχές που υπόκεινται στη μεθοδολογία πρόβλεψης διαλείπουσας ζήτησης ανταλλακτικών. Η ενότητα αυτή ανακεφαλαιώνει τη συλλογιστική πορεία και χαρτογραφεί ρητά τη σύνδεση κάθε υποενότητας με τα επόμενα κεφάλαια της εργασίας.

2.13.1 Συνοπτική επισκόπηση

Η αλυσίδα συλλογισμών του Κεφαλαίου 2 ακολουθεί μια συνεκτική πορεία, από τα θεμέλια προς τις εξειδικευμένες τεχνικές:

- **Από τα γενικά στα εξειδικευμένα:** Οι ενότητες 2.1 και 2.2 θεμελίωσαν το πλαίσιο της ανάλυσης χρονοσειρών (αποσύνθεση, στασιμότητα, μετρικές αξιολόγησης MASE/WAPE) και της διαχείρισης αποθεμάτων (πολιτικές αναπλήρωσης, πίνακας SBC για κατηγοριοποίηση κωδικών). Η ενότητα 2.3 ανέδειξε τη **λογοκριμένη ζήτηση** (*censored demand*) ως δομικό πρόβλημα, ενώ η 2.4 παρουσίασε τις κλασικές μεθόδους (Croston, SBA, TSB) ως μοντέλα αναφοράς, εξηγώντας τους περιορισμούς τους.
- **Από τη Στατιστική στη Μηχανική Μάθηση:** Οι ενότητες 2.5–2.7 παρουσίασαν τη γενική θεωρία της Μηχανικής Μάθησης (χρονικά συνεπής διαχωρισμός, *bias-variance*) και την οικογένεια των **Δέντρων Διαβαθμισμένης Ενίσχυσης** (*Gradient Boosted Trees* - XGBoost, LightGBM, CatBoost), η οποία αποτελεί τον αλγοριθμικό κορμό του προτεινόμενου συστήματος.
- **Από τη μέση τιμή στην κατανομή και τα διαστήματα:** Οι ενότητες 2.8 και 2.9 αποτελούν το θεωρητικό κέντρο βάρους της εργασίας. Η ενότητα 2.8 ανέπτυξε τα **δι-σταδιακά μοντέλα Εμποδίου** (Hurdle *models*) και την παραμετρική διόρθωση εξομάλυνσης (*smearing*) ως τις καταλληλότερες δομές για την ακραία διαλείπουσα φύση των δεδομένων. Η ενότητα 2.9 παρείχε το πλαίσιο της **πιθανοκρατικής πρόβλεψης** εξηγώντας τη μαθηματική διασφάλιση της μονοτονίας μέσω του αλγορίθμου PAVA και την παραγωγή διαστημάτων πρόβλεψης βάσει εμπειρικών καταλοίπων εκτός-δείγματος.
- **Πρόσφατες εξελίξεις:** Τέλος, η ενότητα 2.10 παρουσίασε σύγχρονες ερευνητικές κατευθύνσεις, όπως η **Υπολογιστική Δεξαμενής** (Reservoir Computing) για τη δυναμική της

χρονοσειράς, τα **Γλωσσικά Μοντέλα** (Transformers) για την εξαγωγή διανυσματικών χαρακτηριστικών (embeddings) και τα Γραφικά Νευρωνικά Δίκτυα (GNN) για την εκμάθηση τοπολογικών σχέσεων.

2.13.2 Χαρτογράφηση Θεωρητικού Υποβάθρου και Μετάβαση στα Επόμενα κεφάλαια

Ο πίνακας που ακολουθεί συνδέει ρητά κάθε θεωρητική έννοια με τα σημεία της μεθοδολογίας στα οποία αξιοποιείται πρακτικά:

Πίνακας 2.4: Χαρτογράφηση θεωρητικών εννοιών στη μεθοδολογία

Έννοια / Θεωρητικό Εργαλείο	**Πρακτική Αξιοποίηση στη Μεθοδολογία**
Μετρικές MASE, WAPE, RMSE *(§2.2.3)*	Αξιολόγηση μοντέλων και σύγκριση με *baselines* (Κεφ. 5)
ADI, CV^2, Πίνακας SBC *(§2.3.2)*	Κατηγοριοποίηση SKUs, στρωματοποιημένη αξιολόγηση *(§4.6, κεφ. 5)*
Πολιτικές (s, S), ROP μέσω ποσοστημορίου *(§2.3.3)*	Προσομοίωση και αποφάσεις αναπλήρωσης (Κεφ. 6)
OOS censoring, indicator features *(§2.4.4)*	Μηχανική χαρακτηριστικών (f*eature engineering*) *(§5.6)*
Croston / SBA / TSB *(§2.5)*	Χρήση ως παραδοσιακά μοντέλα αναφοράς *(§5.6.7)*
Walk-forward validation *(§2.6.1)*	Διαδικασία αναδρομικού ελέγχου (Backtesting) *(§5.6.9)*
Global GBT models *(§2.6.4, §2.7)*	Αλγοριθμικός κορμός του κεντρικού *pipeline* *(§5.4.5, §5.6.5)*
Stacking με meta-learner *(§2.7.5)*	Κατασκευή του τελικού συνόλου (*Ensemble*) *(§5.6.10)*
Hurdle two-stage architecture *(§2.9.2)*	Κύρια αρχιτεκτονική του `True Forecast Pipeline` *(§5.6)*
Παραμετρικό Smearing correction *(§2.9.5)*	Διόρθωση λογαριθμικής μεροληψίας (*Post-prediction*) *(§5.4, §5.6)*
Tweedie loss *(§2.9.4)*	Εναλλακτικός μονοσταδιακός εκτιμητής *(§5.4, §5.5)*
PAVA / Isotonic Regression *(§2.10.2)*	Βαθμονόμηση πιθανοτήτων & χρονική μονοτονία *(§5.6)*
Διαστήματα μέσω εμπειρικών καταλοίπων *(§2.10.1)*	Υπολογισμός ορίων αβεβαιότητας *(§5.6.12)*
Γλωσσικά Μοντέλα (Embeddings) & ESN *(§2.11)*	Εξαγωγή χαρακτηριστικών και δυναμική μοντελοποίηση *(§5.5, §5.6.13)*
Γραφικά Νευρωνικά Δίκτυα (GNNs) *(§2.12)*	Εκμάθηση τοπολογικών σχέσεων (*Graph* Embeddings) *(§5.4)*

2.13.3 Γέφυρα προς το Κεφάλαιο 3

Ενώ το παρόν κεφάλαιο επιχείρησε να παρουσιάσει το θεωρητικό υπόβαθρο σε ευρύτητα και βάθος, χωρίς κριτική σύγκριση εμπειρικών αποτελεσμάτων, το Κεφάλαιο 3 (Βιβλιογραφική Ανασκόπηση) εστιάζει αποκλειστικά σε συγκεκριμένες μεθοδολογικές προτάσεις και εμπειρικές μελέτες που εξειδικεύουν τα θεωρητικά αυτά εργαλεία στο πλαίσιο της διαλείπουσας ζήτησης ανταλλακτικών. Πιο συγκεκριμένα, το Κεφάλαιο 3 διερευνά:

- **Εμπειρικές συγκρίσεις κλασικών έναντι ML μεθόδων** σε δεδομένα ανταλλακτικών [82], **Error! Reference source not found.**, με έμφαση στις συνθήκες υπό τις οποίες υπερτερεί η εκάστοτε προσέγγιση.
- **Προηγμένες αρχιτεκτονικές για διαλείπουσα ζήτηση**, όπως οι βαθιές διαδικασίες ανανέωσης (*deep renewal processes*) και οι ενοποιημένες πιθανοκρατικές επεκτάσεις των παραδοσιακών μεθόδων [72], [91].
- **Χρήση Γραφικών Νευρωνικών Δικτύων (GNNs)** για τη μοντελοποίηση αλληλεξαρτήσεων και την αξιοποίηση συσχετίσεων μεταξύ προϊόντων μέσω διανυσματικών αναπαραστάσεων (embeddings), τα οποία τροφοδοτούνται ως χαρακτηριστικά σε αλγορίθμους δέντρων [92], [93].
- **Σύζευξη της πρόβλεψης με τις αποφάσεις ελέγχου αποθεμάτων** μέσω ολοκληρωμένων πλαισίων κανονιστικής ανάλυσης (*prescriptive analytics*), τα οποία μεταφράζουν το στατιστικό σφάλμα σε πραγματικό κόστος ελλείψεων και διακράτησης [94], [95].

Στο τέλος του Κεφαλαίου 3 αναγνωρίζονται ρητά τα *ερευνητικά κενά* που υφίστανται στη σύγχρονη βιβλιογραφία και τοποθετείται με σαφήνεια η συμβολή της παρούσας διπλωματικής εργασίας ως προς αυτά. Έτσι, εάν το Κεφάλαιο 2 αποτελεί το «Εγχειρίδιο Αναφοράς» (*Reference Manual*) των αλγορίθμων, το Κεφάλαιο 3 παρέχει τον «Χάρτη του Ερευνητικού Πεδίου» (*State-of-the-Art Map*) οδηγώντας με φυσικό και τεκμηριωμένο τρόπο στη μεθοδολογική παρουσίαση και τα πειραματικά αποτελέσματα των Κεφαλαίων 4, 5 και 6.

Κεφάλαιο 3ο: Βιβλιογραφική Ανασκόπηση

Στο τρέχον κεφάλαιο λαμβάνει χώρα μια ανασκόπηση της συναφούς διεθνούς ερευνητικής βιβλιογραφίας με το αντικείμενο της παρούσας εργασίας εστιάζοντας στην πρόβλεψη της μελλοντικής ζήτησης ανταλλακτικών αυτοκινήτων (*automotive spare parts*) και την παραγωγή προτάσεων για παραγγελίες αναπλήρωσης (*replenishment orders*). Η ανασκόπηση εκκινεί από την ανάλυση των δομικών χαρακτηριστικών που διέπουν τη ζήτηση των ανταλλακτικών (§3.1) και συνεχίζει με την παρουσίαση των κλασικών στατιστικών μεθόδων πρόβλεψης διαλείπουσας ζήτησης (§3.2). Έπειτα, εξετάζονται οι σύγχρονες προσεγγίσεις Μηχανικής Μάθησης (Machine Learning – ML) και Βαθιάς Μάθησης (Deep Learning – DL) (§3.3). Δίνουν ιδιαίτερη έμφαση στα μοντέλα Εμποδίου (Hurdle *models*) και στα πλαίσια που επικεντρώνονται στη μηχανική χαρακτηριστικών (*feature-engineering-centric frameworks*), καθώς αυτά αποτελούν τον πλησιέστερο μεθοδολογικό σύνδεσμο με την προτεινόμενη ροή εργασιών (*pipeline*) της παρούσας μελέτης (§3.4).

Επίσης, παρουσιάζεται το πλαίσιο της πιθανοκρατικής πρόβλεψης (*probabilistic forecasting*) και οι διαγωνισμοί αναφοράς (*benchmarking competitions*) που καθορίζουν την τρέχουσα κατάσταση της επιστήμης (§3.5), ενώ διερευνώνται οι πλέον πρόσφατες κατευθύνσεις που αφορούν τη χρήση Μεγάλων Γλωσσικών Μοντέλων (Large Language Models – LLMs) και της Υπολογιστικής Δεξαμενής (Reservoir Computing – RC) (§3.6). Στη συνέχεια, η πρόβλεψη συνδέεται άρρηκτα με τη διαχείριση αποθεμάτων (*inventory management*) (§3.7). Το κεφάλαιο ολοκληρώνεται με τη σύνθεση των ερευνητικών κενών που εντοπίστηκαν και τα οποία στοχεύει να καλύψει η παρούσα διπλωματική εργασία (§3.8).

3.1 Ιδιαιτερότητες της ζήτησης ανταλλακτικών αυτοκινήτων

Η ζήτηση ανταλλακτικών αναγνωρίζεται διαχρονικά στη βιβλιογραφία ως ένα χαρακτηριστικό παράδειγμα διαλείπουσας (*intermittent*) και, συχνά, ακανόνιστης (*lumpy*) ζήτησης. Συχνά παρατηρούνται εκτεταμένες περίοδοι μηδενικής ζήτησης να διακόπτονται από σποραδικές, στοχαστικές αιχμές μεταβλητού μεγέθους [45][96]. Σε αντίθεση με τα καθημερινά καταναλωτικά αγαθά, η ζήτηση για ανταλλακτικά δεν απορρέει από προγραμματισμένες αγοραστικές αποφάσεις, αλλά συνδέεται άρρηκτα με τη φθορά, τις βλάβες και τις πολιτικές προληπτικής συντήρησης. Η ιδιαιτερότητα αυτή καθιστά την εν λόγω διαδικασία εξαρτημένη από τους ρυθμούς αστοχίας (*failure rates*) και τις πολιτικές συντήρησης (*maintenance policies*) και όχι από τη συμπεριφορά του εκάστοτε αγοραστή [96], [97]. Για τη συστηματική ταξινόμηση αυτών των προφίλ, η βιβλιογραφία υιοθετεί εκτενώς το ταξινομικό πλαίσιο των Syntetos, Boylan και Croston [5], το οποίο αξιοποιεί το μέσο διάστημα μεταξύ των απαιτήσεων (Average Demand Interval — ADI) και τον τετραγωνικό συντελεστή μεταβλητότητας των μεγεθών της ζήτησης (squared Coefficient of Variation — CV^2) για τη διάκριση της ζήτησης σε ομαλή (*smooth*), ασταθή (*erratic*), διαλείπουσα (*intermittent*) και ακανόνιστη (*lumpy*). Στις εφαρμογές ανταλλακτικών, η συντριπτική πλειονότητα των κωδικών εμπίπτει στις δύο τελευταίες, και πλέον απαιτητικές, κατηγορίες [45][96].

Στο οικοσύστημα της δευτερογενούς αγοράς (*aftermarket*) των ανταλλακτικών αυτοκινήτων, τα δομικά αυτά χαρακτηριστικά εντείνονται από τη λειτουργική κρισιμότητα (*criticality*) πολλών κωδικών για τη διαθεσιμότητα του οχήματος. Για παράδειγμα, η έλλειψη ενός και μόνο ζωτικού εξαρτήματος μπορεί να προκαλέσει ακινητοποίηση του στόλου (*vehicle downtime*) και, κατά συνέπεια, υψηλό επιχειρησιακό κόστος [96], [98]. Ως εκ τούτου, οι επιχειρήσεις στοχεύουν παραδοσιακά σε εξαιρετικά υψηλά επίπεδα εξυπηρέτησης (*service levels*) ακόμη και για κωδικούς με ακανόνιστη ζήτηση. Αυτή η πρακτική διογκώνει τα αποθέματα και το κόστος δέσμευσης κεφαλαίου εντείνοντας, παράλληλα, τον κίνδυνο

απαξίωσης (*obsolescence risk*) εξαιτίας μεταβολών στη γκάμα ή την τεχνολογία των οχημάτων [95], [97]. Εμπειρικές μελέτες σε πραγματικά βιομηχανικά δεδομένα επιβεβαιώνουν ότι οι κατάλογοι περιλαμβάνουν δεκάδες χιλιάδες κωδικούς με ακραία ετερογενή προφίλ ζήτησης. Ενδεικτικά, το πρόσφατο σύνολο δεδομένων που αξιοποιείται από τους Nathan *et al.* [98] για τη δευτερογενή αγορά αυτοκινήτου (*automotive aftermarket*) περιλαμβάνει περίπου 56.000 χρονοσειρές και 1,4 εκατομμύρια μηνιαίες παρατηρήσεις, με κυρίαρχο χαρακτηριστικό την ακραία σποραδικότητα (*extreme sparsity*). Αντίστοιχα ευρήματα αναφέρονται και σε προσαρμοσμένες εφαρμογές σε δίκτυα «4S (*Sale*, *Spare part*, *Service*, *Survey*)» στην Κίνα [12] και σε βιομηχανικούς καταλόγους με ετερογενείς χρόνους παράδοσης, όπου η συνύπαρξη κωδικών με άμεση διαθεσιμότητα και κωδικών με πολύμηνες καθυστερήσεις προμήθειας επιβάλλει τη χρήση ευέλικτων μοντέλων πρόβλεψης σε πολλαπλούς χρονικούς ορίζοντες [45].

Πέρα από την εγγενή διαλείπουσα φύση τους, οι ρυθμοί ζήτησης επηρεάζονται από τη σύνθεση και την ηλικία του στόλου, τις συνθήκες λειτουργίας (κλίμα, οδικό δίκτυο, ένταση χρήσης), τα πρότυπα συντήρησης και την εμφάνιση τεχνικών οδηγιών ανάκλησης (*service bulletins*), που μπορούν να προκαλέσουν απότομες μεταβολές στη ζήτηση συγκεκριμένων κωδικών (*part numbers*) [45], [96]. Παράλληλα, η άνοδος του ηλεκτρονικού εμπορίου και της αυτοσχέδιας συντήρησης (DIY *maintenance*) επιτείνουν την αβεβαιότητα σχετικά με την κατανομή της ζήτησης μεταξύ των οργανωμένων δικτύων συντήρησης και των ανεξάρτητων καναλιών [99]. Σε αυτό το πλαίσιο, τα κλασικά μοντέλα χρονοσειρών που προϋποθέτουν ομαλές ροές ή τακτική εποχικότητα (όπως η απλή εκθετική εξομάλυνση και τα μοντέλα ARIMA) παρουσιάζουν σημαντικούς περιορισμούς, οδηγώντας είτε σε συστηματικό σφάλμα υπερεκτίμησης (*over-forecasting bias*) και υπεραποθέματα, είτε σε υποεκτίμηση και αυξημένο κίνδυνο ελλείψεων [96], [98].

Τέλος, είναι κρίσιμο να επισημανθεί ότι η μαθηματική πρόοδος στα πλαίσια πρόβλεψης δεν μεταφράζεται πάντα σε επιτυχή βιομηχανική εφαρμογή. Η συστηματική ανασκόπηση των Van der Auweraer, Boute και Syntetos **Error! Reference source not found.** καταγράφει ότι τα συστήματα υποστήριξης αποφάσεων συχνά παρακάμπτονται από έμπειρους σχεδιαστές (*planners*) μέσω υποκειμενικών παρεμβάσεων (*judgmental overrides*). Η πρακτική αυτή αναδεικνύει ένα επίμονο χάσμα μεταξύ των ακαδημαϊκών υποδειγμάτων και της καθημερινής πρακτικής. Η διαπίστωση αυτή υπογραμμίζει την ανάγκη για ροές εργασίας (*pipelines*) που δεν εξασφαλίζουν μόνο στατιστική ακρίβεια, αλλά ενσωματώνονται οργανικά στα πληροφοριακά συστήματα, ώστε να παρέχουν διαφανείς και ερμηνεύσιμες προτάσεις αναπλήρωσης [95], [97].

3.2 Κλασικές μέθοδοι πρόβλεψης διαλείπουσας ζήτησης

Η πρώτη συστηματική προσπάθεια αντιμετώπισης της διαλείπουσας ζήτησης οφείλεται στον Croston [4], ο οποίος εισήγαγε έναν μηχανισμό πρόβλεψης ειδικά προσαρμοσμένο σε χρονοσειρές με έντονη **σποραδικότητα (*sparsity*)** και πολλά μηδενικά λόγω παρατεταμένης έλλειψης ζήτησης. Η θεμελιώδης ιδέα της μεθόδου έγκειται στον διαχωρισμό της διαδικασίας σε δύο συνιστώσες, το μέσο μέγεθος της ζήτησης (*demand size*), όταν αυτή εμφανίζεται, και το μέσο διάστημα μεταξύ δύο διαδοχικών μη μηδενικών παρατηρήσεων (*inter-demand interval*). Οι δύο αυτές ποσότητες εκτιμώνται μέσω Απλής Εκθετικής Εξομάλυνσης (Simple Exponential Smoothing – SES) και ενημερώνονται μόνο όταν παρατηρείται θετική ζήτηση. Η πρόβλεψη του ρυθμού ζήτησης ανά περίοδο προκύπτει, στη συνέχεια, ως ο λόγος των εξομαλυμένων ποσοτήτων του μεγέθους και του χρονικού διαστήματος [4]. Η μέθοδος αυτή ενσωματώθηκε ευρέως σε πληροφοριακά συστήματα διαχείρισης αποθεμάτων και παραμένει μέχρι σήμερα το βασικό σημείο αναφοράς (*baseline*) σε σενάρια ανταλλακτικών.

Μεταγενέστερες αναλύσεις ανέδειξαν ότι η αρχική διατύπωση του Croston είναι μεροληπτική και οδηγεί σε συστηματική υπερεκτίμηση της μέσης ζήτησης. Οι Syntetos και Boylan [1] πρότειναν τη διορθωμένη παραλλαγή, γνωστή ως Προσέγγιση Syntetos–Boylan (Syntetos–Boylan Approximation – SBA), εισάγοντας έναν διορθωτικό συντελεστή στον τελικό λόγο, ώστε να επιτευχθεί ασυμπτωτικά αμερόληπτη πρόβλεψη. Η ανάλυσή τους, η οποία βασίστηκε σε ένα εκτενές βιομηχανικό δείγμα περίπου 3.000 χρονοσειρών από τον κλάδο του αυτοκινήτου, κατέδειξε ότι η SBA υπερέχει της αρχικής μεθόδου τόσο σε όρους μέσου τετραγωνικού σφάλματος όσο και σε μετρικές απόδοσης αποθεμάτων (*inventory performance*), ιδιαίτερα σε προφίλ με μέτρια έως υψηλή ένταση σποραδικότητας [1][98]. Την ίδια περίοδο, οι Shenstone και Hyndman [36] διερεύνησαν το στοχαστικό υπόβαθρο της μεθόδου Croston και κατέληξαν στο συμπέρασμα ότι τα γραμμικά μοντέλα είναι σε μεγάλο βαθμό ασύμβατα με τις ιδιότητες των διαλειπουσών σειρών, όπως η μη-αρνητικότητα και η ακέραια φύση των τιμών. Παρ' όλα αυτά, οι σημειακές προβλέψεις (*point forecasts*) και τα διαστήματα πρόβλεψης (*prediction intervals*) των μεθόδων αυτών παραμένουν πρακτικά χρήσιμα για τον υπολογισμό του αποθέματος ασφαλείας (*safety stock*) [36], [97].

Για την αντιμετώπιση φαινομένων όπως οι παρατεταμένες περίοδοι μηδενικής ζήτησης και η σταδιακή απαξίωση των κωδικών, οι Teunter, Syntetos και Babai [2] εισήγαγαν τη μέθοδο TSB (Teunter–Syntetos–Babai). Η προσέγγιση αυτή αντικαθιστά την εξομάλυνση των διαστημάτων με την απευθείας εξομάλυνση της πιθανότητας εμφάνισης ζήτησης ανά περίοδο. Η πιθανότητα αυτή ενημερώνεται σε κάθε χρονικό βήμα, ακόμη και όταν η ζήτηση είναι μηδενική. Με αυτό τον τρόπο καθίσταται δυνατή η σταδιακή και ομαλή μείωση της εκτιμώμενης τιμής σε περιπτώσεις φθίνοντος κύκλου ζωής του προϊόντος. Παράλληλα, στην ίδια κατεύθυνση προτάθηκαν εναλλακτικές παραλλαγές, όπως η Υπερβολική-Εκθετική Εξομάλυνση (Hyperbolic–Exponential Smoothing – HES) και η Γραμμική Εκθετική Εξομάλυνση (Linear Exponential Smoothing – LES), οι οποίες διαφοροποιούνται ως προς τον μαθηματικό χειρισμό του λόγου μεγέθους/διαστήματος σε περιπτώσεις εκτεταμένων περιόδων απραξίας επιχειρώντας να ελέγξουν την αστάθεια που εμφανίζουν οι προσεγγίσεις τύπου Croston υπό ακραίες συνθήκες σποραδικότητας [45], [72], [101].

Η βιβλιογραφία επεκτάθηκε συστηματικά προς τη βελτιστοποίηση αυτών των στατιστικών υποδειγμάτων και τη διασύνδεσή τους με επιχειρησιακά κριτήρια. Ο Kourentzes [101] εστίασε στη βελτιστοποίηση των παραμέτρων εξομάλυνσης με βάση την απόδοση του αποθέματος (*inventory performance*) αντί για αποκλειστικά στατιστικά κριτήρια σφάλματος προσπαθώντας να αποδείξει ότι η παραμετροποίηση επιδρά με διαφορετικό τρόπο στην ακρίβεια πρόβλεψης και στα τελικά επίπεδα αποθεμάτων. Στο ίδιο πλαίσιο, οι Teunter και Sani [102], καθώς και οι Syntetos, Babai και Gardner **Error! Reference source not found.**, τεκμηρίωσαν ότι η αποτελεσματικότητα των στατιστικών μεθόδων πρόβλεψης δεν πρέπει να αξιολογείται απομονωμένα, καθότι εξαρτάται άμεσα από τη δομική και μαθηματική συμβατότητά τους με τη συνοδευτική πολιτική ελέγχου αποθεμάτων που εφαρμόζει η επιχείρηση. Συγκεκριμένα, σε μια πολιτική συνεχούς ελέγχου (*continuous review policy* – (*Q, R*)), κατά την οποία εκδίδεται αυτόματα μια σταθερή ποσότητα παραγγελίας (*Q*), μόλις η θέση του αποθέματος υποχωρήσει στο προκαθορισμένο σημείο αναπαραγγελίας (reorder point – *R*), η μέθοδος πρόβλεψης καλείται να εκτιμήσει με ακρίβεια τη ζήτηση αποκλειστικά κατά τη διάρκεια του χρόνου παράδοσης (*lead time – L*). Αντιθέτως, σε μια πολιτική περιοδικού ελέγχου (*periodic review policy – (T, S)*), όπου η κατάσταση του αποθέματος εξετάζεται ανά τακτά χρονικά διαστήματα (*T*) και εκδίδεται μεταβλητή παραγγελία για την αναπλήρωσή του έως ένα μέγιστο επίπεδο (*order-up-to level – S*), η απαιτούμενη πρόβλεψη πρέπει να καλύψει έναν διευρυμένο ορίζοντα προστασίας, ίσο με το άθροισμα του μεσοδιαστήματος ελέγχου και του χρόνου παράδοσης ($T + L$). Κατά συνέπεια, επειδή οι παραδοσιακοί μαθηματικοί τύποι των πολιτικών αυτών βασίζονται σε τυπικές θεωρητικές κατανομές (όπως η

κατανομή Poisson), σε συνθήκες έντονης σποραδικότητας αποτυγχάνουν να συλλάβουν τις βαριές ουρές (*heavy tails*) της πραγματικής ζήτησης. Το γεγονός αυτό αναδεικνύει ότι η απλή σημειακή πρόβλεψη (*point forecast*) είναι ανεπαρκής για στοχαστικές επιχειρησιακές αποφάσεις, καθιστώντας αναγκαία είτε την εκτίμηση της πλήρους πιθανοκρατικής κατανομής (*probabilistic forecasting*) είτε τη χρήση μη παραμετρικών προσεγγίσεων αναδειγματοληψίας (*bootstrapping*) [103].

Επιπλέον, το ταξινομικό σχήμα *ADI–CV²* των Syntetos, Boylan και Croston [5] καθιερώθηκε ως ένα αξιόπιστο εργαλείο επιλογής μοντέλου, σύμφωνα με το οποίο προκρίνονται διαφορετικές μέθοδοι (SES, Croston, SBA, TSB) ανάλογα με τον συνδυασμένο βαθμό σποραδικότητας και μεταβλητότητας της εκάστοτε χρονοσειράς [5][96].

Σε πιο πρόσφατες μελέτες, επιχειρήθηκε η ενσωμάτωση των κλασικών αυτών προσεγγίσεων σε σύγχρονα υπολογιστικά περιβάλλοντα. Οι Türkmen et al. [72] γενίκευσαν τις μεθόδους τύπου Croston σε ένα ενιαίο πιθανοκρατικό πλαίσιο, βασισμένο σε βαθιές διαδικασίες ανανέωσης (*deep renewal processes*) και αναδρομικά νευρωνικά δίκτυα (Recurrent Neural Networks – RNN). Στο διμερές αυτό μοντέλο, η εμφάνιση ζήτησης και το αντίστοιχο μέγεθος αυτής μοντελοποιούνται από κοινού, με αποτέλεσμα οι κλασικές στατιστικές μέθόδοι (Croston, SBA, TSB) να προκύπτουν ως ειδικές περιπτώσεις του γενικότερου πιθανοκρατικού υποδείγματος βαθιάς μάθησης. Η εξέλιξη αυτή επιτρέπει τη διατήρηση της μεθοδολογικής συνέχειας από τα κλασικά υποδείγματα προς τις σύγχρονες αρχιτεκτονικές δίχως να ακυρώνεται η αξία των στατιστικών μοντέλων αναφοράς.

Η αντικειμενική αξιολόγηση των κλασικών μεθόδων συνδέεται άρρηκτα με την επιλογή των κατάλληλων μετρικών σφάλματος. Οι Hyndman και Koehler [20] πρότειναν τη χρήση μετρικών ελεύθερων κλίμακας (*scale-free metrics*), όπως το Μέσο Απόλυτο Κλιμακωτό Σφάλμα (Mean Absolute Scaled Error – MASE). Μάλιστα αναδεικνύουν ότι οι παραδοσιακοί δείκτες, όπως το Μέσο Απόλυτο Ποσοστιαίο Σφάλμα (MAPE) ή το Ριζικό Μέσο Τετραγωνικό Σφάλμα (RMSE), εμφανίζουν αστάθεια ή αδυναμία μαθηματικού προσδιορισμού σε χρονοσειρές πυκνής διατύπωσης με προσθήκη μηδενικών κατά την έλλειψη ζήτησης. Οι πρόσφατοι διαγωνισμοί προβλέψεων M4 και M5 [21], [104] υιοθέτησαν σχεδόν καθολικά κλιμακωτές μετρικές, όπως το Ριζικό Μέσο Τετραγωνικό Κλιμακωτό Σφάλμα (Root Mean Squared Scaled Error – RMSSE) και τη σταθμισμένη εκδοχή αυτού (WRMSSE), οι οποίες κυριαρχούν σε περιβάλλοντα λιανικής αγοράς και διαλείπουσας ζήτησης.

Επιπλέον, ο Kolassa [14], [15] επισημαίνει ότι η επιλογή της μετρικής καθορίζει ουσιαστικά τη φύση της «βέλτιστης σημειακής πρόβλεψης». Δηλαδή, η ελαχιστοποίηση των μετρικών τύπου L_2 (RMSE, RMSSE) οδηγεί σε προβλέψεις που στοχεύουν στη μέση τιμή (*mean*), ενώ οι μετρικές τύπου L_1 (MAE, MASE, WAPE) στοχεύουν στη διάμεσο (*median*). Η διαπίστωση αυτή είναι εξαιρετικά κρίσιμη για σποραδικές χρονοσειρές, καθώς η βέλτιστη πρόβλεψη υπό την έννοια της διαμέσου είναι συχνά το μηδέν. Αυτό σημαίνει ότι η επιλογή της μετρικής δεν αξιολογεί απλώς, αλλά καθοδηγεί τη συμπεριφορά και την επιλογή του ίδιου του μοντέλου πρόβλεψης [14], [45], [96].

Συνολικά, οι μέθοδοι τύπου Croston και οι στατιστικές παραλλαγές τους παραμένουν αναπόσπαστο στοιχείο κάθε συγκριτικής αξιολόγησης στο πεδίο της διαλείπουσας ζήτησης. Παρά τους γνωστούς περιορισμούς τους, όπως η υψηλή ευαισθησία στη ρύθμιση των παραμέτρων και η αδυναμία ενσωμάτωσης εξωγενών μεταβλητών, παρέχουν το απαραίτητο υπόβαθρο (*baseline*) για την αποτίμηση των μοντέλων Μηχανικής Μάθησης, Βαθιάς Μάθησης και Μεγάλων Γλωσσικών Μοντέλων που εξετάζονται στη συνέχεια [45], [96], [98].

3.3 Μέθοδοι Μηχανικής Μάθησης για διαλείπουσα ζήτηση

Κατά την τελευταία δεκαετία, η διεθνής βιβλιογραφία έχει στραφεί έντονα προς τις μεθόδους Μηχανικής Μάθησης (Machine Learning – ML) για την πρόβλεψη της διαλείπουσας ζήτησης. Με τον τρόπο αυτό, επιχειρεί να υπερβεί τους εγγενείς περιορισμούς των στατιστικών υποδειγμάτων τύπου Croston και να ενσωματώσει πρόσθετα πληροφοριακά στοιχεία, όπως εξωγενείς μεταβλητές, ποιοτικά χαρακτηριστικά προϊόντων και συσχετίσεις των διαφορετικών κωδικών ειδών. Η συστηματική ανασκόπηση των Giannopoulos, Dasaklis, Tsantilis και Patsakis [45] καταγράφει ενενήντα εννέα στοχευμένες μελέτες που καλύπτουν ένα ευρύ φάσμα τεχνικών, από δέντρα αποφάσεων (*decision trees*), Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines – SVM) και τον αλγόριθμο των *k*-Πλησιέστερων Γειτόνων (*k*-Nearest Neighbors – *k*-NN), έως βαθιά νευρωνικά δίκτυα, μεθόδους διαβαθμισμένης ενίσχυσης (*gradient boosting*) και σχήματα μετα-μάθησης (*meta-learning*). Έτσι, αναδεικνύουν ότι η χρήση της Μηχανικής Μάθησης είναι πλέον διάχυτη τόσο στη διαχείριση ανταλλακτικών (*spare parts*) όσο και σε κωδικούς λιανεμπορίου (*retail* SKUs) με σποραδικά προφίλ.

Η ειδοποιός διαφορά σε σχέση με τα κλασικά υποδείγματα έγκειται στο ότι τα μοντέλα Μηχανικής Μάθησης αναδιαμορφώνουν το πρόβλημα της πρόβλεψης ως ένα γενικότερο πρόβλημα Επιβλεπόμενης Μάθησης (Supervised Learning). Στο πλαίσιο αυτό, οι ιστορικές υστερήσεις (*lags*) της χρονοσειράς, τα κατηγορικά χαρακτηριστικά (όπως η κλάση προϊόντος, ο προμηθευτής και η λειτουργική κρισιμότητα), καθώς, και μακροοικονομικοί ή λειτουργικοί δείκτες μετασχηματίζονται σε διανύσματα χαρακτηριστικών (*feature vectors*), τα οποία τροφοδοτούν τον εκάστοτε αλγόριθμο [3], [45], [99].

3.3.1 Νευρωνικά δίκτυα πρόσθιας τροφοδοσίας (Feed-Forward Neural Networks)

Μία από τις πρώτες στοχευμένες προσπάθειες εφαρμογής τεχνητών νευρωνικών δικτύων σε προβλήματα σποραδικής ζήτησης αποτελεί η εργασία του Kourentzes [82], ο οποίος πρότεινε μια μεθοδολογία Νευρωνικών Δικτύων Πρόσθιας Τροφοδοσίας (Feed-Forward Neural Networks – FFNN) για τη δυναμική εκτίμηση του ρυθμού ζήτησης (*demand rate*). Το μοντέλο δεν χρησιμοποιεί έναν υποτιθέμενο σταθερό ή γραμμικό ρυθμό, αλλά εκπαιδεύεται να συλλαμβάνει ταυτόχρονα τις μη γραμμικές αλληλεπιδράσεις μεταξύ των θετικών ποσοτήτων και των μεσοδιαστημάτων εμφάνισης της ζήτησης. Αυτό επιτυγχάνεται με την αξιοποίηση τεχνικών εκπαίδευσης με κανονικοποίηση (*regularised training*) και συνάρθροισης διαμέσου (*median ensembles*), ώστε να μετριάζονται οι επιπτώσεις του μικρού δείγματος. Το πλέον αξιοσημείωτο εύρημα της μελέτης αυτής είναι ότι, αν και τα FFNN δεν παρουσίαζαν σαφή υπεροχή στις κλασικές στατιστικές μετρικές σφάλματος, επέτυχαν σημαντικά υψηλότερα επίπεδα εξυπηρέτησης (*service levels*) σε σύγκριση με την καλύτερη στατιστική παραλλαγή του Croston και χωρίς, μάλιστα, να αυξηθεί το μέσο υπάρχον απόθεμα. Το γεγονός αυτό παρείχε την πρώτη ισχυρή ένδειξη ότι η αξιολόγηση μοντέλων αραιής ζήτησης αποκλειστικά με παραδοσιακές στατιστικές μετρικές σφάλματος είναι επιχειρησιακά ανεπαρκής [82].

Σε μεταγενέστερη μελέτη, οι Hoffmann, Lasch και Meinig [105] συνέκριναν συστηματικά οκτώ διαφορετικές αρχιτεκτονικές τεχνητών νευρωνικών δικτύων με μονό κρυφό στρώμα (*single hidden layer* ANNs) έναντι έντεκα κλασικών στατιστικών μεθόδων αξιολογώντας την απόδοσή τους ξεχωριστά ανά κατηγορία ζήτησης κατά το σχήμα των Syntetos *et al.* [5]. Βασιζόμενοι σε πραγματικά δεδομένα ανταλλακτικών μηχανολογικού εξοπλισμού, εξέτασαν την επίδραση του πλήθους των κρυφών νευρώνων και των αλγορίθμων εκπαίδευσης, αντιπαραβάλλοντας την κλασική οπισθοσκέδαση (backpropagation) με τον αλγόριθμο Levenberg–Marquardt. Εν τέλει, κατέληξαν ότι τα κατάλληλα ρυθμισμένα νευρωνικά δίκτυα εμφανίζουν υψηλή ανταγωνιστικότητα, κυρίως σε προφίλ με έντονα ακανόνιστη (*lumpy*) συμπεριφορά [45], [105]. Ανάλογα συμπεράσματα διατυπώνονται και σε μελέτες

για ανταλλακτικά της αεροπορικής βιομηχανίας (*civil aerospace spare parts*) [106], **Error! Reference source not found.**, στις οποίες τεκμηριώνεται ότι η ενσωμάτωση περιγραφικών δεικτών σποραδικότητας (όπως διάστημα που μεσολάβησε από την τελευταία θετική ζήτηση) και η εφαρμογή Μπεϊζιανής κανονικοποίησης (*Bayesian regularisation*) βελτιώνουν αισθητά τη σταθερότητα των FFNN έναντι απλών/«αφελών» (*naive*) αρχιτεκτονικών.

Πέρα από τις βασικές εφαρμογές, η βιβλιογραφία περιλαμβάνει και πιο εξειδικευμένες μελέτες που αφορούν επιμέρους κλάδους, όπως το ηλεκτρονικό εμπόριο, η μόδα και ο χώρος της αυτοκινητοβιομηχανίας ή πιο σύνθετα μοντέλα που συνδυάζουν νευρώνες τύπου «*Perceptron*» πολλαπλών επιπέδων (Multi-Layer Perceptron – MLP) με κλασικές μεθόδους, όπως ARIMA, Theta ή τύπου Croston [12], [45]. Σε όλες αυτές τις περιπτώσεις επιβεβαιώνεται ότι η επιτυχία των δικτύων πρόσθιας τροφοδοσίας εξαρτάται σε κρίσιμο βαθμό από την ποιότητα του σχεδιασμού των χαρακτηριστικών, τον όγκο των διαθέσιμων δεδομένων και τη συστηματική ρύθμιση των υπερπαραμέτρων.

3.3.2 Μοντέλα ακολουθιών (RNN, LSTM και Transformers)

Η δεύτερη μεθοδολογική κατεύθυνση εντοπίζεται στο πεδίο της Βαθιάς Μάθησης (Deep Learning) και αφορά τα μοντέλα ακολουθιών (*sequence models*). Εδώ κυριαρχούν τα Αναδρομικά Νευρωνικά Δίκτυα (Recurrent Neural Networks – RNN) και τα Δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (Long Short-Term Memory – LSTM), τα οποία σχεδιάστηκαν για την άμεση εκμετάλλευση των χρονικών εξαρτήσεων. Στο πεδίο του λιανεμπορίου, οι Ma και Fildes [108] εισήγαγαν ένα Συνελικτικό Νευρωνικό Δίκτυο Διπλού Καναλιού (Double-Channel Convolutional NN – DCCNN) στο πλαίσιο της μετα-μάθησης (*meta-learning*). Το μοντέλο αυτό εξάγει αυτόματα αναπαραστάσεις χαρακτηριστικών από τις ιστορικές πωλήσεις και εξωγενείς παράγοντες (όπως προωθητικές ενέργειες και μεταβολές τιμών) μαθαίνοντας να συνδυάζει βέλτιστα ένα σύνολο βασικών μοντέλων για κάθε σειρά. Παρά το γεγονός ότι το σύνολο δεδομένων τους δεν ήταν αποκλειστικά διαλείπον, περιλάμβανε πλήθος σειρών με υψηλή σπανιότητα και μεταβλητότητα αποδεικνύοντας, έτσι, ότι τα βαθιά σχήματα μετα-μάθησης μπορούν να υπερέχουν των ανεξάρτητων στατιστικών μοντέλων σε μεγάλης κλίμακας καταλόγους ειδών [45], [108].

Ειδικά για τη διαλείπουσα ζήτηση, οι Praveena και Prasanna Devi [109] πρότειναν ένα υβριδικό μοντέλο που συνδυάζει δίκτυα LSTM με μοντέλα ARIMA ή Croston. Το LSTM αναλαμβάνει τη σύλληψη των μη γραμμικών, μακροχρόνιων χρονικών εξαρτήσεων, ενώ οι κλασικές μέθόδοι τη διαχείριση της βασικής δομής. Επίσης, πρόσφατες μελέτες, που αφορούν τον τομέα του ηλεκτρονικού εμπορίου, αναδεικνύουν τη δυναμική των μοντέλων καθολικής μάθησης (*cross-learning* LSTM), όπου μια ενιαία αρχιτεκτονική εκπαιδεύεται ταυτόχρονα στο σύνολο των διαθέσιμων χρονοσειρών [45]. Με τον τρόπο αυτό, το δίκτυο εκμεταλλεύεται κοινά συμπεριφορικά μοτίβα μεταξύ διαφορετικών προϊόντων επιτυγχάνοντας, έτσι, ανώτερη γενίκευση σε σχέση με τα ανεξάρτητα μονής μεταβλητής (*univariate*) μοντέλα ανά κωδικό είδους (SKU), ιδιαίτερα όταν το ιστορικό ανά κωδικό είναι περιορισμένο. Αντίστοιχες προσεγγίσεις για ανταλλακτικά αεροσκαφών [110], στις οποίες συνδυάζονται πολυεπίπεδες δομές RNN-LSTM με τροποποιημένους βελτιστοποιητές τύπου `Adam` για τη βελτίωση της σύγκλισης σε μακροχρόνιες εξαρτήσεις (*long-term dependencies*).

Η εμφάνιση των αρχιτεκτονικών τύπου Transformers [83] εισήγαγε ένα νέο παράδειγμα στην πρόβλεψη χρονοσειρών. Μοντέλα όπως τα «Informer, Autoformer, Temporal Fusion Transformer (TFT) και PatchTST» πέτυχαν εντυπωσιακά αποτελέσματα σε προβλήματα εκτεταμένου ορίζοντα (*long-range*

forecasting) σε σύγκριση με τα RNN και CNN, βασιζόμενα σε μηχανισμούς αυτο-προσοχής (*self-attention*) [3], **Error! Reference source not found.**.

Ωστόσο, η σύγχρονη βιβλιογραφία εμφανίζεται ιδιαίτερα επιφυλακτική ως προς την καθολική υπεροχή των Transformers σε συνθήκες ακραίας σποραδικότητας. Οι Tan, Merrill, Gupta, Althoff και Hartvigsen [112] τεκμηρίωσαν, μέσω εκτενών πειραμάτων αφαίρεσης (*ablation studies*), ότι η απλή αφαίρεση των συνιστωσών από πολύπλοκες αρχιτεκτονικές βαθιάς μάθησης όχι μόνο δεν μείωσε την ακρίβεια, αλλά, συχνά, βελτίωσε τις επιδόσεις. Το εύρημα αυτό ενισχύει τη θέση ότι σε πολύ αραιές σειρές, η αρχιτεκτονική πολυπλοκότητα υποχωρεί έναντι της σωστής κατασκευής χαρακτηριστικών (*feature engineering*).

Τέλος, ένα διακριτό ρεύμα αφορά τα ενοποιημένα πιθανοκρατικά πλαίσια, όπως αυτό των Türkmen *et al.* [72]. Στο πλαίσιο αυτό, RNN αρχιτεκτονικές εκπαιδεύονται για την εκτίμηση ολόκληρης της κατανομής της μελλοντικής ζήτησης μέσω βαθιών διαδικασιών ανανέωσης (*deep renewal processes*). Έτσι παράγονται διαστήματα εμπιστοσύνης (*quantiles*) αντί για απλές σημειακές προβλέψεις, τα οποία επιτρέπουν την απευθείας ενσωμάτωση των προβλέψεων σε πολιτικές διαχείρισης αποθεμάτων υψηλού ρίσκου [72].

3.3.3 Δέντρα αποφάσεων, μέθοδοι διαβαθμισμένης ενίσχυσης και σύνολα μοντέλων

Ένας κυρίαρχος όγκος ερευνητικών και εφαρμοσμένων εργασιών εστιάζει σε δομές δέντρων αποφάσεων, όπως τα Τυχαία Δάση (Random Forests – RF) και τα Δέντρα Διαβαθμισμένης Ενίσχυσης (Gradient Boosted Trees – GBT). Αλγόριθμοι αιχμής στα GBTs θεωρούνται τα πλαίσια (*frameworks*) XGBoost [54], LightGBM [55] και CatBoost [56]. Οι μέθοδοι αυτές είναι ιδιαίτερα ελκυστικές, όσον αφορά βιομηχανικές εφαρμογές μεγάλης κλίμακας, διότι ενσωματώνουν με φυσικό τρόπο εκατοντάδες κατηγορικά και αριθμητικά χαρακτηριστικά (όπως χρόνους παράδοσης, κατηγοριοποίηση προϊόντων, μακροοικονομικούς δείκτες), προσφέρουν ιδανική ισορροπία μεταξύ ακρίβειας και υπολογιστικού κόστους και υποστηρίζονται από ώριμα εργαλεία αποτίμησης της σπουδαιότητας των χαρακτηριστικών (*feature importance*) και αυτοματοποιημένης ρύθμισης υπερπαραμέτρων [113].

Χαρακτηριστικό παράδειγμα αποτελεί η εργασία των S. Wang *et al.* [91], όπου GBT μοντέλα εμπλουτίζονται με πληροφορίες συσχέτισης μεταξύ διαφορετικών προϊόντων (*cross-sectional*) και χωρικές πληροφορίες για τη βελτίωση των προβλέψεων σε αραιά δεδομένα πωλήσεως λιανικού εμπορίου (*intermittent retail*). Αντίστοιχα, οι Lainder και Wolfinger [114] παρουσίασαν τη νικήτρια λύση του παγκόσμιου διαγωνισμού «*M5 Uncertainty*», η οποία βασίστηκε αποκλειστικά στον αλγόριθμο LightGBM σε συνδυασμό με εντατικό feature engineering και τεχνικές επαύξησης δεδομένων (*data augmentation*) επιβεβαιώνοντας ότι τα GBT αποτελούν το «*de facto state-of-the-art*» σε περιβάλλοντα διαλείπουσας ζήτησης.

Βασιζόμενοι στο γενικότερο πλαίσιο συνδυασμού προβλέψεων με γνώμονα τη διαφορετικότητα που εισήγαγαν οι Kang *et al.* [115], οι δημιουργοί κατέδειξαν ότι η επιλογή επιμέρους μοντέλων με υψηλή ανομοιογένεια και συμπληρωματικά σφάλματα είναι συχνά πιο κρίσιμη για τη συνολική ευστάθεια του συστήματος από την ατομική ακρίβεια του κάθε μοντέλου ξεχωριστά. Πάνω σε αυτή τη φιλοσοφία, οι Li *et al.* **Error! Reference source not found.** πρότειναν ένα εξειδικευμένο συνδυαστικό πλαίσιο βασισμένο στα χαρακτηριστικά και τη διαφορετικότητα (*feature- and diversity-based combination framework*) για περιπτώσεις διαλείπουσας ζήτησης. Στην προσέγγιση αυτή, ένα μοντέλο μετα-μάθησης (*meta-learner*) βασισμένο στον αλγόριθμο XGBoost εκπαιδεύεται να συνδυάζει στατιστικά μοντέλα και μοντέλα μηχανικής μάθησης (ML) με βάση τα στατιστικά χαρακτηριστικά των χρονοσειρών, επιτυγχάνοντας κορυφαίες επιδόσεις στα διαδεδομένα σύνολα δεδομένων αναφοράς «RAF» και «M5».

Παράλληλα, οι Wang *et al.* [91] κατέδειξαν ότι ο συνδυασμός LightGBM και LSTM επιτρέπει την παραγωγή πλήρους έκτασης πιθανοκρατικών πυκνοτήτων (όπως η Αρνητική Διωνυμική κατανομή – Negative Binomial), με αποτέλεσμα να παράγονται διακριτές συναρτήσεις μάζας πιθανότητας (*probability mass functions*) άμεσα αξιοποιήσιμες από στοχαστικά μοντέλα διαχείρισης αποθεμάτων.

Σε καθαρά εφαρμοσμένο βιομηχανικό επίπεδο, η διατριβή του Manjunath [117] για την Atlas Copco Industrial Technique AB συνέκρινε στατιστικές μεθόδους, όπως η Exponential Triple Smoothing, με μοντέλα βασισμένα στους αλγορίθμους Random Forest και XGBoost πάνω σε ένα πραγματικό παγκόσμιο δίκτυο βιομηχανικών προϊόντων και κατέγραψε την κυριαρχία των Random Forest ως προς τη σχετική ακρίβεια, ενώ ταυτόχρονα ανέδειξε πρακτικά ζητήματα υλοποίησης (ποιότητα ανεπεξέργαστων δεδομένων, υποδομές, χάσμα δεξιοτήτων). Σε ανάλογη τροχιά, οι Dodin *et al.* [118] περιέγραψαν την ολοκληρωμένη ροή εργασιών (*integrated pipeline*) της Bombardier για τη δευτερογενή αγορά της πολιτικής αεροπορίας (*civil aerospace aftermarket*), όπου GBT μοντέλα συνδυάζονται με στατιστικές μεθόδους μέσω ensemble στρατηγικών ανά ομάδα μοτίβου ζήτησης (*demand pattern group*) και δημιουργούν ένα σύστημα εξαιρετικά ανθεκτικό στον θόρυβο. Το κοινό συμπέρασμα όλων αυτών των μελετών είναι ότι **τα GBT μοντέλα υπερέχουν συστηματικά των καθαρών αρχιτεκτονικών Βαθιάς Μάθησης,** όταν το μέγεθος του συνόλου δεδομένων είναι μεσαίο προς μεγάλο και τα χαρακτηριστικά έχουν σχεδιαστεί επιμελώς [3], [98].

3.3.4 Νευρωνικά δίκτυα γράφων και εξαρτήσεις μεταξύ χρονοσειρών (cross-series dependencies)

Μια πρόσφατη και πολλά υποσχόμενη κατεύθυνση στη βιβλιογραφία επιχειρεί να ξεπεράσει τη μονομεταβλητή (*univariate*) προσέγγιση μοντελοποιώντας ρητά τις συσχετίσεις μεταξύ προϊόντων μέσω Νευρωνικών Δικτύων Γράφων (Graph Neural Networks – GNN). Η βασική διαφορά των GNNs είναι ότι επιτρέπουν τη ροή πληροφορίας μεταξύ παρόμοιων ή συσχετιζόμενων κωδικών (λόγω κοινού προμηθευτή, ίδιας κατηγορίας ή προτύπων συναγοράς) τόσο κατά την εκπαίδευση όσο και κατά τη φάση της εξαγωγής συμπερασμάτων (*inference*) [92][93].

Οι Kozodoi *et al.* [92] πρότειναν το μοντέλο «GraphDeepAR», μια ολοκληρωμένη (*end-to-end*) αρχιτεκτονική, όπου ένας GNN κωδικοποιητής (*encoder*) βασισμένος στην ομοιότητα των χαρακτηριστικών των προϊόντων (*article attribute similarity*) τροφοδοτεί έναν πιθανοκρατικό DeepAR αποκωδικοποιητή (*decoder*). Το μοντέλο κατασκευάζει τον γράφο δυναμικά μέσω της ομοιότητας των μεταβλητών και παράγει ενσωματώσεις προϊόντων (*article* embeddings) που μεταφέρουν τη συλλογική πληροφορία. Τα πειράματά τους έδειξαν συστηματική υπεροχή έναντι των μη-γραφικών μοντέλων αναφοράς. Στην ίδια κατεύθυνση, οι Gandhi *et al.* [119] ανέπτυξαν GNN μοντέλα για πλατφόρμες ηλεκτρονικού εμπορίου καταγράφοντας σημαντικά κέρδη σε κωδικούς με πρόβλημα ψυχρής εκκίνησης (*cold-start* SKU*s*), δηλαδή κωδικούς χωρίς πρότερο ιστορικό πωλήσεων.

Οι Fatemi *et al.* [93], στο πλαίσιο του μοντέλου «Cold Causal Demand Forecasting (CDF-cold)», αντιμετώπισαν το πρόβλημα της ψυχρής εκκίνησης συνδυάζοντας την αιτιακή συμπερασματολογία (*causal inference*) με τη Βαθιά Μάθηση για την εκτίμηση των σχέσεων μεταξύ χρονοσειρών. Αν και η αρχική εφαρμογή αφορούσε τη δικτυακή κίνηση στα κέντρα δεδομένων της Google, η μεθοδολογική προσέγγιση, δηλαδή η χρήση αιτιακών σχέσεων για τη γεφύρωση της απουσίας ιστορικού, είναι άκρως σχετική με την εισαγωγή νέων κωδικών ανταλλακτικών (*new part numbers*) στην αγορά αυτοκινήτου. Αυτό αποδεικνύει ότι η ρητή μοντελοποίηση των αλληλοεξαρτήσεων μπορεί να αντισταθμίσει την παντελή έλλειψη ιστορικών παρατηρήσεων.

3.3.5 Υβριδικές προσεγγίσεις, μετα-μάθηση και μεταφορά γνώσης

Ένα σημαντικό τμήμα της πρόσφατης έρευνας εστιάζει στη σύνθεση διαφορετικών μεθοδολογικών εργαλείων. Το πλαίσιο UNISON των Fu και Chien [120] αποτελεί χαρακτηριστικό παράδειγμα. Πρόκειται για ένα πλαίσιο καθοδηγούμενο από τα δεδομένα (*data-driven framework*), το οποίο συνδυάζει τη χρονική συσσωμάτωση (*temporal aggregation*) με μοντέλα μηχανικής μάθησης για τη βελτίωση της πρόβλεψης ηλεκτρονικών εξαρτημάτων ενισχύοντας την ανθεκτικότητα της εφοδιαστικής αλυσίδας. Οι συγγραφείς κατέδειξαν ότι η συγκέντρωση των δεδομένων σε ευρύτερες χρονικές θυρίδες (*time buckets*), όπως, για παράδειγμα, η μετάβαση από την ημερήσια στην εβδομαδιαία ή μηνιαία κλίμακα, μειώνει δραστικά τον στοχαστικό θόρυβο και επιτρέπει στα ML μοντέλα να λειτουργήσουν με μεγαλύτερη αποτελεσματικότητα στο επίπεδο του ελέγχου των αποθεμάτων [45], [120].

Παράλληλα, οι Kiefer, Grimm και van Dinther [121] εξέτασαν σχήματα μεταφοράς γνώσης (*transfer learning*), όπου μοντέλα βαθιάς μάθησης εκπαιδεύονται σε ένα «πλούσιο» σύνολο δεδομένων από συσχετιζόμενες χρονοσειρές και, έπειτα, προσαρμόζονται σε σειρές με εξαιρετικά περιορισμένες παρατηρήσεις. Αυτή είναι μια πρακτική με υψηλή αξία για σπάνιους ή εξειδικευμένους κωδικούς ανταλλακτικών, γιατί επιτρέπει την επαναχρησιμοποίηση γνώσης μεταξύ παρεμφερών προϊόντων ή κατηγοριών μειώνοντας την ανάγκη για μεγάλα ιστορικά δείγματα ανά κωδικό. Επίσης, οι Giannopoulos, Dasaklis και Rachaniotis [122] πρότειναν ένα πλαίσιο βελτιστοποίησης με χρήση του k-NN αλγορίθμου σε συνθήκες υψηλής σπανιότητας ενσωματώνοντας μηχανισμούς επιλογής τόσο του μήκους του διανύσματος ζήτησης όσο και του πλήθους των γειτονικών παραδειγμάτων μέσω τεχνικής αναζήτησης πλέγματος (*grid-search*). Ο στόχος είναι η επίλυση του προβλήματος των στρεβλών αποστάσεων (*distorted distances*) που προκαλεί η υπερβολική παρουσία μηδενικών.

Τέλος, η υιοθέτηση σχημάτων συσσωρευτικής γενίκευσης (*stacked generalization – stacking*) και συνάθροισης με αναδειγματοληψία (*bootstrap aggregating – bagging*) με τη χρήση γραμμικών ή μη γραμμικών μοντέλων μετα-μάθησης (*meta-learners*) για τον συνδυασμό στατιστικών μοντέλων και μοντέλων Μηχανικής Μάθησης έχει βρει εκτεταμένη εφαρμογή στις σύγχρονες ροές εργασιών (*pipelines*) πρόβλεψης διαλείπουσας ζήτησης. Οι προσεγγίσεις αυτές, οι οποίες θεμελιώθηκαν θεωρητικά από τις κλασικές εργασίες των Wolpert [58] και Dietterich [123] και προσφέρουν αυξημένη σταθερότητα και ελαχιστοποίηση της διακύμανσης του σφάλματος θωρακίζοντας το σύστημα έναντι του στοχαστικού θορύβου που χαρακτηρίζει τα σποραδικά δεδομένα [13].

3.3.6 Συγκριτική αποτίμηση

Συνολικά, η διεθνής βιβλιογραφία συγκλίνει στο συμπέρασμα ότι **δεν υφίσταται μία καθολικά κυρίαρχη μέθοδος Μηχανικής Μάθησης** για όλα τα προφίλ διαλείπουσας ζήτησης. Η απόδοση κάθε αλγορίθμου εξαρτάται άμεσα από τα ιδιαίτερα γεωμετρικά χαρακτηριστικά της ζήτησης (διαλείπουσα έναντι ακανόνιστη), τον επιδιωκόμενο ορίζοντα πρόβλεψης, τη διαθεσιμότητα εξωγενών μεταβλητών και το μέγεθος του συνόλου των δεδομένων. Οι βαθιές αρχιτεκτονικές ακολουθιών (RNN/LSTM/Transformers) εμφανίζονται ιδιαίτερα ισχυρές όταν υπάρχει μεγάλος όγκος δεδομένων και πλούσιο πληροφοριακό περιεχόμενο. Από την άλλη, τα Δέντρα Διαβαθμισμένης Ενίσχυσης (GBT), τα υβριδικά σχήματα και οι συνδυασμοί μοντέλων (*ensembles*) προσφέρουν τη βέλτιστη ισορροπία μεταξύ ακρίβειας, υπολογιστικής ταχύτητας και ερμηνευσιμότητας σε βιομηχανικά περιβάλλοντα [3], [45], [114].

Σε κάθε περίπτωση, οι σύγχρονες ανασκοπήσεις υπογραμμίζουν με έμφαση ότι η αξιολόγηση των ML μεθόδων πρέπει να εκτελείται ταυτόχρονα με στατιστικές μετρικές σφάλματος (*forecasting metrics*) και επιχειρησιακούς δείκτες αποθεμάτων (*inventory* KPIs), καθότι η βελτίωση στη μία διάσταση δεν

συνεπάγεται νομοτελειακά βελτίωση και στην άλλη [82], [95], [97]. Αυτή η διαπίστωση αποτελεί τον πυρήνα της παρούσας διπλωματικής εργασίας, όπου ο τελικός στόχος είναι η βέλτιστη λήψη αποφάσεων αναπλήρωσης στη δευτερογενή αγορά αυτοκινήτου.

3.4 Μοντέλα Εμποδίου (*Hurdle models*), μηχανική χαρακτηριστικών και αρχιτεκτονική πολυπλοκότητα

Μια διακριτή και ιδιαίτερα ισχυρή κατηγορία μεθοδολογικών προσεγγίσεων για την αντιμετώπιση της διαλείπουσας ζήτησης προέρχεται από την οικογένεια των μοντέλων Εμποδίου (*Hurdle models*) και των μοντέλων διογκωμένου μηδενός (*zero-inflated models*). Στις αρχιτεκτονικές αυτές, η εκτίμηση της ζήτησης διαχωρίζεται ρητά σε δύο διαδοχικά στάδια. Αρχικά εκτιμάται η πιθανότητα εμφάνισης θετικής ζήτησης, $\mathbb{P}(y > 0)$, και εν συνεχεία, υπό τη συνθήκη ότι η ζήτηση είναι υπαρκτή, εκτιμάται το αναμενόμενο μέγεθός της, $\mathbb{E}[\, y \mid y > 0\,]$. Η θεωρητική θεμελίωση αυτής της προσέγγισης ανάγεται στις κλασικές οικονομετρικές μελέτες του Cragg [71] σχετικά με τα μοντέλα διπλού Εμποδίου (*double-*Hurdle *models*) και του Mullahy [70] σχετικά με τις διμερείς διατυπώσεις (*two-part formulations*) και τις κατανομές Poisson διογκωμένου μηδενός (*zero-inflated Poisson*). Η σύνδεση των υποδειγμάτων αυτών με το πεδίο της διαλείπουσας ζήτησης είναι άμεση και φυσική, καθώς η δομή του προβλήματος, δηλαδή εκτεταμένες περίοδοι μηδενικής δραστηριότητας που διακόπτονται σποραδικά από θετικά γεγονότα μεταβλητής έντασης, αντιστοιχεί πλήρως στη φιλοσοφία του διαχωρισμού των δύο αυτών στοχαστικών διαδικασιών.

3.4.1 Σύνδεση μοντέλων εμποδίου με κλασικές μεθόδους τύπου Croston

Υπό το πρίσμα αυτό, η μέθοδος TSB των Teunter *et al.* [2], η οποία αναλύθηκε στην ενότητα 3.2, αποτελεί στην πραγματικότητα ένα υπόδειγμα εμποδίου (*hurdle scheme*) ενσωματωμένο σε ένα πλαίσιο εκθετικής εξομάλυνσης, καθώς ενημερώνει ανεξάρτητα την πιθανότητα εμφάνισης και το υπό συνθήκη μέγεθος της ζήτησης, ακολουθώντας τη δομή του Cragg (1971). Παρομοίως, το πλαίσιο βαθέων διαδικασιών ανανέωσης (*deep renewal framework*) των Türkmen *et al.* [72] μοντελοποιεί από κοινού την εμφάνιση ζήτησης και τα εκάστοτε μεγέθη αυτής αποδεικνύοντας ότι οι παραδοσιακές στατιστικές προσεγγίσεις (Croston, SBA, TSB) συνιστούν ειδικές περιπτώσεις ενός ενιαίου πιθανοκρατικού μοντέλου εμποδίου.

Στη σύγχρονη βιβλιογραφία της Μηχανικής Μάθησης, ο διαχωρισμός αυτός αξιοποιείται με δύο προσεγγίσεις. Είτε άμεσα, μέσω της παράλληλης ή διαδοχικής εκπαίδευσης δύο διακριτών αλγορίθμων, είτε έμμεσα, μέσω της χρήσης εξειδικευμένων πιθανοκρατικών κατανομών που ενσωματώνουν τη διπλή αυτή φύση, όπως η κατανομή Poisson, η Αρνητική Διωνυμική κατανομή διογκωμένου μηδενός (Zero-Inflated Negative Binomial – ZINB), καθώς, και η κατανομή Tweedie [45][73].

Ενδεικτικά, οι Rožanec, Fortuna και Mladenić [124] πρότειναν ένα σχήμα δύο σταδίων, επιστρατεύοντας έναν ταξινομητή CatBoost (CatBoost *classifier*) για την πρόβλεψη της εμφάνισης ζήτησης και έναν παλινδρομητή LightGBM (LightGBM *regressor*) για την εκτίμηση του μεγέθους της επιτυγχάνοντας υψηλή ακρίβεια σε εξαιρετικά σποραδικούς κωδικούς λιανεμπορίου. Επίσης, οι Chien, Ku και Lu [13] ανέπτυξαν μια αρχιτεκτονική συσσωρευτικής γενίκευσης (*stacked generalization – stacking*), συνδυάζοντας τις μεθόδους SBA, Holt-Winters, ARIMA και XGBoost μέσω ενός γραμμικού μοντέλου μετα-μάθησης (*meta-learner*) για εφαρμογές σε ανταλλακτικά αυτοκινήτων.

Η διεθνής έρευνα έχει τεκμηριώσει ότι οι δομές εμποδίου είναι ιδανικές για σποραδικά και διαλείποντα δεδομένα, διότι επιτρέπουν τη ρητή αναγνώριση και αυτόνομη διαχείριση των δύο υποκατηγοριών πληθυσμού (μηδενικών και θετικών παρατηρήσεων). Έτσι, διευκολύνουν την αντιμετώπιση της έντονης

ανισορροπίας κλάσεων (*class imbalance*) μέσω της χρήσης βαρών στη θετική κλάση (*positive class weights*) ή μέσω μάθησης ευαίσθητης στο κόστος (*cost-sensitive learning*) [45], [98].

3.4.2 Σύγχρονες εξειδικευμένες αρχιτεκτονικές: Το πλαίσιο «Switch-Hurdle»

Στην αιχμή της πρόσφατης βιβλιογραφίας για την πιθανοκρατική πρόβλεψη της διαλείπουσας ζήτησης, οι Mușat και Căbuz [111] εισάγουν το εξελιγμένο πλαίσιο «Switch-Hurdle». Η συγκεκριμένη προσέγγιση γεφυρώνει την αρχιτεκτονική πολυπλοκότητα των σύγχρονων μοντέλων τύπου *Transformer* με τη στατιστική στιβαρότητα των υποδειγμάτων Εμποδίου (Hurdle *models*) επιλύοντας το πρόβλημα της μαθηματικής παράλυσης που προκαλεί η υπερεκπροσώπηση των μηδενικών τιμών (*zero-inflation*).

Η δομή του μοντέλου Switch-Hurdle είναι αρθρωτή και αποτελείται από τρία διακριτά στάδια, έναν κωδικοποιητή, έναν αποκωδικοποιητή και μια κοινή κεφαλή εμποδίου που αναλαμβάνει να κάνει τις προβλέψεις. Ο κωδικοποιητής του συστήματος απομακρύνεται από τη συμβατική λογική των «πυκνών» επιπέδων και ενσωματώνει ένα πλαίσιο Μίγματος Ειδικών (*Mixture-of-Experts* – MoE). Οπότε, αντί για την καθολική επεξεργασία όλων των δεδομένων από ένα ενιαίο, ογκώδες δίκτυο, η πληροφορία δρομολογείται σε E ανεξάρτητους «Ειδικούς». Το μοντέλο εφαρμόζει μια στρατηγική αραιής δρομολόγησης κορυφαίου ειδικού (*Top-1 sparse routing*) κατά την πρόσθια διέλευση (*forward pass*), δηλαδή ένας γραμμικός δρομολογητής (*router*) υπολογίζει τις πιθανότητες κατανομής και κατευθύνει κάθε δομικό στοιχείο της πληροφορίας (*token*) αποκλειστικά στον έναν, πλέον κατάλληλο ειδικό, εκμηδενίζοντας το άσκοπο υπολογιστικό κόστος. Από την άλλη, κατά την εκπαίδευση/οπισθοσκέδαση (*backward pass*) χρησιμοποιεί πυκνή δρομολόγηση, ώστε να μαθαίνουν όλοι οι ειδικοί και να εξασφαλίζεται η σταθερότητα της εκπαίδευσης. Έπειτα, ο αποκωδικοποιητής λαμβάνει τα συμπεράσματα των ειδικών και λειτουργεί ως αυτοπαλινδρομητής (*autoregressor*). Δηλαδή χρησιμοποιεί την αμέσως προηγούμενη κατάστασή του για να προβλέψει την επόμενη. Τέλος, η παραγόμενη κατάσταση του αποκωδικοποιητή τροφοδοτείται σε μια κοινή κεφαλή εμποδίου (*shared hurdle head*), η οποία διαχωρίζει αυστηρά τη στοχαστική διαδικασία σε δύο συνιστώσες, τη δυαδική συνιστώσα (*binary component*), που εκτιμά την πιθανότητα εμφάνισης θετικής ζήτησης p_t^+, και την υπό συνθήκη συνιστώσα (*conditional component*), που χρησιμοποιεί μια αρνητική διωνυμική κατανομή για την πρόβλεψη της ποσότητας.

Η εμπειρική αξιολόγηση του Switch-Hurdle στον διαγωνισμό αναφοράς M5 και σε μεγάλης κλίμακας πραγματικά δεδομένα της πλατφόρμας eMAG κατέδειξε σαφή υπεροχή και μείωση του δείκτη WRMSSE έναντι ανταγωνιστικών αρχιτεκτονικών βασισμένων σε Transformers (όπως τα μοντέλα TFT και PatchTST) και απέδειξε ότι η σύγχρονη αρχιτεκτονική πολυπλοκότητα αποδίδει τα μέγιστα, όταν ευθυγραμμίζεται με τη στατιστική φύση της σποραδικότητας [111].

3.4.3 Μηχανική χαρακτηριστικών έναντι αρχιτεκτονικής πολυπλοκότητας: «Το πλαίσιο SHOS»

Ένα κρίσιμο ερώτημα, τόσο από θεωρητικής όσο και από πρακτικής σκοπιάς, που απασχολεί τη σύγχρονη έρευνα είναι κατά πόσο η αύξηση της αρχιτεκτονικής πολυπλοκότητας δικαιολογείται επιχειρησιακά, ή είναι προτιμότερος ο εμπλουτισμός των μοντέλων με κατάλληλα σχεδιασμένα χαρακτηριστικά γνωρίσματα (*features*). Οι Nathan *et al.* [98] τάσσονται σθεναρά υπέρ της υπεροχής της μηχανικής χαρακτηριστικών (*primacy of feature engineering*) και προτείνουν το πλαίσιο «Smoothed Hybrid Occurrence-Size (SHOS)», το οποίο αξιολόγησαν σε ένα μεγάλης κλίμακας σύνολο δεδομένων της δευτερογενούς αγοράς αυτοκινήτου.

Το πλαίσιο SHOS συνιστά μια επέκταση της λογικής TSB, ειδικά σχεδιασμένη για συνθήκες υψηλής σποραδικότητας, η οποία παράγει δύο εξειδικευμένες ανά χρονοσειρά (*series-specific*) εκτιμήσεις, την πιθανότητα εμφάνισης ζήτησης ανά περίοδο και το μέσο μέγεθος αυτής υπό τη συνθήκη εμφάνισης. Αντί οι εκτιμήσεις αυτές να αποτελέσουν το τελικό αποτέλεσμα, ενσωματώνονται ως πρόσθετα μετα-χαρακτηριστικά (*meta-features*) σε ένα μοντέλο μηχανικής μάθησης ενός σταδίου (όπως το LightGBM ή το XGBoost), το οποίο εκπαιδεύεται στο σύνολο των δεδομένων. Η βασική καινοτομία έγκειται στο ότι τα χαρακτηριστικά SHOS κωδικοποιούν τοπικά πρότυπα σποραδικότητας και τη δυναμική επαναδραστηριοποίησης (*reactivation dynamics*) των ανενεργών προϊόντων. Αυτά αποτελούν πληροφορία, την οποία τα δέντρα διαβαθμισμένης ενίσχυσης (GBT) αδυνατούν να εξάγουν αυτόματα από τις πρωτογενείς ιστορικές υστερήσεις (*raw lags*).

Το πλέον αξιοσημείωτο εύρημα της μελέτης τους είναι ότι το μοντέλο ενός σταδίου με ενσωματωμένα τα χαρακτηριστικά SHOS επέδειξε ισοδύναμη ή και ανώτερη απόδοση σε σύγκριση με πολύπλοκα μοντέλα Εμποδίου δύο σταδίων (*two-stage* Hurdle *models*), υπογραμμίζοντας ότι η ποιότητα της αναπαράστασης των δεδομένων υπερτερεί συχνά της αρχιτεκτονικής πολυπλοκότητας [98]. Η τάση αυτή προσφέρει μια ισχυρή επιστημονική νομιμοποίηση για την ανάπτυξη ροών εργασίας (*pipelines*) που επενδύουν συστηματικά στην κατασκευή προηγμένων χαρακτηριστικών (όπως *rolling statistics*, δείκτες ADI/CV^2 κ.ά.).

3.4.4 Έλλειψη αποθέματος και αποκομμένη/λογοκριμένη ζήτηση (*censored demand*) ως αντικείμενο μηχανικής χαρακτηριστικών

Ένα εξαιρετικά κρίσιμο, αλλά συστηματικά παραμελημένο, ζήτημα στη βιβλιογραφία, αφορά τη διαφοροποίηση μεταξύ των καταγεγραμμένων ιστορικών πωλήσεων και της πραγματικής ζήτησης. Σε περιπτώσεις, κατά τις οποίες ένας κωδικός παρουσιάζει έλλειψη αποθέματος (*stock-out*), η καταγεγραμμένη πώληση είναι εύλογα μηδενική. Αυτό, όμως δεν συνεπάγεται και μηδενική ζήτηση. Είναι πιθανό, η πραγματική ζήτηση της αγοράς να παραμένει θετική αλλά να μην μπορεί να ικανοποιηθεί. Τα δεδομένα αυτά χαρακτηρίζονται στη στατιστική ως λογοκριμένα (*censored data*) και η αφελής εκπαίδευση μοντέλων Μηχανικής Μάθησης πάνω σε αυτά εισάγει ένα έντονο συστηματικό σφάλμα υποεκτίμησης (*under-forecasting bias*), ιδιαιτέρως, όταν πρόκειται για προϊόντα με συχνές ελλείψεις [45], [100].

Η βιβλιογραφία αναγνωρίζει ρητά ότι η ενσωμάτωση δεικτών έλλειψης αποθέματος (*Out-Of-Stock* (OOS) *flags*) και μεταβλητών διαθεσιμότητας στο ράφι (*shelf-availability*) αποτελεί ένα από τα σημαντικότερα ανοικτά ζητήματα της σύγχρονης έρευνας [96]. Πρόσφατες μελέτες προτείνουν διάφορες τεχνικές για την αποκατάσταση των περιορισμών της ζήτησης (*unconstraining demand*). Αυτές ποικίλουν από απλούς καταλογισμούς τιμών (*imputations*) και παλινδρομήσεις ευαίσθητες στη λογοκρισία (*censoring-aware regression*) [6], [7], έως τη ρητή εισαγωγή δυαδικών μεταβλητών ένδειξης έλλειψης αποθέματος (*stock-out indicators*) κατά τη φάση της εκπαίδευσης [125]. Στο ίδιο πλαίσιο, οι Svetunkov και Sroginis [8] ανέπτυξαν μια προσέγγιση βασισμένη σε μοντέλα (*model-based approach*) για την ποιοτική ταξινόμηση των μηδενικών παρατηρήσεων. Τα διακρίνουν σε μηδενικά που οφείλονται στην εγγενή σποραδικότητα και σε εκείνα που προκύπτουν λόγω έλλειψης αποθέματος, αποδεικνύοντας εμπειρικά ότι η διάκριση αυτή βελτιώνει θεαματικά την ικανότητα γενίκευσης των μοντέλων Μηχανικής Μάθησης σε πραγματικά δεδομένα που αντλούνται από λογιστικά συστήματα (ERP).

3.4.5 Σύνοψη περιορισμών και ερευνητικά κενά

Παρά την αδιαμφισβήτητη θεωρητική τους συνεισφορά και το γεγονός ότι επιβεβαιώνουν τις δομικές επιλογές της παρούσας μελέτης, οι πλέον σύγχρονες προσεγγίσεις (συμπεριλαμβανομένων των Switch-Hurdle και SHOS) παρουσιάζουν συγκεκριμένους περιορισμούς, όταν εξετάζονται υπό το πρίσμα της ρεαλιστικής βιομηχανικής εφαρμογής. Στη συντριπτική τους πλειονότητα, τα μοντέλα αυτά περιορίζονται σε βραχυπρόθεσμους ορίζοντες πρόβλεψης (συνήθως μίας περιόδου μπροστά – *single-step ahead*), δεν ενσωματώνουν συστηματικά την πληροφορία της λογοκριμένης ζήτησης (*censored demand/stock-outs*) στην παραγωγική τους ροή. Επίσης, σπανίως επεκτείνονται στον κρίσιμο υπολογισμό επιχειρησιακών παραμέτρων, όπως τα δυναμικά αποθέματα ασφαλείας (*safety stocks*).

Κατά συνέπεια, παραμένει ανοιχτή η επιστημονική πρόκληση για την ανάπτυξη μιας ενιαίας αρχιτεκτονικής εμποδίου (*end-to-end true forecast pipeline*), η οποία αφενός θα ενσωματώνει τα πλεονεκτήματα της προηγμένης μηχανικής χαρακτηριστικών και των μοντέλων Μίγματος Ειδικών, και αφετέρου θα τα γεφυρώνει άμεσα με τη Διαχείριση Αποθεμάτων μεγάλης κλίμακας σε πολλαπλούς χρονικούς ορίζοντες. Ακριβώς σε αυτό το θεωρητικό και εφαρμοσμένο κενό τοποθετείται η παρούσα διπλωματική εργασία σχεδιάζοντας ένα σύστημα που αποκαθιστά τους περιορισμούς της ζήτησης και εξάγει άμεσα αξιοποιήσιμες πολιτικές αποθεμάτων.

3.5 Πιθανοκρατική πρόβλεψη και διαγωνισμοί αναφοράς

Κατά τα τελευταία έτη, η επιστημονική συζήτηση γύρω από τη διαλείπουσα ζήτηση έχει μετατοπιστεί αποφασιστικά από τις καθαρές σημειακές προβλέψεις (*point forecasts*) προς τις πλήρεις πιθανοκρατικές προσεγγίσεις (*probabilistic forecasting*). Στο νέο αυτό παράδειγμα, ο αντικειμενικός στόχος δεν περιορίζεται στην εκτίμηση της αναμενόμενης μέσης τιμής, αλλά επεκτείνεται στην πλήρη μοντελοποίηση της υποκείμενης κατανομής πιθανότητας της μελλοντικής ζήτησης. Όπως υπογραμμίζουν οι συστηματικές ανασκοπήσεις των Giannopoulos *et al.* [45] και Pinçe *et al.* [96], η επιχειρησιακή λήψη αποφάσεων σε περιβάλλοντα εφοδιαστικής αλυσίδας ανταλλακτικών και λιανεμπορίου, στα οποία καθορίζονται τα αποθέματα ασφαλείας (*safety stocks*), τα επιδιωκόμενα επίπεδα εξυπηρέτησης (*service levels*) και οι ανεκτές ανεκπλήρωτες παραγγελίες (*backorders*), εξαρτάται άμεσα από τη συμπεριφορά των ουρών (*tails*) της κατανομής και όχι από έναν μοναδικό σημειακό εκτιμητή.

Παράλληλα, η ραγδαία ανάπτυξη των μοντέλων βαθιάς πιθανοκρατικής πρόβλεψης μέσω μάθησης (*deep probabilistic forecasting*) έχει προσφέρει ευέλικτα υπολογιστικά εργαλεία για τη διαχείριση αυτών των προβλημάτων. Πλαίσια όπως το «DeepAR» [46], καθώς, και προηγμένες αρχιτεκτονικές αναδρομικών δικτύων βασισμένες σε συναρτήσεις εκατοστημορίων τύπου spline (*spline quantile functions*) [126] επιτρέπουν την άμεση εξαγωγή προβλέψεων εκατοστημορίων (*quantile forecasts*) και πυκνοτήτων πιθανότητας (probability densities) σε δεδομένα με σύνθετη στοχαστική δομή.

Σε αυτό το πλαίσιο, οι παγκόσμιοι διαγωνισμοί forecasting της σειράς M (*Makridakis competitions*) λειτούργησαν ως καταλύτες για την αντικειμενική αποτίμηση και σύγκριση στατιστικών υποδειγμάτων και αλγορίθμων Μηχανικής και Βαθιάς Μάθησης σε μεγάλης κλίμακας χρονοσειρές. Ο διαγωνισμός M4 [104], ο οποίος περιλάμβανε 100.000 χρονοσειρές και εξέτασε 61 μεθοδολογίες, κατέδειξε ότι τα συνδυαστικά σύνολα μοντέλων (*ensembles*) στατιστικών προσεγγίσεων υπερέχουν κατά μέσο όρο των μεμονωμένων αλγορίθμων Μηχανικής Μάθησης. Η διαπίστωση αυτή ανέδειξε τη σπουδαιότητα της αρχιτεκτονικής συνδυασμού έναντι της επιλογής ενός και μόνο πολύπλοκου μοντέλου. Η νικήτρια λύση του Smyl [127] συνδύαζε την κλασική εκθετική εξομάλυνση με δίκτυα Μακράς Βραχυπρόθεσμης Μνήμης (LSTM) θεμελιώνοντας την αξία των υβριδικών σχημάτων.

Ο μεταγενέστερος διαγωνισμός M5 [21] εξειδικεύτηκε ακόμη περισσότερο στο πρακτικό περιβάλλον του λιανεμπορίου χρησιμοποιώντας τις ιεραρχικές χρονοσειρές πωλήσεων της Walmart. Η σημαντικότερη συνεισφορά του M5 για το πεδίο της διαλείπουσας ζήτησης ήταν η καθιέρωση της κλιμακωτής μετρικής του Ριζικού Μέσου Τετραγωνικού Κλιμακωτού Σφάλματος (*Root Mean Squared Scaled Error* – RMSSE) και της σταθμισμένης εκδοχής της (WRMSSE) ως πρωτευόντων κριτηρίων αξιολόγησης. Τα συμπεράσματα του διαγωνισμού απέδειξαν ότι οι σύνθετες συνθέσεις μοντέλων, στις οποίες κυριαρχεί η χρήση δέντρων διαβαθμισμένης ενίσχυσης (*Gradient Boosted Trees* – GBT μέσω του LightGBM), επιτυγχάνουν θεαματικά κέρδη ακρίβειας σε επίπεδα του καταλόγου, όπου η ζήτηση εμφανίζει ακραία σποραδικότητα ή υψηλή μεταβλητότητα [21], [114]. Επακόλουθες μελέτες πάνω στα δεδομένα του M5 επιβεβαίωσαν ότι η ακρίβεια των προβλέψεων εκατοστημορίων διαφοροποιείται ανάλογα με το εξεταζόμενο επίπεδο. Με άλλα λόγια, ένα μοντέλο μπορεί να εμφανίζει βέλτιστη συμπεριφορά για τη διάμεσο (50ο εκατοστημόριο) αλλά να υστερεί στις ακραίες τιμές (10ο και 90ο εκατοστημόριο). Αυτό αποτελεί κρίσιμο στοιχείο για τον ορθό σχεδιασμό των αποθεμάτων ασφαλείας [128].

Η συμβολή των τελευταίων (Spiliotis και συνεργατών) υπήρξε καθοριστική για τη σύνδεση της πιθανοκρατικής θεωρίας με την επιχειρησιακή πράξη. Μέσα από μια σειρά συνδυαστικών μελετών [128], [129] τεκμηριώθηκε ότι η επιλογή προσαρμοσμένων εκατοστημορίων, καθώς, και η συνάρθροιση ποσοστιαίων προβλέψεων από ετερογενή μοντέλα, δύναται να βελτιστοποιήσει απ' ευθείας το συνολικό κόστος διατήρησης αποθεμάτων και να επιτύχει με ακρίβεια το επιθυμητό επίπεδο εξυπηρέτησης. Η προσέγγιση αυτή υπερέχει των απλών σημειακών προβλέψεων που βασίζονται σε αυθαίρετους, σταθερούς κανόνες ασφαλείας (*safety-factor heuristics*). Η στροφή αυτή ενισχύει την ανάπτυξη «επιχειρησιακά ενσυνείδητων» πιθανοκρατικών μοντέλων, κατά τα οποία η εκτίμηση των κατανομών στοχεύει στην άμεση τροφοδότηση στοχαστικών συστημάτων ελέγχου αποθεμάτων [72], [91].

Σε περιβάλλοντα με έντονη σποραδικότητα, όπως η δευτερογενής αγορά ανταλλακτικών αυτοκινήτων, η μετάβαση προς την πιθανοκρατική οπτική καθίσταται επιτακτική για την επαρκή διαχείριση του κινδύνου ελλείψεων (*stock-out risk*) και της απαξίωσης των κωδικών ειδών. Σύγχρονα πλαίσια, όπως το ενοποιημένο μοντέλο βαθιάς ανανέωσης (*unified deep renewal model*) των Türkmen *et al.* [72] και τα συνδυαστικά μοντέλα πυκνότητας πιθανότητας (*ensemble density models*) των Wang *et al.* [91], αποδεικνύουν ότι η από κοινού μοντελοποίηση της εμφάνισης ζήτησης και του μεγέθους αυτής επιτρέπει τον άμεσο υπολογισμό συναρτήσεων απώλειας (*loss functions*) που συνδέονται μαθηματικά με το κόστος διατήρησης αποθεμάτων (*holding costs*) και τις ελλείψεις. Επιπροσθέτως, η βιβλιογραφία της διαχείρισης αποθεμάτων (*inventory management*) μέσα από κλασικές εργασίες όπως των Ghobbar και Friend [130], Syntetos *et al.* [95] και Babai *et al.* [97], ενισχύει τη θέση ότι η αξιολόγηση των συστημάτων πρόβλεψης πρέπει να εκτελείται μέσω ενός διττού πλαισίου, στο οποίο συνδυάζονται κλιμακωτά στατιστικά σφάλματα πρόβλεψης (*forecasting errors*), όπως οι δείκτες MASE και RMSSE, με αμιγώς επιχειρησιακούς δείκτες αποθεμάτων (*inventory-oriented* KPIs), όπως το ποσοστό κάλυψης ζήτησης (*fill rate*) και η πιθανότητα έλλειψης αποθέματος (*stock-out probability*).

Από την άλλη, οι κλιμακωτές μετρικές που καθιερώθηκαν από τους διαγωνισμούς M έχουν υποστεί αυστηρή κριτική ανάλυση από τον Kolassa [14], [15] και τους Syntetos, Nikolopoulos και Boylan [131]. Οι συγκεκριμένες μελέτες αναδεικνύουν ότι η επιλογή της μετρικής επηρεάζει βαθύτατα τη συμπεριφορά του μοντέλου. Δηλαδή, μια μετρική βασισμένη στην απόλυτη απόκλιση (τύπου L_1, όπως η WAPE ή η MASE) εξαναγκάζει το μοντέλο σε προβλέψεις που στοχεύουν στη διάμεσο (*median-targeted*), ενώ μια μετρική βασισμένη στο τετραγωνικό σφάλμα (τύπου L_2, όπως η RMSSE) οδηγεί σε προβλέψεις που στοχεύουν στη μέση τιμή (*mean-targeted*). Σε συνθήκες σποραδικότητας, οι δύο αυτές

μαθηματικές ιδιότητες παράγουν ριζικά διαφορετικά αποτελέσματα, διότι η βέλτιστη πρόβλεψη υπό την έννοια της διαμέσου είναι συχνά το μηδέν. Κατά συνέπεια, η επιλογή της μετρικής και του αντίστοιχου στόχου του μοντέλου πρέπει να καθοδηγείται από το πραγματικό οικονομικό κόστος που επιφέρει κάθε τύπος σφάλματος στο σύστημα αποθεμάτων της επιχείρησης [3], [14], [15].

3.6 Μεγάλα Γλωσσικά Μοντέλα και Υπολογισμός Δεξαμενής για διαλείπουσα ζήτηση

Οι πλέον πρόσφατες εξελίξεις στο πεδίο της Βαθιάς Μάθησης έχουν οδηγήσει σε δύο νέες, αλλά πολλά υποσχόμενες, ερευνητικές κατευθύνσεις για τη διαχείριση της διαλείπουσας ζήτησης. Αυτές είναι η χρήση Μεγάλων Γλωσσικών Μοντέλων (Large Language Models – LLMs) ως γενικοί μηχανισμοί διαδοχικής πρόβλεψης και η αξιοποίηση της αρχιτεκτονικής της Υπολογιστικής Δεξαμενής (Reservoir Computing – RC) με αιχμή τα Δίκτυα Κατάστασης Ηχούς (Echo State Networks – ESNs). Και οι δύο αυτές κατευθύνσεις επανεξετάζουν σε θεμελιώδες επίπεδο το ερώτημα του ποια αρχιτεκτονική δομή είναι η πλέον κατάλληλη για εξαιρετικά αραιές χρονοσειρές, οι οποίες χαρακτηρίζονται από ακραία διόγκωση μηδενικών παρατηρήσεων (*zero-inflation*).

3.6.1 Μεγάλα Γλωσσικά Μοντέλα για πρόβλεψη χρονοσειρών

Το βασικό μεθοδολογικό κίνητρο για την εφαρμογή των LLMs σε προβλήματα χρονοσειρών έγκειται στην παραδοχή ότι τόσο τα κείμενα όσο και οι χρονοσειρές μπορούν να αναπαρασταθούν ως διακριτές ακολουθίες (*discrete sequences*) δομικών στοιχείων (*tokens*), όπου η παραγωγή κάθε επόμενου στοιχείου εξαρτάται άμεσα από το ιστορικό παρελθόν. Τα θεμελιώδη μοντέλα Transformers [83] απέδειξαν μια εξαιρετική ικανότητα αποτύπωσης μακροχρόνιων εξαρτήσεων (*long-range dependencies*) μέσω μηχανισμών αυτο-προσοχής (*self-attention*) και αποτέλεσαν τη βάση για σύγχρονες αρχιτεκτονικές που βασίζονται αποκλειστικά σε αποκωδικοποιητές (*decoder-only*, π.χ. GPT), αποκλειστικά σε κωδικοποιητές (*encoder-only*, π.χ. BERT) ή σε συνδυασμούς κωδικοποιητή-αποκωδικοποιητή (*encoder-decoder*, π.χ. T5) [132].

Πρόσφατες εργασίες, όπως το πλαίσιο «*AutoTimes*» [133], προτείνουν την επαναχρησιμοποίηση ενός *decoder-only* LLM ως αυτο-αναδρομικού συστήματος πρόβλεψης. Σε αυτή την προσέγγιση, οι αριθμητικές χρονοσειρές χαρτογραφούνται στον χώρο διανυσματικών αναπαραστάσεων (*embedding space*) των γλωσσικών tokens και το μοντέλο παράγει πολυβηματικές προβλέψεις αξιοποιώντας τον ίδιο ακριβώς μηχανισμό πρόβλεψης επόμενου στοιχείου (*next-token prediction*) που χρησιμοποιείται κατά την παραγωγή φυσικού κειμένου. Η μέθοδος αυτή εκμεταλλεύεται την εγγενή ικανότητα των LLMs να μοντελοποιούν πολύπλοκες, μη γραμμικές εξαρτήσεις και προσφέρει μοναδική ευελιξία ως προς το μήκος του παραθύρου αναδρομής (*lookback window*) και τον επιθυμητό ορίζοντα πρόβλεψης.

Άλλες ερευνητικές προσπάθειες εστιάζουν σε στοχευμένες τεχνικές προσαρμογής. Το πλαίσιο «*TEST*» των Sun *et al.* [134] εισάγει ενσωματώσεις ευθυγραμμισμένες με γλωσσικά πρότυπα (*text-prototype aligned* embeddings) που «ενεργοποιούν» τις ικανότητες πρόβλεψης των LLMs μέσω αντιθετικής μάθησης (*contrastive learning*). Αντίστοιχα, οι Bian *et al.* [135] προτείνουν την πρόβλεψη πολλαπλών τμημάτων (*multi-patch prediction*), ενώ οι German-Morales *et al*. [136] διερευνούν τεχνικές μεταφοράς γνώσης (*transfer learning*) με μεθόδους προσαρμογής πινάκων χαμηλής τάξης (*Low-Rank Adaptations* – LoRA) για τη δραστική μείωση του υπολογιστικού κόστους της μικρο-ρύθμισης.

Ωστόσο, η διεθνής βιβλιογραφία διατηρεί επιφυλάξεις απέναντι στον ενθουσιασμό γύρω από τα LLMs στην πρόβλεψη χρονοσειρών. Οι Tan *et al.* [112], μέσα από συστηματικά πειράματα αφαίρεσης (*ablation studies*), απέδειξαν ότι σε αρκετές περιπτώσεις η πλήρης αφαίρεση της γλωσσικής συνιστώσας όχι μόνο δεν μειώνει, αλλά συχνά βελτιώνει την απόδοση των ολοκληρωμένων πλαισίων.

Το εύρημα αυτό υποδηλώνει ότι τα παρατηρούμενα οφέλη οφείλονται πρωτίστως στη γενική αρχιτεκτονική προσοχής (*attention mechanisms*) των *Transformers* και όχι στην ίδια τη γλωσσική προεκπαίδευση. Παρότι η παρατήρηση αυτή επαναπροσδιορίζει τον τρόπο αξιολόγησης των LLMs, δεν αναιρεί τη δυνητική τους αξία σε εξειδικευμένα (niche) σενάρια όπως η σποραδική ζήτηση.

Επί του προκειμένου, η εργασία των Tsantilis *et al.* [132] παρουσιάζει, εξ όσων γνωρίζουμε, την πρώτη εφαρμογή ενός *decoder-only* LLM, ειδικά σχεδιασμένου για την πρόβλεψη διαλείπουσας ζήτησης σε πραγματικό βιομηχανικό σύνολο δεδομένων. Το προτεινόμενο πλαίσιο, «GPT2ForForecasting», προσαρμόζει από την αρχή ένα ελαφρύ μοντέλο GPT-2 (περίπου 124 εκατ. παραμέτρων) εισάγοντας δύο καίριες καινοτομίες:

- **Στρώμα άμεσης ενσωμάτωσης (*Direct embedding*):** Ένα γραμμικό επίπεδο (`nn.Linear(1, 768)`) που χαρτογραφεί απευθείας τις αριθμητικές τιμές της χρονοσειράς στον 768-διάστατο χώρο αναπαραστάσεων του GPT-2 παρακάμπτοντας τη διαδικασία κερματισμού (*tokenization*) και διατηρώντας την αριθμητική εγγύτητα μεταξύ γειτονικών τιμών.
- **Κεφαλή διογκωμένου μηδενός (*Zero-inflated head*):** Ένα προαιρετικό τελικό επίπεδο που συνδυάζει τη Δυαδική Διασταυρούμενη Εντροπία (*Binary Cross-Entropy*) για την πιθανότητα μηδενικής ζήτησης με το Μέσο Τετραγωνικό Σφάλμα για το μέγεθος προσαρμοσμένο ρητά στη *zero-inflated* φύση των σποραδικών σειρών.

Η αξιολόγηση του συστήματος σε κωδικούς του κλάδου πλαστικών συσκευασιών (με ποσοστό μηδενικών 70–80%) κατέδειξε ότι το LLM μειώνει τον δείκτη RMSSE έως και 27% σε σύγκριση με κορυφαίες μεθόδους (Croston, TSB, LightGBM, LSTM).

3.6.2 Υπολογιστική Δεξαμενής και Δίκτυα Κατάστασης Ηχούς

Παράλληλα με τα γλωσσικά μοντέλα, η βιβλιογραφία έχει αναδείξει τον Υπολογισμό Δεξαμενής (Reservoir Computing – RC) και, ειδικότερα, τα Δίκτυα Κατάστασης Ηχούς (Echo State Networks – ESNs) ως μια εξαιρετικά ελαφριά αλλά υπολογιστικά ισχυρή εναλλακτική για την αποτύπωση πολύπλοκων χρονικών δυναμικών [81], **Error! Reference source not found.**. Ο πυρήνας της μεθοδολογίας έγκειται στη χρήση μιας «δεξαμενής» (*reservoir*) τυχαία αρχικοποιημένων και μη εκπαιδεύσιμων νευρώνων, η οποία λειτουργεί ως ένα δυναμικό σύστημα που προβάλλει τις εισόδους σε έναν χώρο καταστάσεων υψηλής διάστασης. Το μοναδικό τμήμα του δικτύου που εκπαιδεύεται είναι ένα απλό γραμμικό επίπεδο ανάγνωσης (*readout layer*) πάνω σε αυτές τις καταστάσεις, γεγονός που μειώνει δραστικά το υπολογιστικό κόστος και επιλύει τα προβλήματα εξαφάνισης της κλίσης (*vanishing gradients*) που ταλανίζουν τα παραδοσιακά RNN.

Η εφαρμογή του RC στην πρόβλεψη αραιής ζήτησης είναι πρωτοποριακή. Η μελέτη των Tsantilis, Giannopoulos, Dasaklis και Patsakis [138] αποτελεί μία από τις πρώτες στοχευμένες εφαρμογές ESNs σε δεδομένα με έντονη παρουσία μηδενικών και απρόβλεπτων αιχμών. Ακολουθώντας μια μεθοδολογία καθοδηγούμενη από την αναλυτική δεδομένων (*analytics-driven*), οι συγγραφείς προτείνουν, αρχικά, τον ενδελεχή στατιστικό χαρακτηρισμό των σειρών μέσω εξειδικευμένων ελέγχων (όπως το «Ljung–Box» για λευκό θόρυβο, το «*Augmented Dickey–Fuller*» για στασιμότητα και το «Lomb–Scargle periodogram» για εποχικότητα) και τον υπολογισμό δεικτών χαοτικής δυναμικής (όπως ο Μέγιστος Εκθέτης Lyapunov, ο Εκθέτης Hurst και η Εντροπία Shannon). Στη συνέχεια, τα χαρακτηριστικά αυτά καθοδηγούν τον σχεδιασμό των ESN ρυθμίζοντας κρίσιμες παραμέτρους όπως η φασματική ακτίνα (*spectral radius*) και η σποραδικότητα (*sparsity*) της δεξαμενής [138].

Σε συνδυασμό με παλαιότερες εργασίες της ως επί το πλείστον ίδιας ερευνητικής ομάδας [139] που τεκμηρίωσαν τη χρησιμότητα των χαοτικών δεικτών, τα ESN πλαίσια σχηματίζουν ένα ερευνητικό «δίπολο» σε σχέση με τα LLMs. Ενώ τα LLMs προσφέρουν την ισχύ της μαζικής προεκπαίδευσης και της αναπαράστασης, τα ESNs παρέχουν μια εξαιρετικά γρήγορη, μαθηματικά ερμηνεύσιμη και δυναμικά πλούσια λύση, ιδανική για βιομηχανικά περιβάλλοντα, στα οποία οι χρονοσειρές είναι μικρού μήκους αλλά ακραία μεταβλητές.

3.7 Σύζευξη πρόβλεψης και ελέγχου αποθεμάτων

Πέρα από τις αμιγώς στατιστικές ή αλγοριθμικές επιδόσεις, ένα σημαντικό και διαρκώς αυξανόμενο τμήμα της βιβλιογραφίας αξιολογεί τις μεθόδους πρόβλεψης της διαλείπουσας ζήτησης με βάση την *επιχειρησιακή τους επίδοση* σε προβλήματα ελέγχου αποθεμάτων. Η αξιολόγηση αυτή πραγματοποιείται μέσω αμιγώς οικονομικών δεικτών, όπως το συνολικό αναμενόμενο κόστος, το ποσοστό κάλυψης ζήτησης (*fill rate*), ο κίνδυνος ελλείψεων αποθέματος και το κόστος απαξίωσης αποθεμάτων (*obsolescence*).

Η θεμελιώδης παρατήρηση που διατρέχει αυτόν τον ερευνητικό κλάδο είναι ότι η πρόβλεψη δεν αποτελεί αυτοσκοπό, αλλά *εισροή* (*input*) σε μια πολιτική αναπλήρωσης, όπως η πολιτική παραγγελίας σταθερής ποσότητας σε σημείο αναπαραγγελίας *(Q, R)*, ή ο περιοδικός έλεγχος με παραγγελία έως το μέγιστο επίπεδο *(T, S)*. Συνεπώς, η τελική αξία της πρόβλεψης κρίνεται σε όρους συνολικού αναμενόμενου κόστους (*expected total cost*) ή επίτευξης επιθυμητού επιπέδου εξυπηρέτησης (*target service level*). Οι Babai, Syntetos και Teunter [140] αναδεικνύουν χαρακτηριστικά ότι μέθοδοι πρόβλεψης με σχεδόν πανομοιότυπο στατιστικό σφάλμα (*forecast error*) μπορούν να προκαλέσουν ριζικά διαφορετικές επιπτώσεις στους βασικούς δείκτες αποθέματος (*inventory* KPIs) καθιστώντας απολύτως αναγκαία τη συνεκτίμηση και των δύο διαστάσεων κατά την επιλογή μοντέλου.

3.7.1 Από την ακρίβεια πρόβλεψης στην ακρίβεια απόφασης

Οι Teunter, Syntetos και Babai [94] εξετάζουν πολιτικές περιοδικού ελέγχου αποθέματος για διαλείποντες κωδικούς και τεκμηριώνουν εμπειρικά την άρρηκτη διασύνδεση ανάμεσα στην ακρίβεια της πρόβλεψης και στο κόστος της πολιτικής αναπλήρωσης. Μικρές βελτιώσεις στο Μέσο Απόλυτο Σφάλμα (MAE) μπορεί να συνεπάγονται δυσανάλογα μεγάλα κέρδη σε επίπεδο κόστους, όταν το προφίλ της ζήτησης είναι δομικά ασύμμετρο. Ο Kourentzes [101] επεκτείνει αυτή την αρχή στο πλαίσιο των μεθόδων τύπου Croston αποδεικνύοντας ότι η βελτιστοποίηση των παραμέτρων εξομάλυνσης με κριτήριο τη διαχείριση αποθεμάτων (*inventory-aware optimization*) οδηγεί σε διαφορετικά, πιο αποδοτικά, επίπεδα αποθέματος σε σχέση με τη στατιστικά «βέλτιστη» ρύθμιση που βασίζεται σε κριτήρια MSE ή MAE. Οι Syntetos, Babai και Gardner [103] καταλήγουν σε αντίστοιχα συμπεράσματα συγκρίνοντας απλές παραμετρικές μεθόδους με σχήματα αναδειγματοληψίας (*bootstrap*) σε πραγματικά δεδομένα. Το κύριο συμπέρασμά τους είναι ότι η στατιστικά βέλτιστη μέθοδος πρόβλεψης δεν ταυτίζεται απαραίτητα με τη βέλτιστη μέθοδο ελέγχου αποθεμάτων.

3.7.2 Ολοκληρωμένα πλαίσια πρόβλεψης και αποθεμάτων (*Forecasting–inventory frameworks*)

Σε περιβάλλοντα λιανεμπορίου και έντονα διαλείπουσας ζήτησης, αρκετές εργασίες έχουν προτείνει ενοποιημένες προσεγγίσεις που ενσωματώνουν ρητά το κόστος των ανεκπλήρωτων παραγγελιών (*backorders cost*) και τη στοχοθεσία εξυπηρέτησης ήδη από τη φάση σχεδιασμού των μοντέλων πρόβλεψης. Οι Syntetos *et al.* [95] καταγράφουν το χάσμα μεταξύ ακαδημαϊκής θεωρίας και

επαγγελματικής πρακτικής στον σχεδιασμό της εφοδιαστικής αλυσίδας (*supply chain forecasting*) και τονίζουν την ανάγκη για πλαίσια που συνδέουν την πρόβλεψη και τις αποφάσεις παραγγελίας υπό μια κοινή μετρική κόστους. Αντίστοιχα, οι Babai, Boylan και Rostami-Tabar [97] παρουσιάζουν μια εκτενή επισκόπηση μεθόδων ιεραρχικής πρόβλεψης (*hierarchical forecasting*) και χρονικής συνάθροισης (*temporal aggregation*) για την αντιμετώπιση της ετερογένειας σε πολυεπίπεδα δίκτυα εφοδιασμού, στα οποία η ίδια ζήτηση πρέπει να υπολογίζεται ταυτόχρονα σε πολλαπλούς ορίζοντες για τις λειτουργίες αναπλήρωσης, κεντρικού σχεδιασμού (*master planning*) και προϋπολογισμού.

Σε επίπεδο μαθηματικής μοντελοποίησης, η προσέγγιση των Wang, Kang και Petropoulos [91] προσφέρει ισχυρή εμπειρική τεκμηρίωση για τη σημασία του υπολογισμού πλήρων κατανομών πυκνότητας ζήτησης (*demand densities*). Τα συνδυαστικά μοντέλα πυκνότητας (*ensemble density models*) που βασίζονται σε αλγορίθμους LightGBM και LSTM συνδέονται άμεσα με μετρικές αναμενόμενου κόστους ελλείψεων και ποσοστού κάλυψης επιδεικνύοντας πως οι πιθανοκρατικές προβλέψεις (βλ. ενότητα 3.5) μεταφράζονται σε βέλτιστες αποφάσεις αποθέματος. Συμπληρωματικά, οι Türkmen *et al.* [72] αποδεικνύουν ότι η άμεση εκτίμηση της κατανομής επιτρέπει τον καθορισμό επιπέδων ασφαλείας βασισμένων σε εκατοστημόρια (*quantile-based safety levels*), χωρίς τη μεσολάβηση κλασικών, στατικών τύπων συντελεστή ασφαλείας (*safety-factor heuristics*). Οι θεμελιώδεις αυτές σχέσεις μεταξύ πρόβλεψης, χρόνου παράδοσης και αποθέματος ασφαλείας, όπως περιγράφηκαν στο κλασικό εγχειρίδιο των Silver, Pyke και Peterson [11], παραμένουν αναλλοίωτες στο σύγχρονο πλαίσιο της Μηχανικής Μάθησης και αξιοποιούνται ευρέως σε υβριδικά συστήματα Επιχειρησιακής Έρευνας [141], [142].

3.7.3 Το χάσμα μεταξύ ακαδημαϊκής έρευνας και βιομηχανικής πρακτικής

Παρά τη ραγδαία μαθηματική πρόοδο, πολυάριθμες εμπειρικές μελέτες καταδεικνύουν ότι οι επιχειρήσεις εξακολουθούν να βασίζονται σε σημαντικό βαθμό σε εμπειρικές προσαρμογές (*judgmental overrides*) και απλούς ευρετικούς κανόνες (*heuristics*) κατά τη διαχείριση της διαλείπουσας ζήτησης. Η συστηματική ανασκόπηση των Van der Auweraer, Boute και Syntetos [100] καταγράφει ότι στα περισσότερα περιβάλλοντα ανταλλακτικών, τα πληροφοριακά συστήματα ενσωματώνουν μεν στατιστικές μεθόδους (Croston, SBA, TSB), αλλά οι τελικές αποφάσεις τροποποιούνται αυθαίρετα από έμπειρους σχεδιαστές, γεγονός που υπογραμμίζει ένα επίμονο χάσμα μεταξύ ακαδημαϊκών υποδειγμάτων και βιομηχανικής πραγματικότητας. Σε ανάλογη κατεύθυνση, οι Syntetos, Nikolopoulos και Boylan [131] εισάγουν μετρικές επιπτώσεων-ακρίβειας (*accuracy-implication metrics*), οι οποίες μετρούν ακριβώς το πόσο η αλγοριθμική πρόβλεψη επηρεάζει την τελική απόφαση αποθέματος προσφέροντας, έτσι, μια ρεαλιστική αξιολόγηση της αποδοχής των μοντέλων.

Η διαπίστωση αυτή καθιστά επιτακτική την ανάγκη για την ανάπτυξη μεθοδολογιών που δεν είναι μόνο στατιστικά ακριβείς, αλλά χαρακτηρίζονται από επιχειρησιακή διαφάνεια (*interpretability*) και ομαλή ενσωμάτωση σε υπάρχοντα συστήματα ERP. Οι Nathan *et al.* [98] με το πλαίσιο SHOS σε 56.000 χρονοσειρές ανταλλακτικών αυτοκινήτων απέδειξαν πώς μια προσέγγιση επικεντρωμένη στη μηχανική χαρακτηριστικών (*feature-engineering-centric*) μπορεί να τεθεί άμεσα σε παραγωγική λειτουργία χωρίς να απαιτεί εξειδικευμένες υποδομές υπολογιστικών πόρων ή χρονοβόρες διαδικασίες μίκρο-ρύθμισης. Αντίστοιχα, οι Dodin *et al.* [118] παρουσιάζουν το ολοκληρωμένο σύστημα πρόβλεψης της Bombardier ως ένα εξαιρετικό παράδειγμα μετάβασης από τα ακαδημαϊκά πρότυπα (*benchmarks*) σε ένα πλήρως παραγωγικό σύστημα αεροπορικών ανταλλακτικών.

3.8 Σύνοψη, ερευνητικά κενά και τοποθέτηση της διπλωματικής εργασίας

Η βιβλιογραφία για την πρόβλεψη της διαλείπουσας ζήτησης σε περιβάλλοντα ανταλλακτικών έχει ωριμάσει θεαματικά κατά τις τελευταίες δεκαετίες. Όπως αναλύθηκε στις προηγούμενες ενότητες, η επιστημονική έρευνα έχει μεταβεί από τις παραδοσιακές μεθόδους τύπου Croston και τις παραλλαγές τους (§3.2), στα σύγχρονα πλαίσια Μηχανικής και Βαθιάς Μάθησης (§3.3), και από εκεί σε εξειδικευμένα μοντέλα Εμποδίου (Hurdle) με επίκεντρο τη μηχανική χαρακτηριστικών (§3.4). Πλέον, η αιχμή της έρευνας επικεντρώνεται σε ενοποιημένα πιθανοκρατικά πλαίσια αξιολόγησης (§3.5), σε καινοτόμες αρχιτεκτονικές Μεγάλων Γλωσσικών Μοντέλων (§3.6), καθώς, και στην άμεση ενσωμάτωση των προβλέψεων στις πολιτικές ελέγχου αποθεμάτων (§3.7).

Οι συστηματικές ανασκοπήσεις των Pinçe *et al.* [96] και Giannopoulos *et al.* [45] συγκλίνουν στο συμπέρασμα ότι το πρόβλημα της διαλείπουσας ζήτησης παραμένει δομικά δυσχερές λόγω της ακραίας σποραδικότητας, του κινδύνου απαξίωσης, του αχανούς πλήθους των διαθέσιμων κωδικών και των αυστηρών περιορισμών αποθέματος. Παράλληλα, οι πιο σύγχρονες κατευθύνσεις, όπως τα πλαίσια βαθιάς ανανέωσης, τα συνδυαστικά μοντέλα πυκνότητας και οι αρχιτεκτονικές Μίγματος Ειδικών (MoE), καταδεικνύουν ρητά ότι η «σωστή» μετρική αξιολόγησης ενός μοντέλου εξαρτάται ευθέως από το κόστος ελλείψεων, το κόστος διατήρησης αποθέματος (*holding cost*) και το επιθυμητό επίπεδο εξυπηρέτησης (*service level*).

3.8.1 Ερευνητικά κενά στη βιβλιογραφία

Παρά τη σημαντική επιστημονική πρόοδο, η εστιασμένη ανάλυση που προηγήθηκε αναδεικνύει πέντε σαφή ερευνητικά κενά, ιδιαίτερα εμφανή στο εξειδικευμένο πλαίσιο της δευτερογενούς αγοράς ανταλλακτικών αυτοκινήτων (automotive aftermarket):

- **Περιορισμένη βιβλιογραφία με εξειδίκευση στον κλάδο του αυτοκινήτου:** Η συντριπτική πλειονότητα των εφαρμοσμένων μελετών για σποραδικά ανταλλακτικά επικεντρώνεται στην αεροπορική συντήρηση [107], [118], [130], σε ηλεκτρονικά εξαρτήματα [120] ή σε γενικά βιομηχανικά ανταλλακτικά [105], [117]. Οι στοχευμένες μελέτες στην αγορά του αυτοκινήτου είναι συγκριτικά ελάχιστες και συχνά περιορίζονται σε κλασικά στατιστικά ή βασικά ML μοντέλα. Πρόσφατες εξαιρέσεις [12], [13], [98] απλώς επιβεβαιώνουν τη σπανιότητα ολοκληρωμένων, μεγάλης κλίμακας ερευνών στο συγκεκριμένο πεδίο.
- **Αποσπασματική ενσωμάτωση της «αλήθειας» των δεδομένων (*Demand censoring*):** Ελάχιστες μελέτες αντιμετωπίζουν συστηματικά το ζήτημα της λογοκρισίας της ζήτησης, όπου τα μηδενικά που οφείλονται σε έλλειψη αποθέματος συγχέονται με τα μηδενικά που προκύπτουν από τη φυσική, διαλείπουσα φύση του προϊόντος. Παρότι οι [6], [7], [8], [125] αναδεικνύουν την κρισιμότητα του φαινομένου, η ενσωμάτωση δεικτών έλλειψης (*out-of-stock flags*) ως ενεργών χαρακτηριστικών ή βαρών δειγματοληψίας (*sample weights*) παραμένει σπάνια στα σύγχρονα πλαίσια υλοποίησης.
- **Έλλειψη ενιαίας αξιολόγησης με μικτά κριτήρια:** Ακόμη και, όταν επιστρατεύονται προηγμένοι αλγόριθμοι, η αξιολόγηση παραμένει κατακερματισμένη. Η αποκλειστική χρήση μετρικών επικεντρωμένων στην πρόβλεψη (όπως MAPE ή RMSE), δίχως τον συνδυασμό τους με κλιμακωτούς δείκτες (MASE, RMSSE) και επιχειρησιακά KPIs (π.χ. κόστος, ποσοστό κάλυψης), δυσχεραίνει την επιχειρησιακή ερμηνεία των αποτελεσμάτων [3], [14], [15]. Επιπλέον, συχνά απουσιάζει η στρωματοποιημένη (*stratified*) αξιολόγηση βάσει κατηγορίας προϊόντων, όσον αφορά τον τζίρο τους (Top-N κωδικοί ειδών), παρότι η πρακτική σημασία των SKUs υψηλής αξίας είναι δυσανάλογα μεγαλύτερη.

- **Απουσία μεγάλης κλίμακας πλαισίων με συσσωρευτική μετα-μάθηση (*Meta-learning stacking*) και βαθμονόμηση:** Οι διαγωνισμοί M [21], [104] επιβεβαίωσαν ότι τα δέντρα διαβαθμισμένης ενίσχυσης (GBT) με προσεγμένη μηχανική χαρακτηριστικών διαπρέπουν σε σποραδικά περιβάλλοντα. Ωστόσο, σπανίζουν οι εργασίες που συνδυάζουν άμεση πρόβλεψη πολλαπλών οριζόντων (*multi-horizon direct forecasting*) με τεχνικές μετα-μάθησης βασισμένες σε αυστηρές εκτός δείγματος δοκιμές (*backtest-driven predictions stacking*) και βαθμονόμηση βάσει εσόδων (*revenue-stratified calibration*).
- **Το κενό στη σύγκριση πλαισίων βαθιάς μάθησης (LLMs / RC) με ισχυρά βιομηχανικά πρότυπα:** Ενώ οι εφαρμογές που βασίζονται στα Μεγάλα Γλωσσικά Μοντέλα (π.χ. AutoTimes [133], GPT2ForForecasting [138]) και στα Δίκτυα Κατάστασης Ηχούς (ESN) (π.χ. [139]) ανοίγουν νέους δρόμους, η εφαρμογή τους σε δεδομένα αυτοκινήτου με ακραία εμφάνιση μηδενικών παρατηρήσεων παραμένει πειραματική. Όπως επισημαίνουν οι Tan *et al.* [112], απουσιάζουν οι συγκριτικές μελέτες που να τοποθετούν αυτές τις νέες προσεγγίσεις δίπλα σε άρτια ρυθμισμένα, παραδοσιακά μοντέλα αναφοράς μηχανικής μάθησης εντός ενός ενιαίου ρεαλιστικού πλαισίου σύγκρισης.

3.8.2 Τοποθέτηση της παρούσας διπλωματικής εργασίας

Η παρούσα διπλωματική εργασία έρχεται να καλύψει τα προαναφερθέντα ερευνητικά κενά μέσω τριών διακριτών μεθοδολογικών αξόνων, οι οποίοι αναπτύσσονται στο Κεφάλαιο 5:

- **Πρώτον,** εστιάζει αποκλειστικά σε ένα μεγάλης κλίμακας πραγματικό σύνολο δεδομένων της αγοράς ανταλλακτικών αυτοκινήτου (περίπου 72.000 SKUs, εκ των οποίων το 96,5% χαρακτηρίζεται ως διαλείπον/ακανόνιστο κατά τον ορισμό των Syntetos και Boylan [1], βλ. κεφάλαιο 4). Στο πλαίσιο αυτό, η προτεινόμενη ροή εργασίας (*pipeline*) ενσωματώνει μια πολυδιάστατη και πλούσια αρχιτεκτονική μηχανικής χαρακτηριστικών (*rich feature engineering*), η οποία ξεπερνά τις απλές ιστορικές υστερήσεις κωδικοποιώντας προηγμένα στατιστικά, χρονικά και συμπεριφορικά μοτίβα της ζήτησης. Παράλληλα, το σύστημα μοντελοποιεί ρητά τις αμοιβαίες συσχετίσεις και αλληλεξαρτήσεις μεταξύ των διαφορετικών κωδικών (*cross-SKU correlations*) και, με αυτό τον τρόπο, επιτρέπει τη μεταφορά γνώσης εντός του προϊοντικού δικτύου για την αποτελεσματικότερη διαχείριση της ακραίας σποραδικότητας. Οι πληροφορίες αυτές τροφοδοτούν αρχιτεκτονικές Εμποδίου δύο σταδίων (*two-stage* Hurdle) και αλγορίθμους διαβαθμισμένης ενίσχυσης (HistGradientBoosting και CatBoost) και δέντρων απόφασης (Random Forests και Extra Trees). Η καινοτομία έγκειται στη χρήση συσσωρευτικής μετα-μάθησης πάνω σε εκτός δείγματος προβλέψεις (*out-of-sample backtest predictions*) και στη στρωματοποιημένη βαθμονόμηση βάσει εσόδων (*revenue-stratified calibration*), διαχειριζόμενη άριστα την ετερογένεια του καταλόγου.
- **Δεύτερον,** αντιμετωπίζει κατά μέτωπο το πρόβλημα της λογοκριμένης ζήτησης με τη ρητή ενσωμάτωση δεικτών έλλειψης αποθέματος (*out-of-stock flags*) κατά τη φάση της μηχανικής χαρακτηριστικών, καθώς, και τεχνικών στάθμισης (weighting) κατά την εκπαίδευση. Η εργασία αναδεικνύει την απόκλιση μεταξύ *καταγεγραμμένων πωλήσεων* και *πραγματικής ζήτησης* ως μεθοδολογικό ζήτημα υψίστης σημασίας.
- **Τρίτον,** υιοθετεί ένα ολιστικό πλαίσιο αξιολόγησης. Χρησιμοποιεί κλιμακωτές μετρικές στα πρότυπα του διαγωνισμού M5 [21] (WAPE, MASE) σε άμεσο συνδυασμό με επιχειρησιακά κριτήρια προσομοίωσης αποθεμάτων (προτεινόμενες ποσότητες αναπλήρωσης, επίπεδα ασφαλείας). Η αξιολόγηση πραγματοποιείται σε επίπεδα βάσει τζίρου (Top-20, κλάσεις A/B/C) διασφαλίζοντας ότι η σύγκριση των μοντέλων δεν περιορίζεται στο θεωρητικό στατιστικό σφάλμα, αλλά επεκτείνεται στον μετρήσιμο επιχειρησιακό τους αντίκτυπο.

Μέσω αυτών των τριών αξόνων, η παρούσα μελέτη φιλοδοξεί να γεφυρώσει τις τρεις ιστορικές διαιρέσεις της τρέχουσας βιβλιογραφίας: (α) τη διαίρεση μεταξύ στατιστικής και επιχειρησιακής αξιολόγησης, (β) τη διαίρεση μεταξύ θεωρητικών μεθοδολογιών και ροών εργασίας έτοιμων για παραγωγική λειτουργία (*production-ready pipelines*), και (γ) τη διαίρεση μεταξύ μεθόδων που αγνοούν τη διαλείπουσα φύση της ζήτησης και αυτών που τη μοντελοποιούν ρητά.

Η αναλυτική μεθοδολογική θεμελίωση αυτών των στρατηγικών επιλογών ακολουθεί στο Κεφάλαιο 5.

3.8.3 Παράλληλη επιστημονική σύγκλιση

Αξίζει να σημειωθεί ένα εξαιρετικά ενδιαφέρον εύρημα που αναδεικνύει τον συγχρονισμό της παρούσας μελέτης με την αιχμή της διεθνούς έρευνας. Παράλληλα με την ολοκλήρωση της μεθοδολογικής ανάπτυξης του προτεινόμενου συστήματος, δημοσιεύτηκαν στη βιβλιογραφία δύο ανεξάρτητες μελέτες, οι οποίες συγκλίνουν εντυπωσιακά σε αντίστοιχες αρχιτεκτονικές επιλογές.

Αφενός, οι Nathan *et al.* [98] τεκμηριώνουν εμπειρικά ότι η **επιμελής μηχανική χαρακτηριστικών** (*feature engineering*) υπερτερεί της αρχιτεκτονικής πολυπλοκότητας στην πρόβλεψη της διαλείπουσας ζήτησης. Επιβεβαιώνουν, δηλαδή, ότι τα ισχυρά μοντέλα διαβαθμισμένης ενίσχυσης (*gradient boosting*) με τον κατάλληλο σχεδιασμό χαρακτηριστικών, ξεπερνούν τις πιο σύνθετες βαθιές αρχιτεκτονικές. Αυτό είναι ένα εύρημα που ταυτίζεται απόλυτα με τη στρατηγική επιλογή του Κεφαλαίου 5 της παρούσας μελέτης.

Αφετέρου, οι Mușat & Căbuz [111] προτείνουν την αρχιτεκτονική Switch-Hurdle, η οποία ενσωματώνει έναν κωδικοποιητή Μίγματος Ειδικών (*Mixture of Experts* - MoE) συνδυασμένο με μια αναδρομική κεφαλή εμποδίου (*hurdle head*). Η δομή αυτή είναι εννοιολογικά απολύτως συγγενής με τη **δι-σταδιακή αρχιτεκτονική Εμποδίου** (*two-stage* Hurdle*: classifier + regressor*) που υιοθετείται αυτόνομα στο δικό μας σύστημα (βλ. ενότητα 5.6.5).

Και οι δύο αυτές εργασίες εντοπίστηκαν εκ των υστέρων και δεν καθοδήγησαν τον αρχικό σχεδιασμό της προτεινόμενης ροής εργασιών. Παρατίθενται, ωστόσο, στο παρόν κεφάλαιο ως ισχυρά **τεκμήρια παράλληλης επιστημονικής σύγκλισης**, τα οποία επικυρώνουν και ενισχύουν με τον πλέον κατηγορηματικό τρόπο την ορθότητα και την εγκυρότητα των μεθοδολογικών επιλογών που ακολουθήθηκαν.

Κεφάλαιο 4ο: Περιγραφή Συνόλου Δεδομένων και Στατιστική Ανάλυση

Το τρέχον κεφάλαιο παρουσιάζει αναλυτικά το σύνολο δεδομένων που χρησιμοποιήθηκε. Τα δεδομένα προέρχονται από πραγματικό επιχειρησιακό περιβάλλον μιας ελληνικής εταιρείας χονδρικής εμπορίας ανταλλακτικών αυτοκινήτων, και καλύπτουν πλήρη ιστορικότητα συναλλαγών πενταετίας (2021–2025). Η στατιστική ανάλυση των δεδομένων εστιάζει σε τρεις άξονες: (α) τη δομή και τον όγκο των δεδομένων, (β) τα οικονομικά χαρακτηριστικά (τζίρος, αποθέματα), και (γ) τα στατιστικά χαρακτηριστικά της ζήτησης (αραιότητα, εποχικότητα, μετατόπιση κατανομής δεδομένων μεταξύ συνόλων δεδομένων εκπαίδευσης και αξιολόγησης).

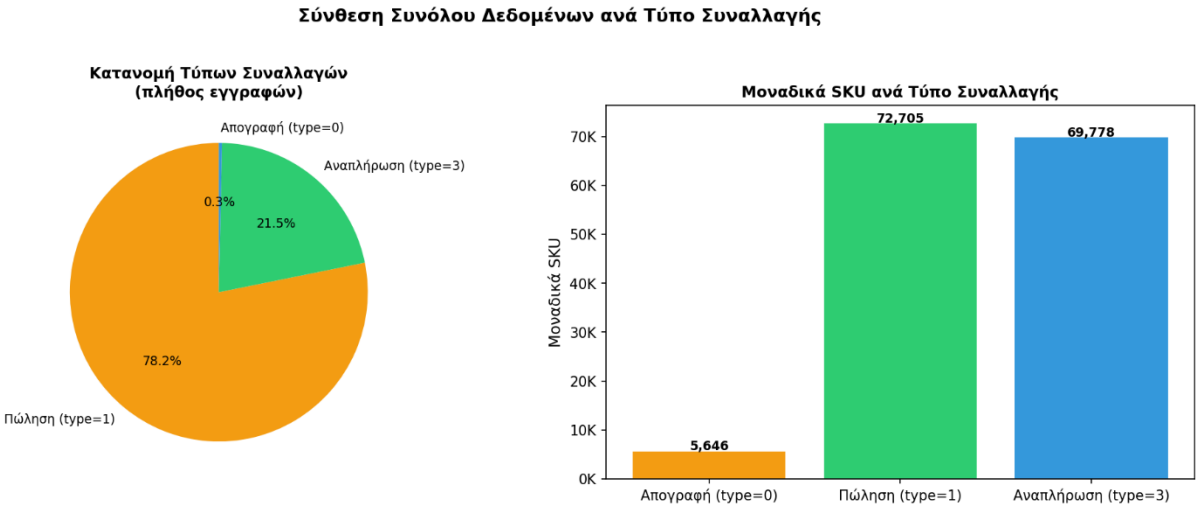
Το σύνολο δεδομένων βρίσκεται σε αρχείο τύπου `csv` και περιλαμβάνει συνολικά **1.617.633** εγγραφές συναλλαγών που αφορούν **79.363** μοναδικούς κωδικούς ειδών (*Stock Keeping Units* – SKU) κατά τη χρονική περίοδο 1/1/2021 έως 31/12/2025. Η κάθε γραμμή του αρχείου αποτελεί και μία συναλλαγή.

4.1 Γενικά χαρακτηριστικά συνόλου δεδομένων

Κάθε εγγραφή χαρακτηρίζεται από ένα πεδίο τύπου συναλλαγής (`Transaction_Type_FKEY`) με τρεις τιμές όπως φαίνεται και στον Πίνακα 4.1 και το Σχήμα 4.1:

Πίνακας 4.1: Σύνθεση συνόλου δεδομένων ανά τύπο συναλλαγής

Τύπος	Περιγραφή	Εγγραφές	% Συνόλου	Μοναδικά SKU
0	Απογραφή αποθήκης (inventory)	~5.000	0,3%	5.646
1	Πώληση (sales)	~1.262.000	78,2%	72.705
3	Αναπλήρωση αποθέματος (replenishment)	~348.000	21,5%	69.778



Σχήμα 4.1: Σύνθεση συνόλου δεδομένων ανά τύπο συναλλαγής

Οι συναλλαγές πώλησης (`type=1`) αποτελούν τον κορμό του συνόλου δεδομένων και χρησιμοποιούνται για τη μοντελοποίηση και πρόβλεψη ζήτησης. Οι αναπληρώσεις (`type=3`) και οι απογραφές (`type=0`) χρησιμοποιούνται συμπληρωματικά για την ανακατασκευή του αποθέματος (*stock reconstruction*). Αξίζει να σημειωθεί ότι υπάρχουν **6.883** SKU που εμφανίζονται αποκλειστικά σε εγγραφές

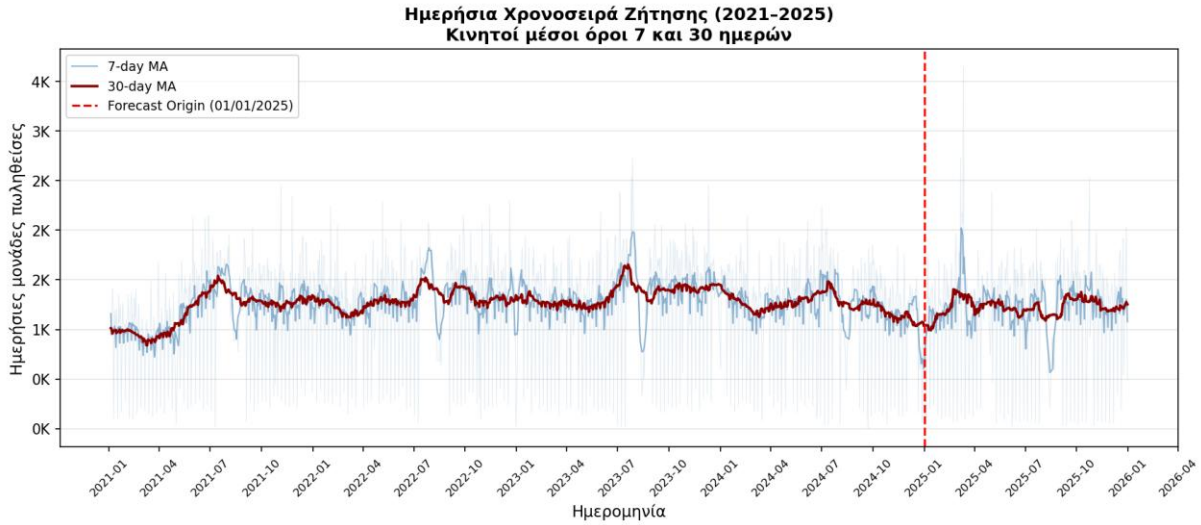
αποθεματικών ή αναπληρώσεων αλλά δεν έχουν καμία πώληση. Αυτά αποτελούν το λεγόμενο «νεκρό απόθεμα» (*dead stock*), δηλαδή αποθέματα που αγοράστηκαν αλλά δεν πωλήθηκαν ποτέ.

Μετά τον καθαρισμό (αφαίρεση εγγραφών με άκυρο ή κενό κωδικό προϊόντος, μηδενική ή αρνητική ποσότητα ή με ημερομηνίες εκτός εύρους), τα δεδομένα πωλήσεων (`type=1`) περιλαμβάνουν:

- 1.162.551 εγγραφές πωλήσεων
- 72.310 μοναδικά SKU με τουλάχιστον μία πώληση
- 418 κατηγορίες προϊόντων

4.2 Ιστορικά στοιχεία ζήτησης

Στο Σχήμα 4.2 παρουσιάζεται η κατανομή της συνολικής ημερήσιας ζήτησης σε μονάδες πλαισιωμένη από κυλιόμενους μέσους όρους 7 και 30 ημερών. Η κόκκινη διακεκομμένη γραμμή αναδικνύει το «*forecast origin*» (01/01/2025), δηλαδή το χρονικό σημείο στο οποίο χωρίζεται η ιστορική (εκπαιδευτική) περίοδος από την περίοδο αξιολόγησης.



Σχήμα 4.2: Ημερήσια κατανομή ζήτησης

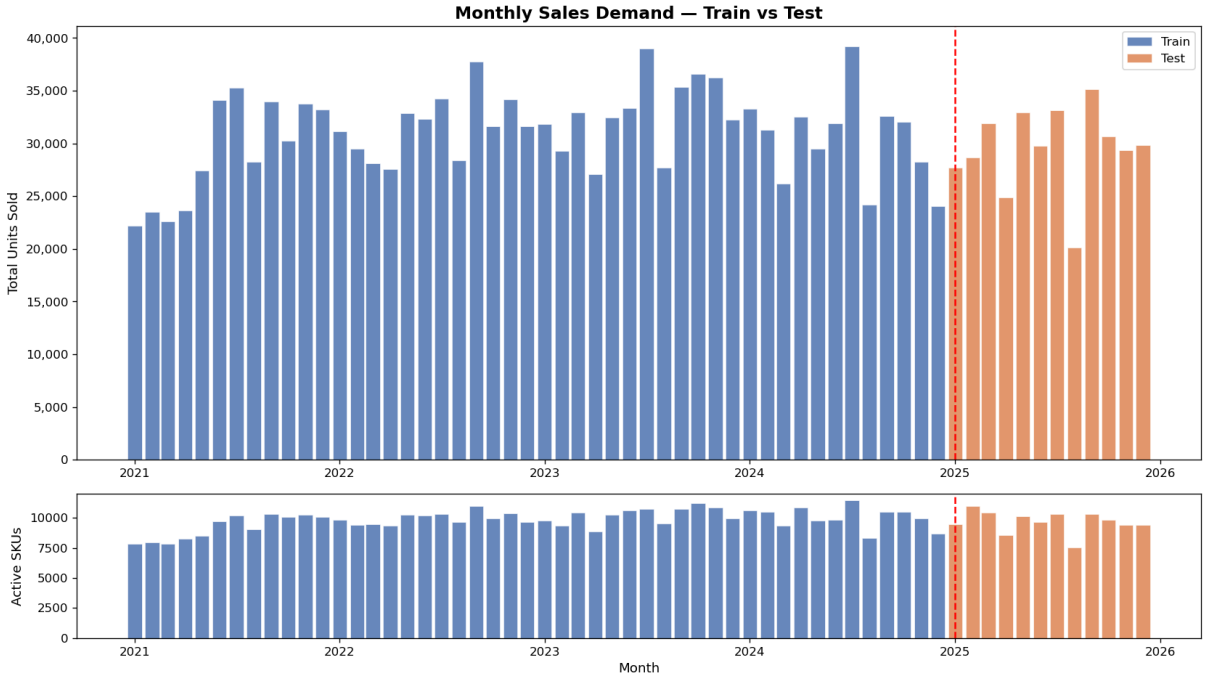
Παρατηρείται ότι η ημερήσια ζήτηση κυμαίνεται γύρω στις 1000 – 2000 μονάδες, με σαφή εβδομαδιαία κυκλικότητα (χαμηλότερη ζήτηση Σαββατοκύριακα) και εποχικό πρότυπο (αυξημένη ζήτηση κατά τους καλοκαιρινούς μήνες και τον Σεπτέμβριο).

Στη συνέχεια, στον Πίνακα 4.2 και το Σχήμα 4.3 παρουσιάζονται οι μέσες μηνιαίες πωλήσεις στα σύνολα δεδομένων εκπαίδευσης και αξιολόγησης.

Πίνακας 4.2: Συγκριτικά στοιχεία μηνιαίων πωλήσεων

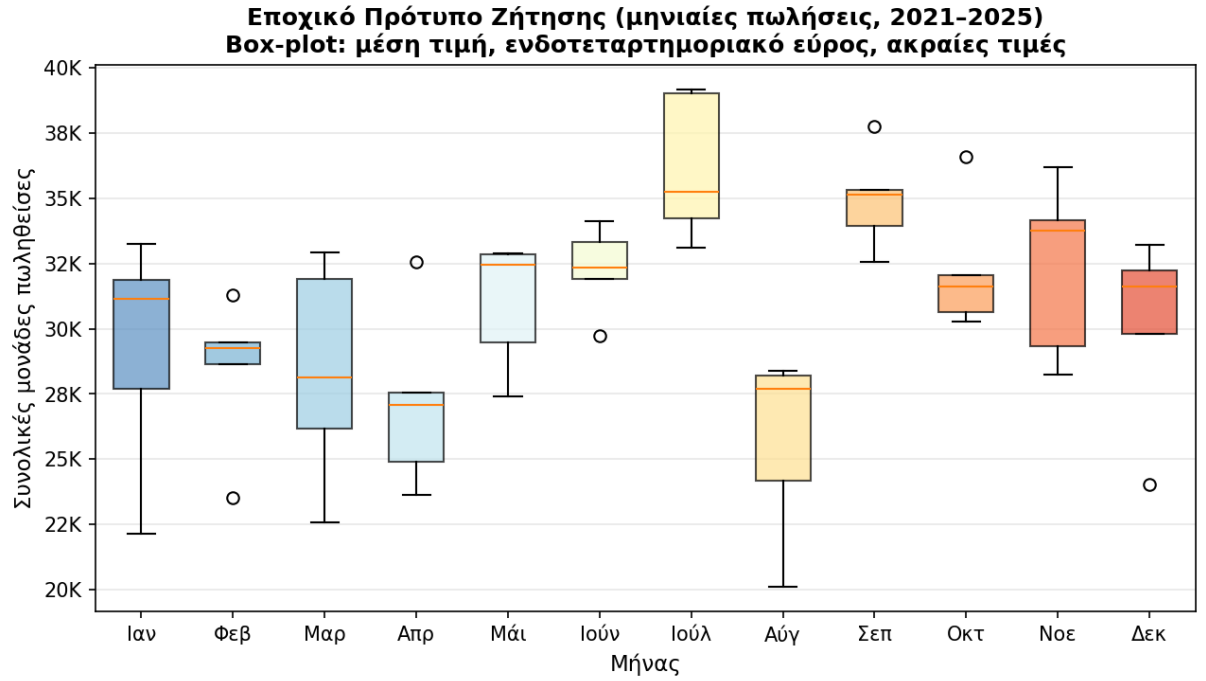
	Μέση Μηνιαία Ζήτηση (μονάδες	**Ενεργά SKU/μήνα**
Σύνολο εκπαίδευσης (2021-2024)	~30.965	~9.900
Σύνολο αξιολόγησης (2025)	~29.500	~9.661
Διαφορά	−4,7%	−2,4%

Η μηνιαία ζήτηση κατά το 2025 είναι ελαφρώς χαμηλότερη (−4,7%) σε σχέση με τον ιστορικό μέσο, γεγονός που υποδεικνύει μέτρια πτωτική τάση στη ζήτηση.



Σχήμα 4.3: Κατανομή μηνιαίων πωλήσεων

Επίσης, η εποχικότητα αποτελεί σημαντική παράμετρο στην πρόβλεψη ζήτησης ανταλλακτικών. Ο κλάδος παρουσιάζει χαρακτηριστικό εποχικό πρότυπο με αυξημένη ζήτηση τους πρώτους καλοκαιρινούς μήνες (Ιούνιος, Ιούλιος) και τον Σεπτέμβριο, ενώ ο Αύγουστος εμφανίζει σημαντική ύφεση λόγω θερινών διακοπών.



Σχήμα 4.4: Εποχικό πρότυπο ζήτησης μηνιαίων πωλήσεων 2021 – 2025

Αυτό το εποχικό πρότυπο αντανακλά τον κύκλο συντήρησης οχημάτων στην Ελλάδα. Δηλαδή, αυξημένη ετοιμασία οχημάτων πριν τις καλοκαιρινές μετακινήσεις (Ιούνιος–Ιούλιος) και μετά τις διακοπές (Σεπτέμβριος). Τα συμπεράσματα αυτά επιβεβαιώνονται και από το γράφημα (*box-plot*) του Σχήματος 4.4, στο οποίο απεικονίζονται όχι μόνο οι διάμεσες (*median*) τιμές ανά μήνα αλλά και η διασπορά που παρατηρείται.

4.3 Ανάλυση τζίρου

Ο συνολικός τζίρος της πενταετίας 2021-2025 για το υπό εξέταση δίκτυο ανέρχεται σε €22.586.443. Προκειμένου να αναλυθεί η δομή των πωλήσεων και να προσδιοριστεί η επιχειρησιακή κρισιμότητα του εκάστοτε κωδικού, κρίνεται απαραίτητη η ταξινόμηση του προϊοντικού καταλόγου. Στο πλαίσιο αυτό, εφαρμόστηκαν καθιερωμένες μέθοδοι της Επιχειρησιακής Έρευνας, όπως η Ανάλυση Pareto, η κατηγοριοποίηση ABC και η αξιολόγηση της συγκέντρωσης του τζίρου.

4.3.1 Ορισμοί κατηγοριών τζίρου κατά Pareto

Η Αρχή του Pareto (γνωστή και ως «Κανόνας 80/20») διατυπώθηκε αρχικά από τον Ιταλό οικονομολόγο Vilfredo Pareto, ενώ ο Juran **Error! Reference source not found.** τη γενίκευσε ως καθολική αρχή διαχείρισης ποιότητας και αποθεμάτων. Παρατηρήθηκε ότι σε πολλά συστήματα, ένα μικρό ποσοστό αιτιών (20%) ευθύνεται για το μεγαλύτερο μέρος των αποτελεσμάτων (80%). Στη διαχείριση αποθεμάτων, η αρχή αυτή υποδεικνύει ότι λίγα προϊόντα παράγουν τον κύριο όγκο του τζίρου, ενώ η συντριπτική πλειονότητα συνεισφέρει ελάχιστα [11].

Η προσέγγιση ABC (*ABC analysis*) αποτελεί την πρακτική εφαρμογή αυτής της αρχής. Τα προϊόντα ταξινομούνται φθίνοντα βάσει τζίρου και κατανέμονται παραδοσιακά σε τρεις κατηγορίες [22], [144]:

- **Κατηγορία A (Pareto-A):** Είδη υψηλής αξίας που αθροιστικά παράγουν το ~80% του τζίρου (τυπικά το 15-20% των κωδικών). Σε αυτή την κατηγορία εντάσσεται και το στενότερο υποσύνολο των **Top-20 SKU**, δηλαδή των 20 απολύτως κορυφαίων προϊόντων σε πωλήσεις.
- **Κατηγορία B:** Είδη μεσαίας αξίας που παράγουν το επόμενο ~15% του τζίρου.
- **Κατηγορία C:** Είδη χαμηλής αξίας που συνεισφέρουν το τελευταίο ~5% του τζίρου.

Για τους σκοπούς της παρούσας μελέτης, και ειδικά για την αξιολόγηση των μοντέλων πρόβλεψης (Κεφάλαιο 5), υιοθετείται μια διμερής κατηγοριοποίηση. Διαχωρίζονται τα «***Pareto-A***» SKU από τα υπόλοιπα (r*est B/C*), τα οποία συλλογικά χαρακτηρίζονται ως «***Long-tail***» (Μακρά Ουρά).

Ο όρος *«Long-tail»* εισήχθη από τον Anderson [145] περιγράφει ακριβώς τον αχανή αριθμό προϊόντων που, αν και σπάνια πωλούμενα, έχουν αθροιστικά σημαντικό οικονομικό αποτύπωμα. Στην εφοδιαστική αλυσίδα ανταλλακτικών, τα *long-tail* SKU χαρακτηρίζονται από αραιή, διαλείπουσα ζήτηση, δυσκολία πρόβλεψης λόγω έλλειψης δεδομένων («ψυχρή εκκίνηση» / *cold-start*) και εξαιρετικά υψηλό κόστος διατήρησης αποθέματος σε σχέση με τον παραγόμενο τζίρο.

4.3.2 Ετήσια εξέλιξη τζίρου και ανισοκατανομή

Βάσει των παραπάνω ορισμών, στον Πίνακα 4.3 παρατίθεται η ετήσια κατανομή του τζίρου για την πενταετία 2021-2025, καθώς και τα μερίδια συμμετοχής των βασικών κατηγοριών.

Όπως προκύπτει από τα δεδομένα , μετά από δύο έτη ισχυρής ανάπτυξης (2021-2023, +28,6% σωρευτικά), ο τζίρος σταθεροποιήθηκε εμφανίζοντας ελαφρά πτωτική τάση κατά την περίοδο 2024-2025. Παράλληλα, παρατηρείται μια ξεκάθαρη δομική μετατόπιση, κατά την οποία το μερίδιο των *Pareto-A* SKU μειώνεται συστηματικά (από 82,6% το 2021 σε 77,0% το 2025) , ενώ ταυτόχρονα το

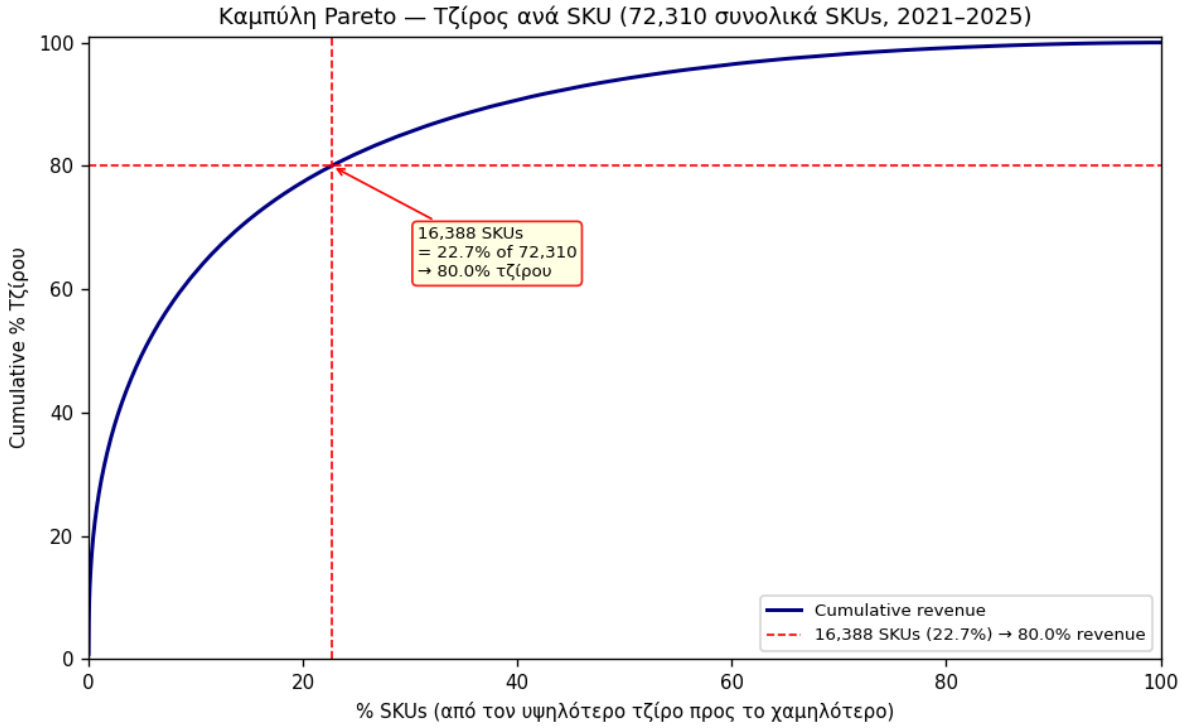
μερίδιο της κλειστής ομάδας των Top-20 SKU αυξάνεται (από 5,3% σε 7,2%). Η αντίθεση αυτή υποδεικνύει ότι τα απολύτως κορυφαία είδη ενισχύουν τη θέση τους, ενώ η μεσαία ζώνη του καταλόγου χάνει έδαφος υπέρ των ειδών της μακράς ουράς (*long-tail*), τα οποία σταδιακά αυξάνουν το μερίδιό τους από 17,4% σε 23,0%.

Πίνακας 4.3: Ετήσιος τζίρος και μερίδια ανά κατηγορία τζίρου προϊόντων

Έτος	Σύνολο (€)	Μεταβολή	Top-20 (€)	Top-20 %	Pareto-A (80%) %	Long-tail / Rest %
2021	3.844.656	—	204.108	5,3%	82,6%	17,4%
2022	4.425.722	+15,1%	234.452	5,3%	81,8%	18,2%
2023	4.946.644	+11,8%	353.154	7,1%	80,7%	19,3%
2024	4.737.160	−4,2%	323.471	6,8%	78,5%	21,5%
2025	4.632.262	−2,2%	335.690	7,2%	77,0%	23,0%

4.3.3 Οπτικοποίηση ευρημάτων: Καμπύλη Pareto και δείκτης Gini

Η κλασική ανάλυση κατά Pareto στο διαθέσιμο σύνολο δεδομένων επιβεβαιώνει τον θεωρητικό κανόνα. Συγκεκριμένα, 16.388 SKU (ήτοι το 22,7% του συνόλου) παράγουν το 80% του συνολικού τζίρου της πενταετίας (Σχήμα 4.5).

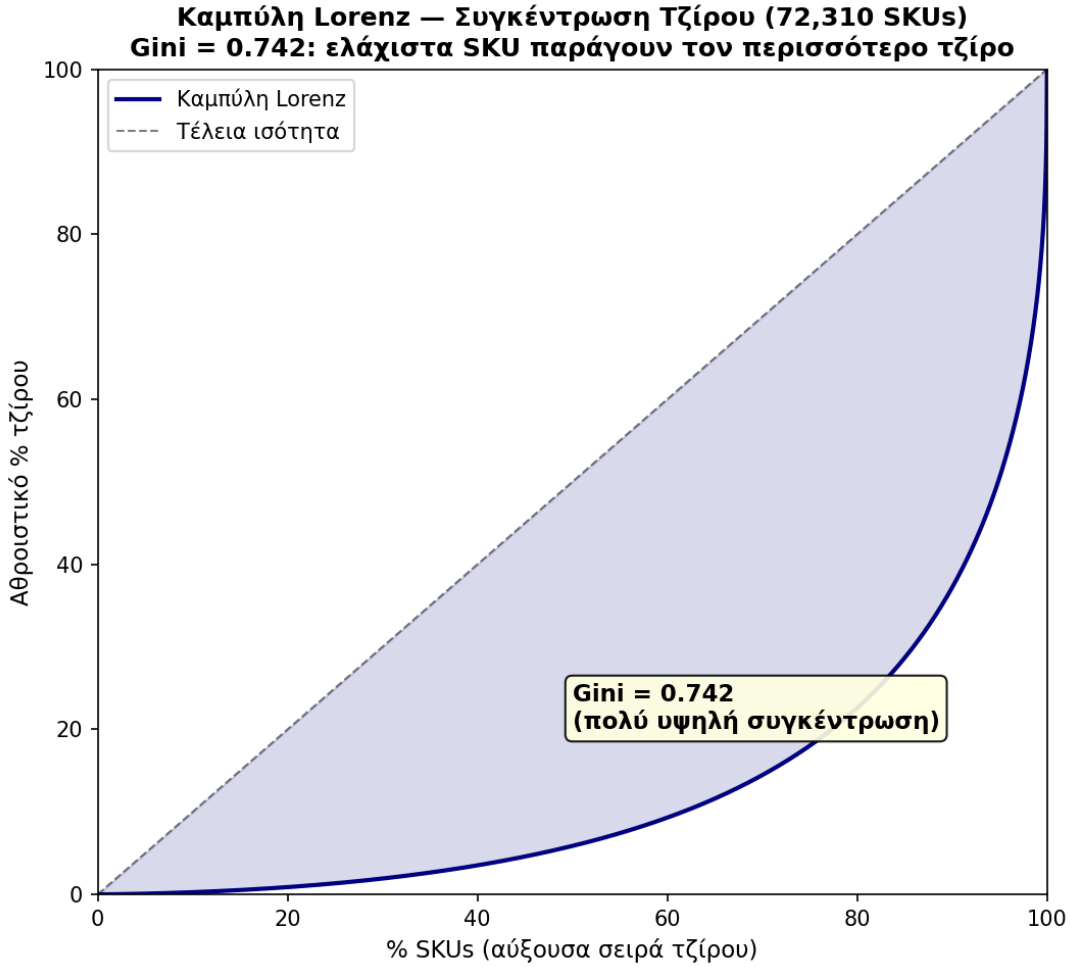


Σχήμα 4.5: Καμπύλη Pareto για το διαθέσιμο σύνολο δεδομένων

Για την αυστηρή μαθηματική ποσοτικοποίηση αυτής της ανισοκατανομής, αξιοποιήθηκαν η καμπύλη Lorenz και ο δείκτης Gini (Σχήμα 4.6). Η καμπύλη Lorenz αποτελεί τη γραφική αναπαράσταση της συγκέντρωσης του τζίρου, όπου η απόκλισή της από τη διαγώνιο («γραμμή τέλειας ισότητας») υποδεικνύει το μέγεθος της ανισοκατανομής. Όσο πιο κοίλη είναι η καμπύλη, τόσο εντονότερη είναι η ανισοκατανομή του τζίρου. Ο δείκτης Gini συμπυκνώνει αυτή τη σχέση στο διάστημα $[0,1]$ μέσω του ολοκληρώματος:

$$G = 1 - 2\int_0^1 L\,(x)dx \tag{4.1}$$

όπου $L(x)$ είναι η συνάρτηση της καμπύλης Lorenz. Γεωμετρικά, ο δείκτης Gini ισούται με τον λόγο της επιφάνειας μεταξύ διαγωνίου και καμπύλης Lorenz προς την επιφάνεια κάτω από τη διαγώνιο. Τιμή $G = 0$ υποδηλώνει τέλεια ισότητα (όλα τα SKU έχουν τον ίδιο τζίρο), ενώ $G = 1$ υποδηλώνει απόλυτη ανισοκατανομή (ένα SKU παράγει όλο τον τζίρο).



Σχήμα 4.6: Καμπύλη Lorenz – Δείκτης Gini

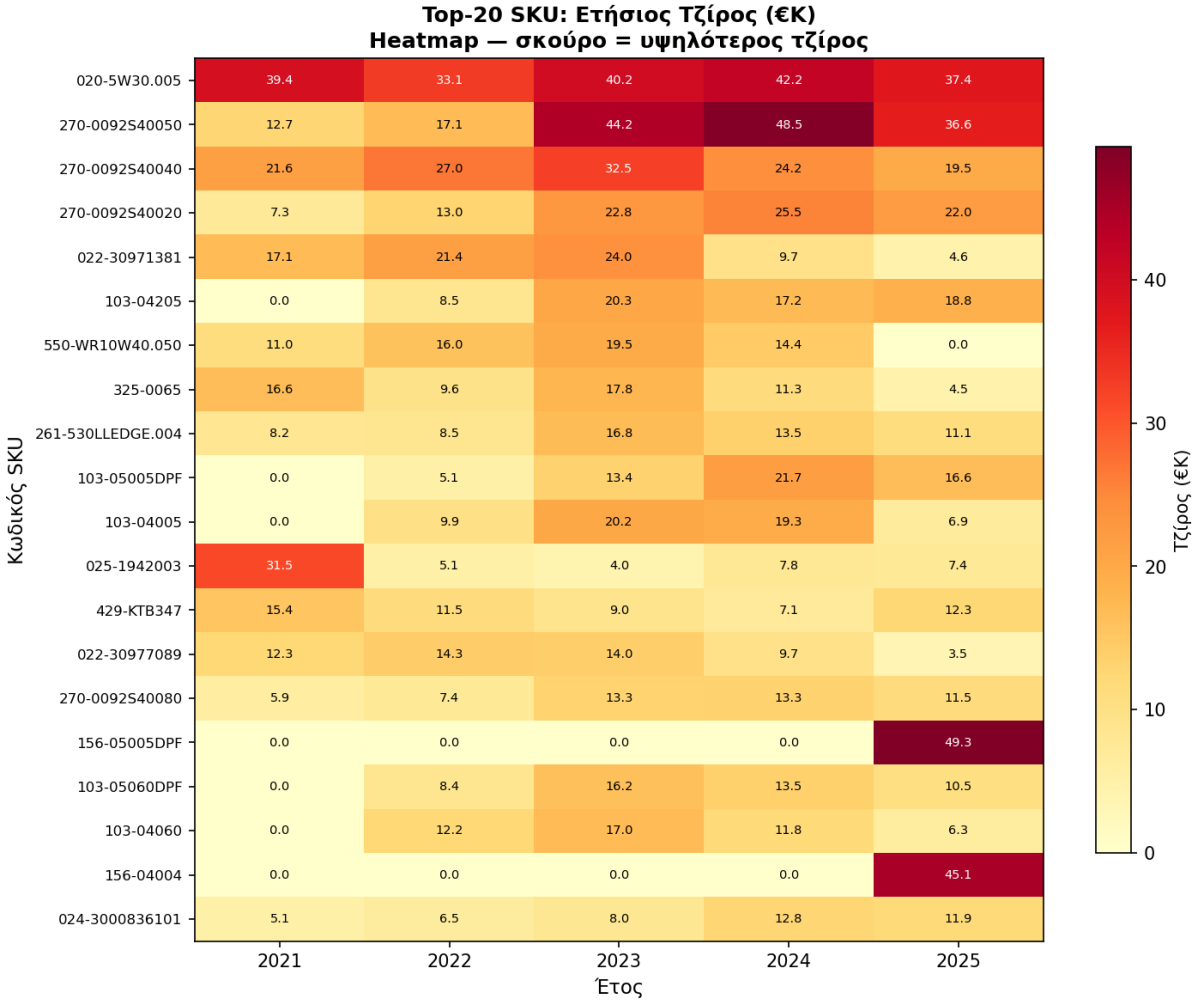
Ο δείκτης Gini αναπτύχθηκε αρχικά για τη μέτρηση ανισοκατανομής εισοδήματος στα οικονομικά, αλλά έχει βρει ευρεία εφαρμογή στη διαχείριση της εφοδιαστικής αλυσίδας ως συμπληρωματικό εργαλείο της ABC κατηγοριοποίησης [11], [144]. Η υπολογισθείσα τιμή $G = 0{,}742$ για το εξεταζόμενο σύνολο δεδομένων καταδεικνύει μια ακραία συγκέντρωση πωλήσεων. Το κατώτερο 75% του καταλόγου παράγει λιγότερο από το 10% του τζίρου. Οι τυπικές τιμές Gini σε βιομηχανικούς κλάδους B2B κυμαίνονται μεταξύ 0,6 και 0,8, γεγονός που επιβεβαιώνει ότι η παρούσα επιχείρηση αντιμετωπίζει την κλασική πρόκληση του κλάδου των ανταλλακτικών, δηλαδή μεγάλη πολυπλοκότητα καταλόγου προϊόντων (>72.000 SKU) συντηρούμενη από έναν ελάχιστο πυρήνα κωδικών υψηλής απόδοσης.

4.3.4 Μικρο-ανάλυση των κορυφαίων 20 προϊόντων (Top-20 SKU)

Ιδιαίτερο επιχειρησιακό ενδιαφέρον παρουσιάζει η στενή ομάδα των 20 απολύτως κορυφαίων κωδικών (Top-20 SKU) βάσει τζίρου, καθώς αντιπροσωπεύουν ένα δυσανάλογα μεγάλο μερίδιο των συνολικών εσόδων, το οποίο ανέρχεται σε 5,3% έως 7,2% ετησίως. Πρόκειται για είδη υψηλής στρατηγικής

σημασίας, τα οποία συνήθως χαρακτηρίζονται από σταθερή και υψηλή ζήτηση και επηρεάζουν άμεσα τη ρευστότητα της επιχείρησης.

Προκειμένου να αποτυπωθεί η διαχρονική εξέλιξη και η δυναμική αυτών των κρίσιμων κωδικών, στο Σχήμα 4.7 παρουσιάζεται ο χάρτης θερμότητας (*heatmap*) του ετήσιου τζίρου τους για την εξεταζόμενη πενταετία.



Σχήμα 4.7: Κατανομή τζίρου των top-20 προϊόντων

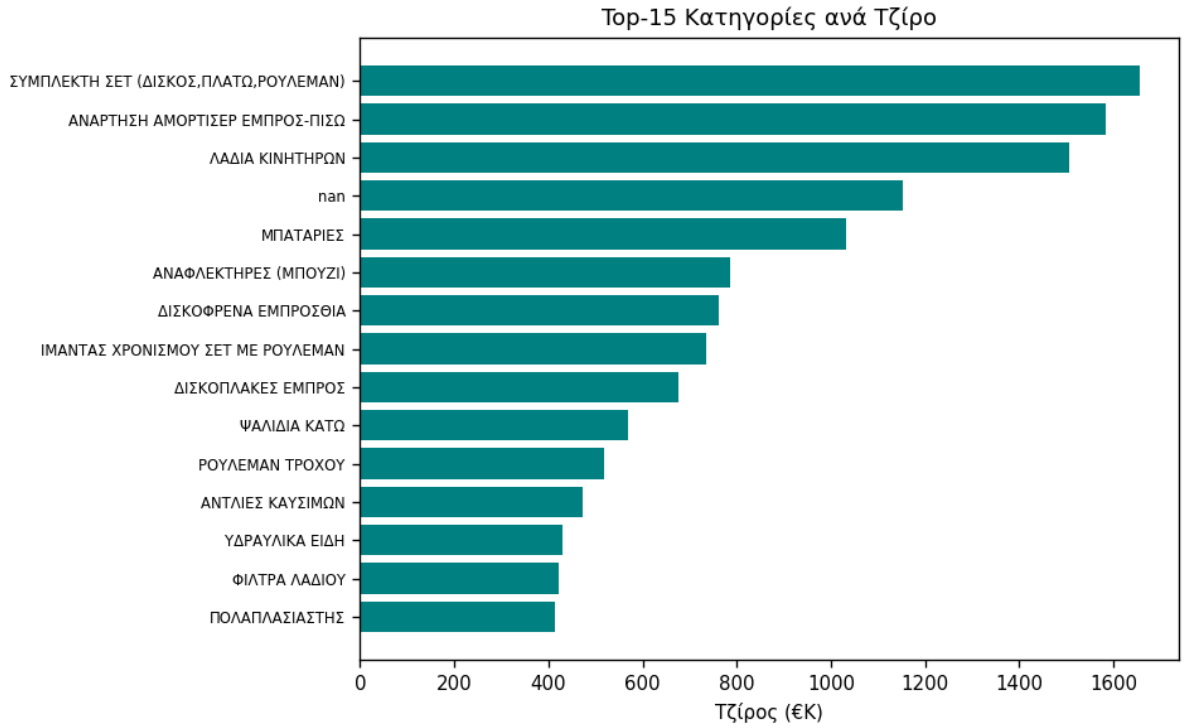
Μέσω της ερμηνείας του χάρτη θερμότητας προκύπτουν σημαντικά συμπεράσματα σχετικά με τον κύκλο ζωής των προϊόντων (*product life cycle*) στη δευτερογενή αγορά αυτοκινήτου:

- **Εισαγωγή νέων προϊόντων υψηλής ζήτησης:** Συγκεκριμένοι κωδικοί (όπως οι `156-05005DPF` και `156-04004`) εμφανίζουν μηδενικές πωλήσεις κατά την περίοδο 2021-2024 και καταγράφουν εξαιρετικά υψηλό τζίρο αποκλειστικά το έτος 2025. Το φαινόμενο αυτό υποδεικνύει την επιτυχή εισαγωγή νέων, κρίσιμων εξαρτημάτων στην αγορά, τα οποία εμφάνισαν απότομη ενεργοποίηση ζήτησης.
- **Αντικατάσταση και απαξίωση κωδικών:** Αντίστροφα, προϊόντα υψηλής κινητικότητας κατά τα πρώτα έτη (όπως το λιπαντικό κινητήρα `550-WR10W40.050`) παρουσιάζουν έντονη δραστηριότητα και υψηλό τζίρο κατά την περίοδο 2021-2024, η οποία όμως μηδενίζεται πλήρως το 2025. Η συμπεριφορά αυτή τεκμηριώνει την αντικατάσταση του κωδικού από νεότερα, συμβατά προϊόντα ή τη σταδιακή απαξίωσή του λόγω τεχνολογικών μεταβολών στον στόλο των οχημάτων.

Οι παρατηρήσεις αυτές αναδεικνύουν ότι ακόμη και εντός της ομάδας των πιο σταθερών και κερδοφόρων προϊόντων της επιχείρησης, η ζήτηση δεν είναι στατική αλλά υπόκειται σε σημαντικές μεταβολές.

4.4 Κατηγορίες προϊόντων

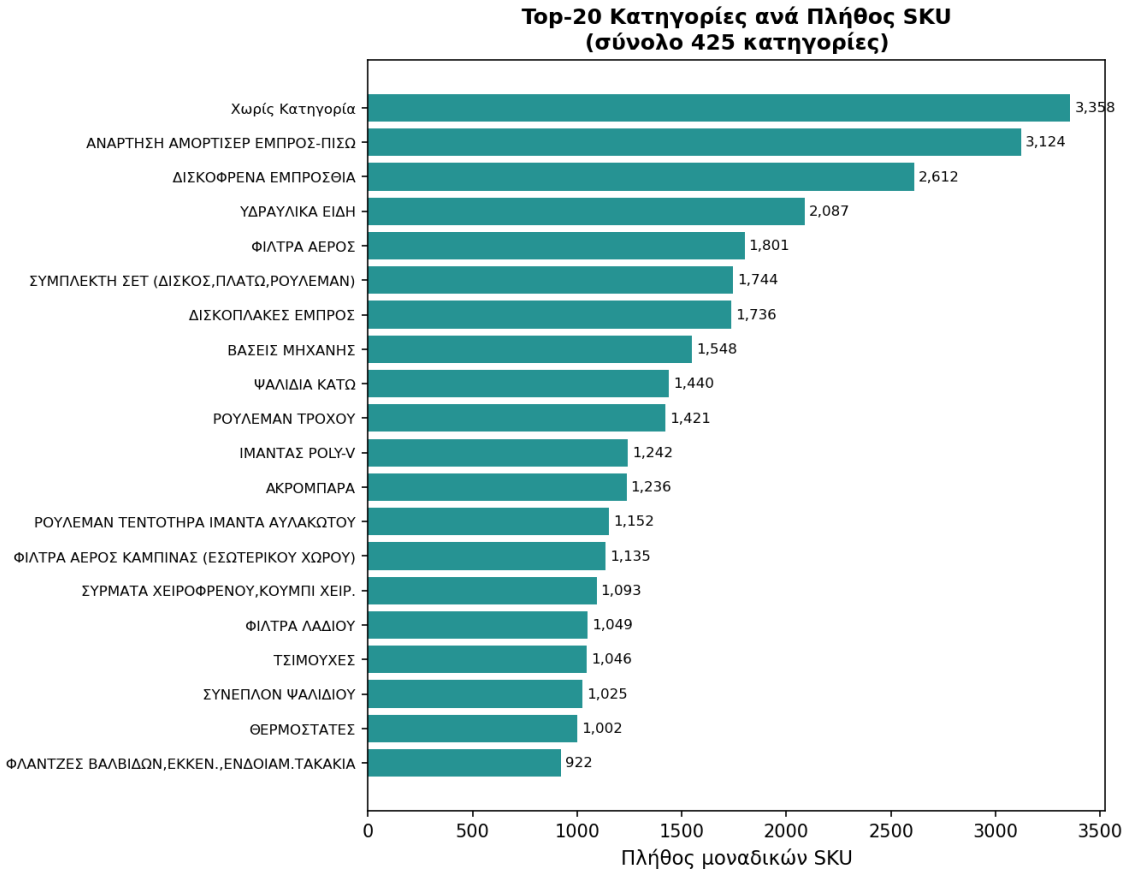
Οι 15 κορυφαίες κατηγορίες ανά τζίρο 5ετίας παρατίθενται στο Σχήμα 4.8. Η πρώτη κατηγορία, σετ συμπλέκτη, αντιπροσωπεύει τζίρο €1,66M, ακολουθούμενη από αμορτισέρ ανάρτησης (€1,58M) και λάδια κινητήρων (€1,51M).



Σχήμα 4.8: Top-15 κατηγορίες βάσει συνολικού τζίρου (€K)

Αξίζει να σημειωθεί ότι υπάρχουν 3.358 μη κατηγοριοποιημένοι κωδικοί (4,6%), οι οποίοι εμφανίζονται ως nan και αθροιστικά αντιπροσωπεύουν τζίρο ~€1,15M που τους κατατάσσει στην τέταρτη θέση της συνολικής λίστας.

Εξίσου σημαντική παρατήρηση προκύπτει από την κατανομή του πλήθους των διακριτών κωδικών (SKU) ανά προϊοντική κατηγορία, η οποία αποκαλύπτει έντονη ετερογένεια. Ενδεικτικά, ο αριθμός των ειδών ανά κατηγορία κυμαίνεται από περισσότερα των 3.000 SKU σε ορισμένες κατηγορίες (όπως τα συστήματα ανάρτησης και οι μη κατηγοριοποιημένοι κωδικοί) έως λιγότερα των 200 SKU σε άλλες (όπως οι συσσωρευτές και τα σετ βολάν). Η δομική αυτή ανομοιογένεια συνιστά μια κρίσιμη πρόκληση για τα μοντέλα Μηχανικής Μάθησης, καθώς οι αλγόριθμοι καλούνται να γενικεύσουν αποτελεσματικά κάτω από δύο αντιδιαμετρικές συνθήκες, αφενός σε υπερπληθείς κατηγορίες που χαρακτηρίζονται από ιδιαίτερα χαμηλό σήμα πληροφορίας ανά κωδικό *(low signal per* SKU), και αφετέρου σε ολιγομελείς κατηγορίες, οι οποίες, ωστόσο, περιλαμβάνουν κωδικούς υψηλής αξίας και επιχειρησιακής κρισιμότητας (*high-value* SKUs). Για την πληρέστερη αποτύπωση αυτής της ανισορροπίας, στο Σχήμα 4.9 παρατίθενται οι 20 πολυπληθέστερες κατηγορίες του καταλόγου ειδών.



Σχήμα 4.9: Οι 20 πολυπληθέστερες κατηγορίες

4.5 Ανάλυση αποθεμάτων

Η αποτιμηθείσα αξία του αποθέματος (*inventory value*) και τα οικονομικά μεγέθη των κωδικών δεν παρέχονται απευθείας στο πρωτογενές σύνολο δεδομένων, αλλά ανακατασκευάζονται αλγοριθμικά μέσω της ακόλουθης διαδικασίας πέντε σταδίων:

- **Ημερήσια καθαρή ροή (*Daily net flow*):** Υπολογίζεται το ημερήσιο ισοζύγιο κίνησης για κάθε κωδικό είδους (SKU), αφαιρώντας τις πωληθείσες ποσότητες (συναλλαγές με την ένδειξη `type=1`) από τις παραληφθείσες ποσότητες αναπλήρωσης (συναλλαγές με την ένδειξη `type=3`).
- **Αθροιστικό απόθεμα (*Cumulative stock*)**: Το τρέχον διαθέσιμο απόθεμα υπολογίζεται μέσω του διαδοχικού αθροίσματος (cumulative sum) της ημερήσιας καθαρής ροής λαμβάνοντας ως σταθερό σημείο αναφοράς (anchor point) το πλέον πρόσφατο στιγμιότυπο φυσικής απογραφής (συναλλαγές με την ένδειξη `type=0`).
- **Μοναδιαίο κόστος κτήσης (*Unit acquisition cost*)**: Εκτιμάται πρωτίστως από τα δεδομένα πωλήσεων (`type=1`) ως το πηλίκο `'Sales_Cost/Trn_Qty'`. Επισημαίνεται ότι η μεταβλητή `Sales_Cost` αντιστοιχεί αυστηρά στο Κόστος Πωληθέντων (*Cost of Goods Sold* – COGS) και όχι στην τιμή πώλησης. Η παραδοχή αυτή επαληθεύεται μέσω της λογιστικής ταυτότητας `'Sales_Gross_Margin = Trn_Net_Value - Sales_Cost'`, η οποία αποδίδει ένα διάμεσο περιθώριο μικτού κέρδους της τάξης του 36%. Για κωδικούς χωρίς καταγεγραμμένες πωλήσεις, το μοναδιαίο κόστος προσεγγίζεται εναλλακτικά μέσω των συναλλαγών αναπλήρωσης (`type=3`), ως το πηλίκο `'Trn_Net_Value/Trn_Qty'`, το οποίο αντικατοπτρίζει πιστά την τιμή αγοράς από τον προμηθευτή στην ίδια βάση κόστους.

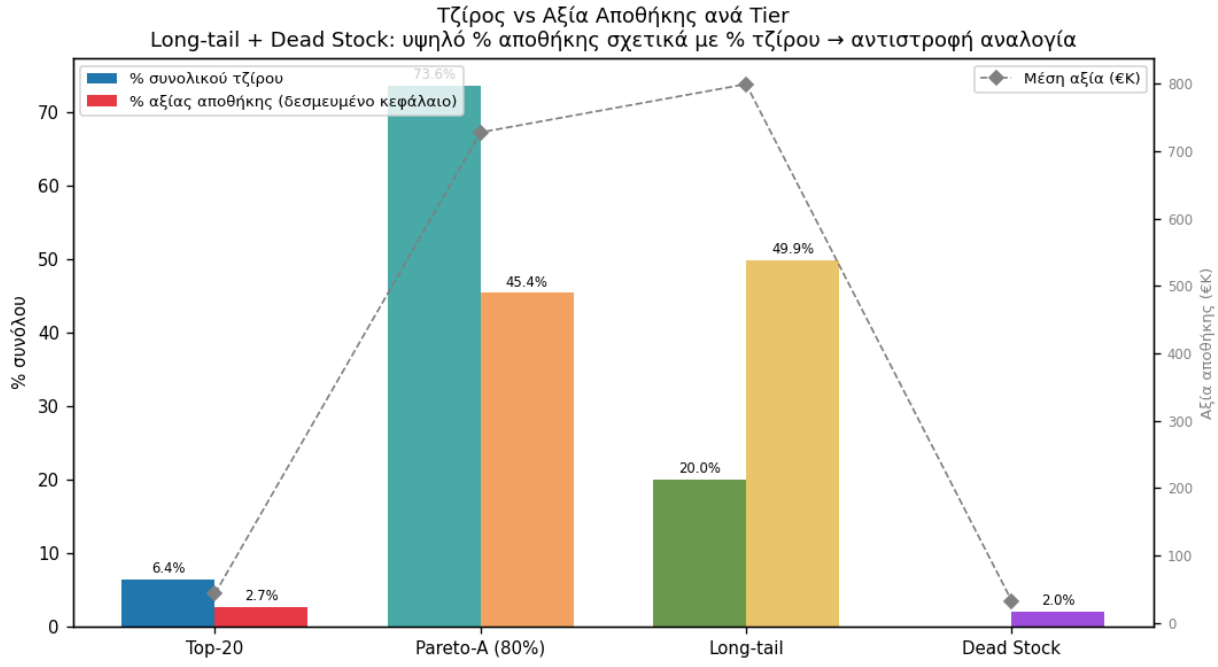
- **Κύκλος εργασιών (*Sales revenue*)**: Ο παραγόμενος τζίρος αντλείται από το πεδίο `Trn_Net_Value` των συναλλαγών πώλησης (`type=1`), το οποίο αποτυπώνει την τελική αξία πώλησης προς τον πελάτη.
- **Αποτιμηθείσα αξία αποθέματος (*Inventory valuation*)**: Το τελικό μέγεθος εξάγεται ως το γινόμενο του μέσου διαθέσιμου αποθέματος της εκάστοτε περιόδου επί το εκτιμώμενο μοναδιαίο κόστος κτήσης (τιμή αγοράς).

Στον Πίνακα 4.4 αποτυπώνεται η μέση αξία της αποθήκης, αναλυτικά ανά κατηγορία εσόδων (*revenue tier*). Σημειώνεται ότι η συνολική μέση αξία του διαθέσιμου αποθέματος για το υπό εξέταση σύνολο δεδομένων ανέρχεται σε 1.603.125 €.

Πίνακας 4.4: Μέση αξία αποθήκης ανά κατηγορία τζίρου

Κατηγορία τζίρου	Πλήθος SKU	% Τζίρου	% Αξίας Αποθήκης	Λόγος Αποθήκης/Τζίρου	Μέρες Κάλυψης
Top-20	20	6,4%	2,7%	0,030	55
Pareto-A (80%)	16.368	73,6%	45,4%	0,044	80
Long-tail	55.922	20,0%	49,9%	0,177	323
Dead stock	6.883	0,0%	2,1%	—	—

Η οπτικοποίηση των δεδομένων αυτών (Σχήμα 4.10) αναδεικνύει ένα κρίσιμο επιχειρησιακό φαινόμενο, την αντιστροφή της αναλογίας μεταξύ παραγόμενου τζίρου και δεσμευμένου κεφαλαίου αποθέματος, καθώς μεταβαίνουμε από τους κορυφαίους κωδικούς προς την ουρά του καταλόγου.



Σχήμα 4.10: Αντιστροφή αναλογίας μεταξύ τζίρου και αξίας αποθήκης

Από τη συνδυαστική ανάλυση του πίνακα και του γραφήματος εξάγονται τα ακόλουθα κεντρικά συμπεράσματα:

- **Τα Top-20 SKU** (τα οποία παράγουν το 6,4% του τζίρου) δεσμεύουν μόλις το 2,7% της αξίας του αποθέματος. Ο μέσος χρόνος παραμονής τους στο ράφι (μέρες κάλυψης) περιορίζεται στις

55 ημέρες, γεγονός που υποδεικνύει εξαιρετικά ταχεία κυκλοφορία και υψηλή αποδοτικότητα κεφαλαίου.

- **Τα *long-tail* SKU**, παρότι συνεισφέρουν μόλις το 20% του τζίρου, δεσμεύουν σχεδόν το ήμισυ (49,9%) της συνολικής αξίας της αποθήκης. Ο μέσος χρόνος παραμονής τους εκτοξεύεται στις 323 ημέρες (περίπου 10,8 μήνες), αναδεικνύοντας τον εξαιρετικά αργό ρυθμό ανακύκλωσής τους.
- **Το νεκρό απόθεμα (*dead stock*)**, το οποίο αποτελείται από 6.883 ανενεργούς κωδικούς με μηδενικές πωλήσεις καθ' όλη τη διάρκεια της εξεταζόμενης περιόδου, εξακολουθεί να δεσμεύει αδρανές κεφάλαιο ύψους περίπου 32.800 € (ήτοι το 2,1% της συνολικής αξίας).

Η παραμονή των προϊόντων της μακράς ουράς στο ράφι για σχεδόν ένα έτος, αποτυπώνει επακριβώς το κεντρικό επιχειρησιακό πρόβλημα του δικτύου, το οποοίο είναι η συσσώρευση υπερβολικού αποθέματος αργά κινούμενων ειδών (*slow movers*). Η κατάσταση αυτή εγκλωβίζει σημαντικό κεφάλαιο κίνησης (*working capital*) και εγκυμονεί κινδύνους ρευστότητας. Επομένως, η ακριβής πρόβλεψη της ζήτησης σε αυτά τα εξαιρετικά αραιά πωλούμενα είδη αποτελεί μονόδρομο για τον εξορθολογισμό των αποφάσεων αναπλήρωσης.

Προκειμένου να ποσοτικοποιηθεί η αξία της αλγοριθμικής βελτιστοποίησης, αρκεί να σημειωθεί ότι, βάσει των ανωτέρω στοιχείων, μια σχετική μείωση κατά 15% στα αποθέματα αποκλειστικά της κατηγορίας *long-tail* (δηλαδή η πτώση του μεριδίου τους από το 49,9% στο ~42,4% της συνολικής αποθήκης) θα απελευθέρωνε άμεσα ρευστότητα της τάξης των 120.000 € ($0{,}15 \times 0{,}499 \times 1.603.125€$). Πρόκειται για ένα ποσό εξαιρετικά σημαντικό για την εξασφάλιση ρευστότητας της επιχείρησης.

4.6 Ανάλυση και κατηγοριοποίηση ζήτησης

4.6.1 Ταξινόμηση κατά Syntetos–Boylan (SBC Matrix)

Η κατηγοριοποίηση των κωδικών (SKU) βάσει της μεθοδολογίας Syntetos–Boylan–Croston (SBC) συνιστά ένα από τα πλέον καθιερωμένα εργαλεία στην ανάλυση αποθεμάτων. Το συγκεκριμένο πλαίσιο ταξινόμησης στηρίζεται στον υπολογισμό δύο θεμελιωδών μεγεθών:

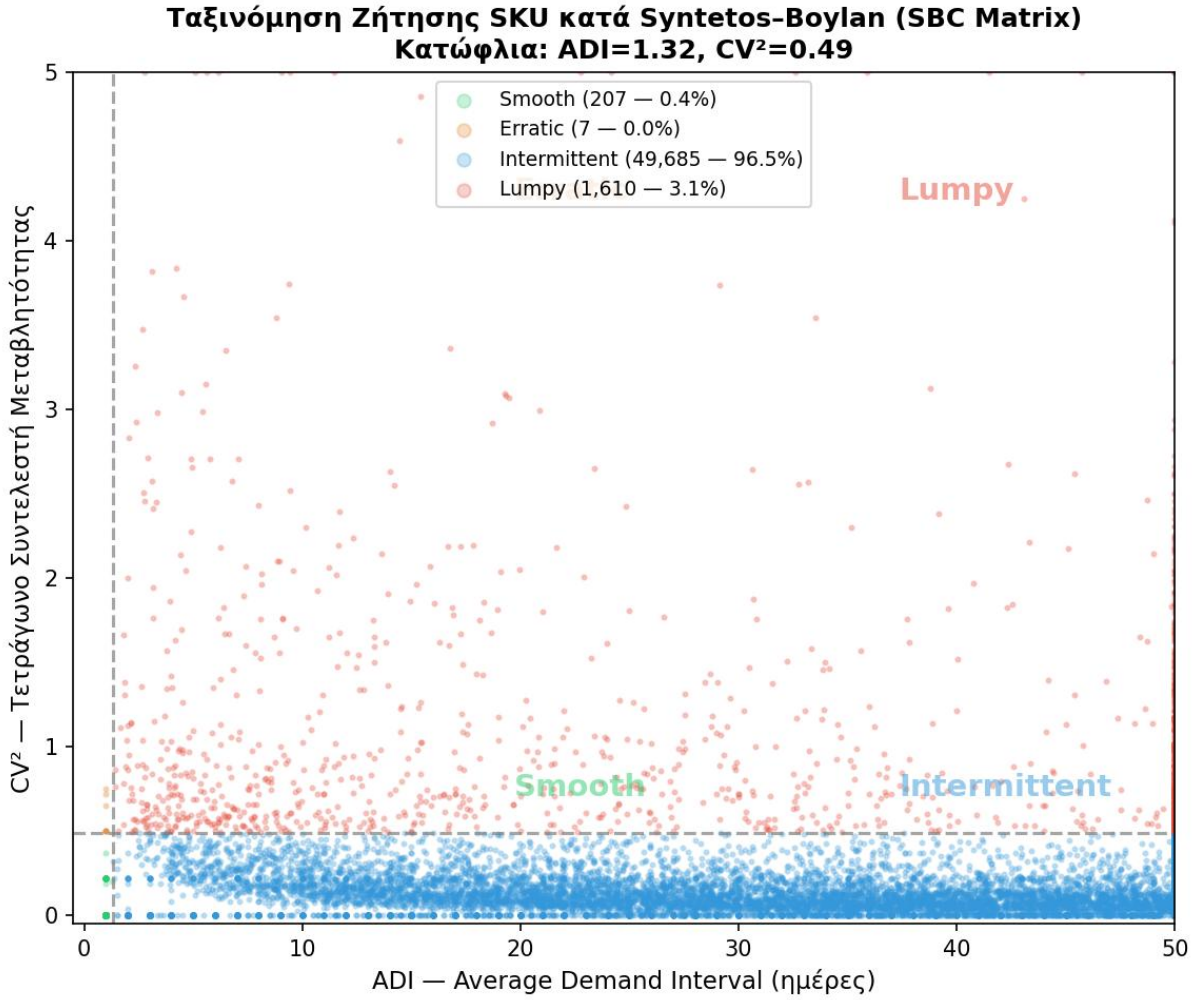
- **ADI (Average Demand Interval):** Το μέσο χρονικό διάστημα (εκφρασμένο σε ημέρες) μεταξύ διαδοχικών γεγονότων μη μηδενικής ζήτησης.
- **CV^2 (Squared Coefficient of Variation):** Το τετράγωνο του συντελεστή μεταβλητότητας των μη μηδενικών ποσοτήτων ζήτησης.

Σύμφωνα με την αρχική μελέτη των Syntetos και Boylan (2005), τα διαχωριστικά κατώφλια (thresholds) ορίζονται στις τιμές $ADI = 1{,}32$ και $CV^2 = 0{,}49$. Η εφαρμογή αυτών των κριτηρίων στο υποσύνολο των 51.509 SKU που παρουσιάζουν τουλάχιστον δύο γεγονότα πώλησης (απαραίτητη συνθήκη για τον υπολογισμό της διακύμανσης), αποδίδει την ταξινόμηση που συνοψίζεται στον Πίνακα 4.5 και οπτικοποιείται στο Σχήμα 4.11.

Πίνακας 4.5: Ταξινόμηση ζήτησης κατά Syntetos – Boylan

Κατηγορία	Χαρακτηριστικό	Πλήθος SKU	%
Ομαλή (Smooth) $\boldsymbol{ADI \leq 1,32, CV^2 \leq 0,49}$	Συνεχής ροή, χαμηλή διακύμανση ποσότητας παραγγελιών	207	0,4%

Κατηγορία	Χαρακτηριστικό	Πλήθος SKU	%
Ασταθής / Ακανόνιστη (Erratic) $\boldsymbol{ADI \leq 1,32, CV^2 > 0,49}$	Συνεχής ροή, εξαιρετικά υψηλή διακύμανση ποσότητας παραγγελιών	7	0,0%
Διακοπτόμενη / Διαλείπουσα (Intermittent) $\boldsymbol{ADI > 1,32, CV^2 \leq 0,49}$	Υψηλή σποραδικότητα εμφάνισης, σταθερότητα ποσότητας παραγγελιών	49.685	**96,5%**
Ανομοιόμορφη (Lumpy) $\boldsymbol{ADI > 1,32, CV^2 > 0,49}$	Συνδυασμός ακραίας σποραδικότητας και υψηλής διακύμανσης μεγέθους	1.610	3,1%



Σχήμα 4.11: SBC Matrix: Ταξινόμηση ζήτησης

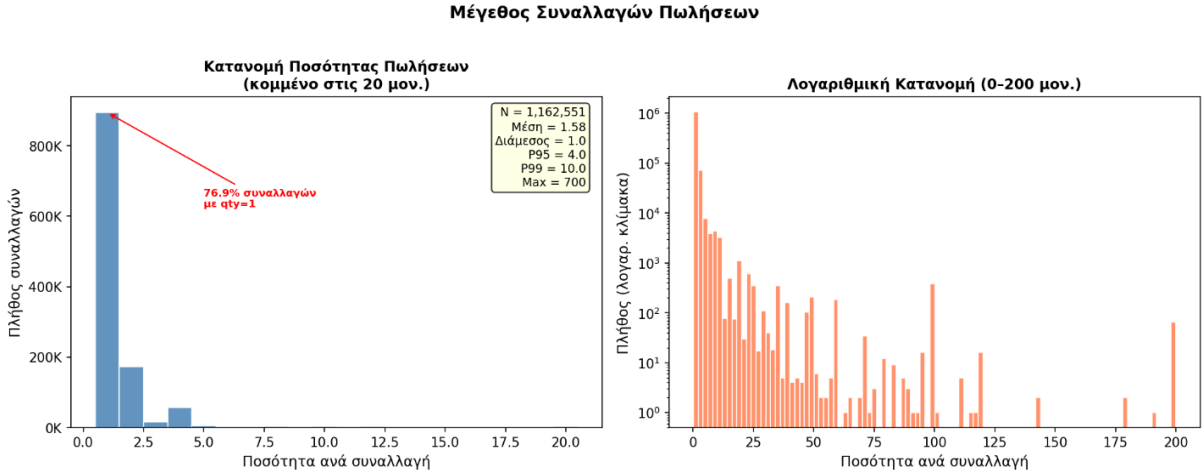
Το συντριπτικό ποσοστό (96,5%) των SKU ταξινομείται στην κατηγορία της διαλείπουσας (*intermittent*) ζήτησης. Αυτό σημαίνει ότι η ζήτηση προκύπτει ιδιαίτερα σποραδικά, ωστόσο, όταν εκδηλώνεται, το μέγεθος της παραγγελίας παραμένει σχετικά σταθερό. Η συγκεκριμένη κατανομή αντικατοπτρίζει απόλυτα τη θεμελιώδη φύση της δευτερογενούς αγοράς ανταλλακτικών, στην οποία η πλειονότητα των εξαρτημάτων εμφανίζει χαμηλή συχνότητα πώλησης (καθώς εξαρτάται από τον ρυθμό βλαβών συγκεκριμένων μοντέλων οχημάτων), αλλά η απαιτούμενη ποσότητα περιορίζεται σχεδόν αποκλειστικά στη μία μονάδα ανά γεγονός αντικατάστασης.

Στον αντίποδα, μόλις το 0,4% των κωδικών (207 είδη) παρουσιάζει ομαλή (smooth) ζήτηση. Όπως είναι αναμενόμενο για τον κλάδο, η εν λόγω κατηγορία απαρτίζεται σχεδόν αποκλειστικά από αναλώσιμα υλικά υψηλής κυκλοφορίας (όπως φίλτρα, λιπαντικά και αναλώσιμα συντήρησης), τα οποία χαρακτηρίζονται από συνεχή ροή πωλήσεων και απουσία παρατεταμένης στασιμότητας.

4.6.2 Ανάλυση συναλλαγών

Η ανάλυση της κατανομής των ποσοτήτων ανά μεμονωμένη συναλλαγή πώλησης επιβεβαιώνει την ακραία σποραδικότητα που χαρακτηρίζει τον κατάλογο. Τα βασικά στατιστικά μεγέθη της διανομής συνοψίζονται ως εξής:

- Μέση ποσότητα παραγγελίας: 1,58 μονάδες.
- Διάμεση ποσότητα παραγγελίας: 1,0 μονάδα.
- Συγκέντρωση: Το 76,9% των συναλλαγών αφορά ποσότητα ακριβώς 1 μονάδας.
- Εκατοστημόρια: Το 95ο εκατοστημόριο (P_{95}) εντοπίζεται στις 4 μονάδες, το 99ο εκατοστημόριο (P_{99}) στις 10 μονάδες, ενώ η μέγιστη καταγεγραμμένη τιμή ανέρχεται στις 700 μονάδες.



Σχήμα 4.12: Κατανομή ποσότητας πωλήσεων

Το Σχήμα 4.12 αποτελείται από δύο επιμέρους γραφήματα, τα οποία απεικονίζουν την ίδια κατανομή σε διαφορετικές κλίμακες (scales), προκειμένου να αναδείξουν διακριτές πτυχές της υποκείμενης στατιστικής συμπεριφοράς:

- **Αριστερό γράφημα – Γραμμική κλίμακα (0–20 μονάδες):** Το ιστόγραμμα αποκαλύπτει την απόλυτη κυριαρχία της μοναδιαίας πώλησης. Η στήλη για $Q = 1$ δεσπόζει στον κατακόρυφο άξονα καταγράφοντας περίπου 895.000 συναλλαγές, ήτοι το 76,9% του συνόλου. Αυτό καθιστά τις υπόλοιπες ποσότητες οπτικά αμελητέες. Συγκεκριμένα, οι παραγγελίες 2 τεμαχίων αποτελούν το ~13,5% (~157.000 εγγραφές), 3 τεμαχίων το ~4,5% (~52.000), 4–5 τεμαχίων το ~3% (~35.000), ενώ ποσότητες άνω των 5 τεμαχίων αφορούν μόλις το ~2,1% των συναλλαγών (~24.000). Η σημαντική απόκλιση μεταξύ της μέσης (1,58) και της διάμεσης (1,0) τιμής υποδεικνύει την ύπαρξη ισχυρής θετικής λοξότητας (right skew), καθώς ορισμένες μεμονωμένες, υπερμεγέθεις παραγγελίες εκτρέπουν τον μέσο όρο προς τα πάνω.
- **Δεξί γράφημα – Λογαριθμική κλίμακα (0–200 μονάδες):** Μέσω της λογαριθμικής κλίμακας, η δομή των ακραίων τιμών καθίσταται ορατή. Παρατηρείται ότι εξαιρετικά μεγάλες παραγγελίες ($Q = 50, 100$ ή 200) εμφανίζονται με αξιοσημείωτη συχνότητα (10 έως 100 φορές) επιβεβαιώνοντας μια συμπεριφορά «βαριάς ουράς» (heavy-tail). Σε ορισμένα τμήματά

της, η κατανομή προσεγγίζει το μοτίβο του Νόμου της Δύναμης (*Power-law*, Clauset et al., 2009), όπου κάθε διπλασιασμός της ποσότητας επιφέρει αναλογική μείωση της συχνότητας εμφάνισης. Παρά τη σπανιότητά τους, αυτές οι μεγάλες παραγγελίες (π.χ. $Q = 100 - 700$) συνεισφέρουν δυσανάλογα στον συνολικό όγκο της ζήτησης.

Επιπτώσεις στη Μοντελοποίηση: Η συντριπτική κυριαρχία της μοναδιαίας παραγγελίας (76,9% με $Q = 1$) συνεπάγεται ότι το πρόβλημα της πρόβλεψης σε ημερήσιο επίπεδο μετασχηματίζεται κατά βάση σε πρόβλημα δυαδικής ταξινόμησης (*binary classification*), δηλαδή εκτίμησης της πιθανότητας εκδήλωσης οποιασδήποτε δραστηριότητας. Το εύρημα αυτό αποτελεί την ισχυρότερη εμπειρική τεκμηρίωση για την υιοθέτηση αρχιτεκτονικής Εμποδίου δύο σταδίων (*two-stage* Hurdle *model*) κατά τον σχεδιασμό του συστήματος πρόβλεψης. Υπό αυτό το πρίσμα, ένα πρωτεύον μοντέλο ταξινόμησης (*classifier*) εκτιμά αρχικά την πιθανότητα εμφάνισης ζήτησης $P(y > 0)$, ενώ ένα ανεξάρτητο μοντέλο παλινδρόμησης (*regressor*) ενεργοποιείται υπό συνθήκη και αναλαμβάνει να εκτιμήσει το μέγεθος της ζήτησης ενσωματώνοντας και μαθαίνοντας ρητά τη συμπεριφορά της «βαριάς ουράς» για τα εν λόγω προϊόντα.

4.6.3 Ποσοστό μηδενικής ζήτησης (Zero-Demand Fraction)

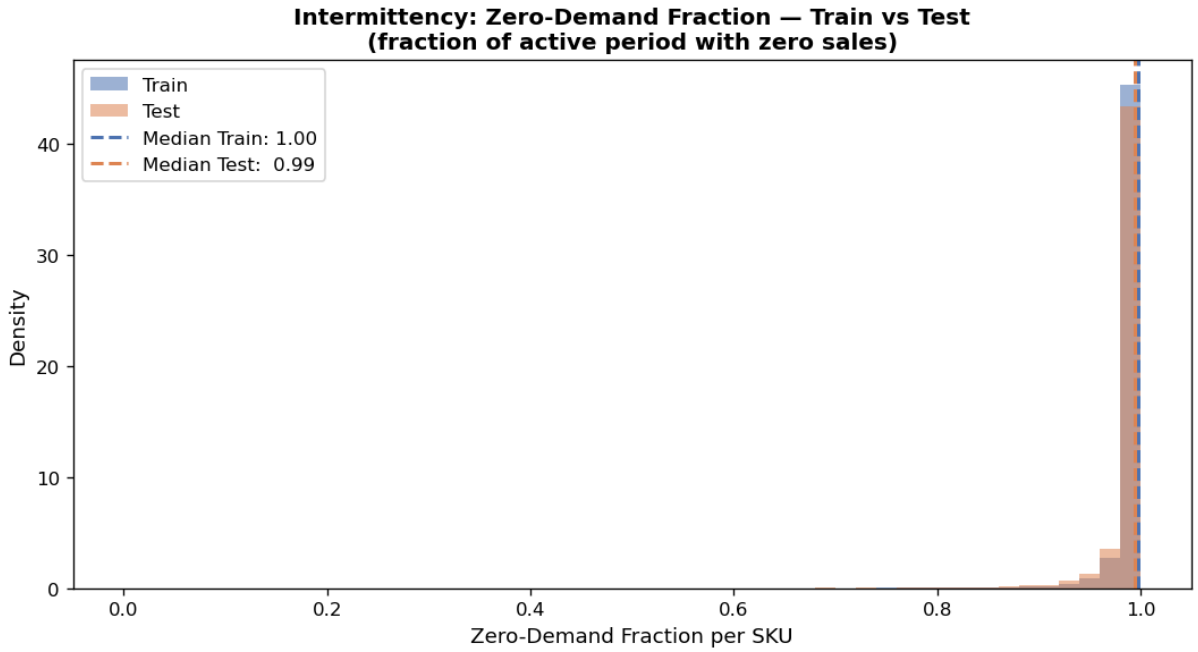
Το Ποσοστό Μηδενικής Ζήτησης (*Zero-Demand Fraction* – ZDF) αποτελεί έναν κρίσιμο στατιστικό δείκτη που αποτυπώνει τον βαθμό αδράνειας ενός κωδικού. Ουσιαστικά, υποδεικνύει το ποσοστό του χρόνου, κατά τον οποίο ένα προϊόν δεν παρουσιάζει καμία απολύτως πώληση.

Ο μαθηματικός τύπος υπολογισμού του δείκτη ορίζεται ως εξής:

$$ZDF = \frac{Ημέρες\ με\ μηδενικές\ πωλήσεις}{Συνολικές\ ημέρες\ που\ είναι\ διαθέσιμο\ το\ προϊόν} \tag{4.2}$$

Βάσει του ορισμού αυτού, τιμή του δείκτη ίση με 1,0 υποδηλώνει πλήρη απουσία πωλήσεων (πλήρως ανενεργός κωδικός για το συγκεκριμένο χρονικό παράθυρο), ενώ τιμή ίση με 0,0 υποδεικνύει αδιάλειπτη, καθημερινή ζήτηση.

Στο Σχήμα 4.13 παρουσιάζεται το ιστόγραμμα πυκνότητας, το οποίο απεικονίζει την κατανομή του συγκεκριμένου δείκτη στο σύνολο του προϊοντικού καταλόγου.



Σχήμα 4.13: Κατανομή zero-demand fraction ανά SKU

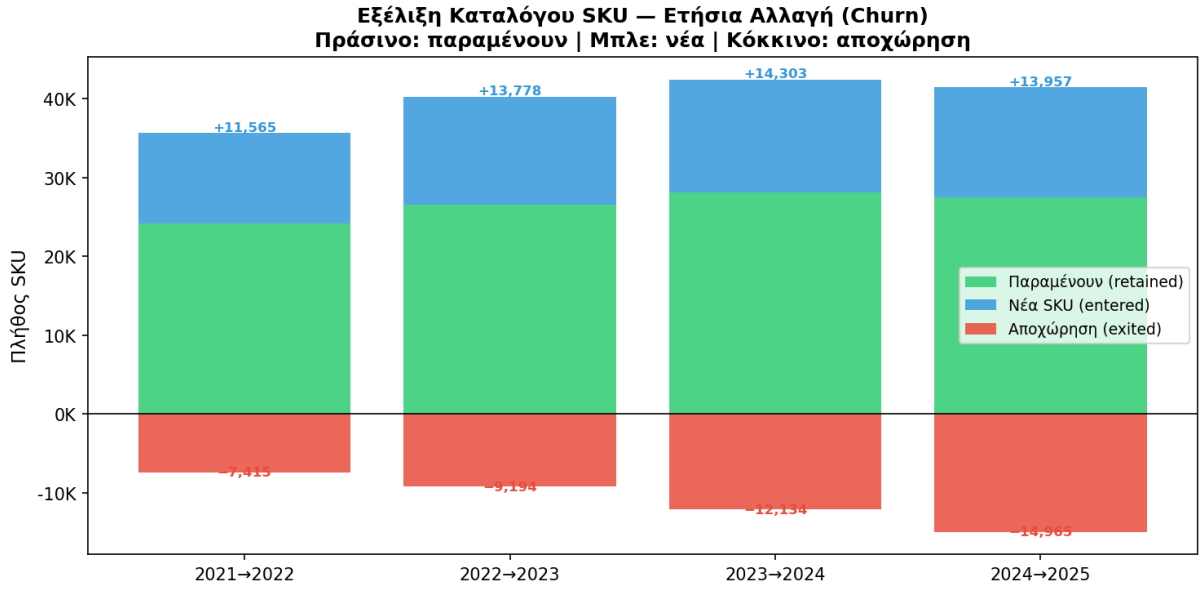
Το θεμελιώδες συμπέρασμα που εξάγεται από το ανωτέρω γράφημα είναι ότι η απουσία πωλήσεων συνιστά τον κανόνα, και όχι την εξαίρεση, στο υπό εξέταση σύνολο δεδομένων. Η κατανομή είναι ακραία ασύμμετρη προς τα αριστερά (*left-skewed*), με τη συντριπτική πλειονότητα των κωδικών να συγκεντρώνεται στο δεξί άκρο του άξονα (κοντά στο 1,0). Συγκεκριμένα, για έναν τυχαία επιλεγμένο κωδικό, παρατηρείται μηδενική ζήτηση κατά το 99% των διαθέσιμων ημερών. Το φαινόμενο αυτό (zero-inflation) συνάδει πλήρως με τα δομικά χαρακτηριστικά των αγορών διαλείπουσας ζήτησης, όπως αυτή της δευτερογενούς αγοράς ανταλλακτικών, και επιβεβαιώνει εκ νέου τη μαθηματική αναγκαιότητα χρήσης εξειδικευμένων μοντέλων Εμποδίου έναντι των κλασικών αλγορίθμων παλινδρόμησης.

4.6.4 Δυναμική καταλόγου προϊόντων (Churn analysis)

Ο κατάλογος των προϊόντων της επιχείρησης χαρακτηρίζεται από έντονη δυναμική και ρευστότητα και δεν μπορεί σε καμία περίπτωση να θεωρηθεί στατικός. Σε ετήσια βάση, παρατηρούνται ριζικές διαφοροποιήσεις μέσω της εισόδου νέων κωδικών (SKU) και της αποχώρησης παλαιότερων, λόγω διακοπής παραγωγής, απαξίωσης ή εξάντλησης. Στον Πίνακα 4.6 και το αντίστοιχο Σχήμα 4.14 αποτυπώνεται αναλυτικά η ετήσια αυτή μεταβολή (*churn*).

Πίνακας 4.6: Ετήσια μεταβολή καταλόγου προϊόντων

Μετάβαση	Παραμένουν	Νέα	Αποχώρηση	Καθαρή Μεταβολή
2021→2022	~24.000	+11.565	−7.815	+3.750
2022→2023	~27.000	+13.778	−9.194	+4.584
2023→2024	~27.500	+14.303	−12.134	+2.169
2024→2025	~27.000	+13.957	−14.963	−1.006



Σχήμα 4.14: Ετήσια μεταβολή καταλόγου προϊόντων

Από τα ανωτέρω δεδομένα προκύπτει μια μέση ετήσια εισαγωγή περίπου 13.500 νέων κωδικών. Στον αντίποδα, καταγράφεται μια μέση ετήσια αποχώρηση άνω των 11.000 ειδών, η οποία μάλιστα παρουσιάζει ισχυρή αυξητική τάση.

Ιδιαίτερο επιχειρησιακό ενδιαφέρον παρουσιάζει η μετάβαση από το 2024 στο 2025, καθώς αποτελεί την πρώτη εξεταζόμενη περίοδο, κατά την οποία καταγράφεται καθαρή συρρίκνωση του καταλόγου. Το γεγονός αυτό υποδεικνύει την πιθανή εφαρμογή μιας στρατηγικής εξορθολογισμού (*rationalization*) ή εκκαθάρισης του χαρτοφυλακίου από πλευράς της επιχείρησης, κατά την οποία αποσύρονται παρωχημένοι ή εξαιρετικά αργά κινούμενοι κωδικοί με ταχύτερο ρυθμό από τον αντίστοιχο ρυθμό εισαγωγής νέων.

4.7 Στατιστική Σύγκριση Συνόλων Εκπαίδευσης και Αξιολόγησης

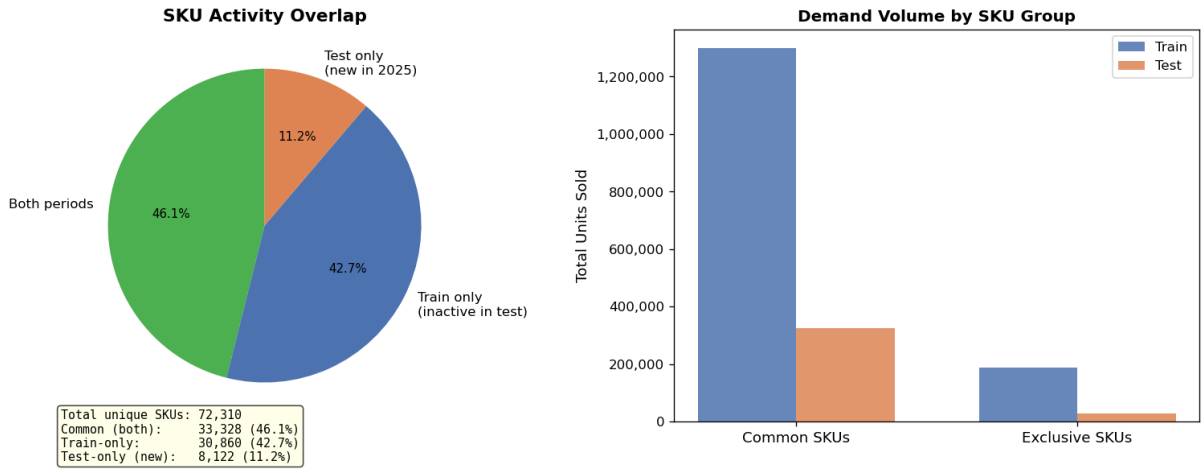
Για την ανάπτυξη και την ορθή αξιολόγηση των αλγορίθμων Μηχανικής Μάθησης (Κεφάλαιο 5), το σύνολο των δεδομένων διαχωρίστηκε χρονικά ως εξής:

- Περίοδος Εκπαίδευσης (Train Set): 01/01/2021 – 31/12/2024 (4 έτη)
- **Περίοδος Αξιολόγησης (Test Set):** 01/01/2025 – 31/12/2025 (1 έτος)
- Σημείο Αναφοράς Πρόβλεψης (Forecast Origin): 01/01/2025

Ένα από τα πλέον κρίσιμα ευρήματα της στατιστικής σύγκρισης αφορά τη μερική αντιστοιχία των ενεργών κωδικών (SKU) μεταξύ των δύο αυτών περιόδων, η οποία συνοψίζεται στον Πίνακα 4.7 και το Σχήμα 4.15.

Πίνακας 4.7: Επικάλυψη ενεργών SKU μεταξύ περιόδων εκπαίδευσης και αξιολόγησης

Κατηγορία	Πλήθος SKU	% Συνόλου
Ενεργά και στις δύο περιόδους	33.328	46,1%
Μόνο Train (ανενεργά το 2025)	30.860	42,7%
Μόνο Test (νέα το 2025)	8.122	11,2%
Σύνολο	**72.310**	100%



Σχήμα 4.15: Επικάλυψη ενεργών SKU μεταξύ περιόδων εκπαίδευσης και αξιολόγησης

Όπως διαπιστώνεται, μόλις το 46,1% των συνολικών SKU εμφανίζει δραστηριότητα πωλήσεων και στις δύο χρονικές περιόδους. Αντιθέτως, 30.860 SKU (ήτοι το 42,7% του συνόλου), τα οποία διέθεταν ιστορικό πωλήσεων κατά την περίοδο εκπαίδευσης, κατέστησαν πλήρως ανενεργά το 2025. Η παρατήρηση αυτή αναδεικνύει μια εξαιρετικά σοβαρή πηγή ολίσθησης κατανομών (*distribution shift*), καθότι τα μοντέλα μηχανικής μάθησης ενδέχεται να προβλέψουν κατά την εκπαίδευση μη μηδενική ζήτηση για αυτά τα ανενεργά SKU προσθέτοντας «πλασματική ζήτηση» (*phantom demand*) στις αθροιστικές προβλέψεις.

4.8 Συμπεράσματα από τη στατιστική ανάλυση και επιπτώσεις στη σχεδίαση του συστήματος

Η ενδελεχής στατιστική και οικονομική ανάλυση του συνόλου δεδομένων πενταετίας (2021–2025) για το δίκτυο χονδρικής διανομής ανταλλακτικών αυτοκινήτων αποκαλύπτει τη σύνθετη φύση της υποκείμενης ζήτησης. Τα εμπειρικά ευρήματα δεν αποτελούν απλώς περιγραφικά στατιστικά στοιχεία, αλλά υπαγορεύουν άμεσα τις αρχιτεκτονικές προδιαγραφές, τη μηχανική χαρακτηριστικών (*feature engineering*) και τις μεθοδολογικές επιλογές του συστήματος πρόβλεψης που αναπτύσσεται στο επόμενο κεφάλαιο (Κεφάλαιο 5).

Τα κεντρικά συμπεράσματα της ανάλυσης και οι άμεσες επιπτώσεις τους στη μοντελοποίηση συνοψίζονται στους ακόλουθους άξονες:

- **Ακραία σποραδικότητα και δυαδική φύση της ζήτησης.** Η ανάλυση μέσω της μήτρας Syntetos–Boylan (SBC *Matrix*) κατέδειξε ότι το συντριπτικό 96,5% των κωδικών (49.685 SKU) ταξινομείται στην κατηγορία της διαλείπουσας (*intermittent*) ζήτησης. Επίσης, από το ποσοστό μηδενικής ζήτησης (ZDF) επιβεβαιώνεται ότι για ένα τυχαίως επιλεγμένο προϊόν παρατηρείται πλήρης απουσία πωλήσεων κατά το 99% των ημερών (*zero-inflation*). Παράλληλα, η μικρο-ανάλυση των συναλλαγών αποκάλυψε ότι το 76,9% των γεγονότων πώλησης αφορά ποσότητα ακριβώς μίας (1) μονάδας. Τα ευρήματα αυτά μετατρέπουν το πρόβλημα της ημερήσιας πρόβλεψης κυρίως σε δυαδικό ερώτημα δραστηριότητας. Συνεπώς, δικαιολογείται απόλυτα η απόρριψη των κλασικών μοντέλων παλινδρόμησης και η **υιοθέτηση αρχιτεκτονικής Εμποδίου δύο σταδίων**, όπου ένα πρωτεύον μοντέλο ταξινόμησης εκτιμά την πιθανότητα εμφάνισης ζήτησης $\mathbb{P}(y > 0)$, ώστε να αποτραπεί το συστηματικό σφάλμα υποεκτίμησης (under-forecasting bias).
- **Χρηματοοικονομική αντιστροφή και η πρόκληση της μακράς ουράς (*long-tail*).** Όπως αναλύθηκε και νωρίτερα, η συγκέντρωση του τζίρου ακολουθεί την Αρχή του Pareto (Δείκτης

Gini = 0,742, με το 22,7% των SKU να παράγει το 80% των εσόδων). Από την άλλη, η αποτίμηση του αποθέματος παρουσιάζει μια πλήρη και προβληματική αντιστροφή. Τα προϊόντα της μακράς ουράς, αν και συνεισφέρουν μόλις το 20% του κύκλου εργασιών, δεσμεύουν το 49,9% της συνολικής αξίας της αποθήκης, καταγράφοντας έναν εξαιρετικά αργό ρυθμό ανακύκλωσης με μέσο χρόνο παραμονής στο ράφι τις 323 ημέρες (~10,8 μήνες). Η συσσώρευση αυτή εγκλωβίζει πολύτιμο κεφάλαιο κίνησης (*working capital*). Ποσοτικοποιώντας τον στόχο, αποδείχθηκε ότι μια αλγοριθμικά καθοδηγούμενη μείωση των αποθεμάτων ασφαλείας κατά 15% αποκλειστικά στην κατηγορία *long-tail*, θα απελευθέρωνε άμεσα **ρευστότητα ύψους 120.000 €**. Το εν λόγω εύρημα αναδεικνύει την υψηλή επιχειρησιακή αξία της παρούσας μελέτης.

- **Ρευστότητα καταλόγου και ο κίνδυνος πλασματικής ζήτησης (*phantom demand*).** Ο προϊοντικός κατάλογος παρουσιάζει έντονη ετήσια μεταβολή, με περίπου 13.500 εισαγωγές και 11.000 αποχωρήσεις κωδικών ανά έτος. Η σύγκριση μεταξύ των συνόλων εκπαίδευσης και αξιολόγησης αποκάλυψε ότι το 42,7% των SKU (30.860 κωδικοί) που διέθεταν ιστορικό πωλήσεων κατά την περίοδο εκπαίδευσης (2021-2024), κατέστησαν πλήρως ανενεργά κατά την περίοδο αξιολόγησης (2025). Η στατιστική αυτή πραγματικότητα εγκυμονεί τον κίνδυνο εισαγωγής πλασματικής ζήτησης (*phantom demand*) στις αθροιστικές προβλέψεις.
- **Ενσωμάτωση κυκλικής εποχικότητας.** Τέλος, η ανάλυση των θηκογραμμάτων (*box-plots*) επιβεβαίωσε την ύπαρξη έντονης κυκλικής εποχικότητας στην ελληνική αγορά, με χαρακτηριστικές αιχμές κατά τους μήνες προετοιμασίας των οχημάτων (Ιούνιος, Ιούλιος, Σεπτέμβριος) και βαθιά ύφεση τον Αύγουστο. Η διαπίστωση αυτή επιβάλλει τη **ρητή εισαγωγή ημερολογιακών και χρονικών χαρακτηριστικών** στο σύστημα, προκειμένου να καταστεί δυνατή η έγκυρη μεσοπρόθεσμη και μακροπρόθεσμη πρόβλεψη.

Συνολικά, η στατιστική χαρτογράφηση του Κεφαλαίου 4 αποδεικνύει ότι η διαχείριση των αποθεμάτων της επιχείρησης δεν μπορεί να βασιστεί σε παραδοσιακές, στατικές μεθόδους. Τα συμπεράσματα αυτά θέτουν τις βάσεις και δικαιολογούν πλήρως την αρχιτεκτονική του προηγμένου pipeline Μηχανικής Μάθησης, η υλοποίηση και η αξιολόγηση του οποίου αναλύονται διεξοδικά στο κεφάλαιο 5.

Κεφάλαιο 5ο: Μεθοδολογία πρόβλεψης ζήτησης

5.1 Επισκόπηση

Όπως αναφέρθηκε και στην εισαγωγή, η παρούσα διπλωματική εργασία πραγματεύεται με την ανάπτυξη ενός έξυπνου συστήματος με χρήση Τεχνητής Νοημοσύνης, το οποίο θα προσφέρει βελτιστοποίηση αποθεμάτων σε είδη με διαλείπουσα και αραιή ζήτηση. Ο τομέας που εστιάζουμε είναι αυτός των ανταλλακτικών, και πιο συγκεκριμένα, των ανταλλακτικών αυτοκινήτων.

Το σύστημα χωρίζεται σε δύο διακριτά τμήματα. Το πρώτο, το οποίο είναι και το εκτενέστερο και δυσκολότερο κομμάτι, είναι η πρόβλεψη της μελλοντικής ζήτησης ανά προϊόν σε ορισμένους χρονικούς ορίζοντες. Έπειτα, χρησιμοποιώντας τα αποτελέσματα αυτού μεταβαίνουμε στη δεύτερη, και τελική, φάση, κατά την οποία παράγονται προτάσεις παραγγελιών αναπλήρωσης αποθεμάτων. Οι προτάσεις αυτές είναι ανά είδος και ανά προμηθευτή για ορισμένο χρονικό ορίζοντα.

Ο τελικός στόχος του συστήματος είναι η διαρκής διαθεσιμότητα προϊόντων χωρίς ελλείψεις και, παράλληλα, η μείωση του υπερβολικού αποθέματος.

Στο τρέχον κεφάλαιο παρουσιάζεται αναλυτικά τα πρώτο τμήμα του προτεινόμενου συστήματος, δηλαδή η πρόβλεψη ζήτησης ανταλλακτικών αυτοκινήτων. Επισημαίνεται ότι δεν αναπτύχθηκε μία και μόνο μεμονωμένη λύση στο εν λόγω πρόβλημα, αλλά μια σειρά τριών διαδοχικών προσεγγίσεων που αναπτύχθηκαν, αξιολογήθηκαν και, τελικά, οδήγησαν στην τελική λύση. Κάθε προσέγγιση σχεδιάστηκε, για να απαντήσει σε συγκεκριμένα ερωτήματα ή να άρει τους περιορισμούς της προηγούμενης. Η τεκμηριωμένη παρουσίαση και των τριών είναι ουσιαστική για την κατανόηση της τελικής αρχιτεκτονικής, διότι οι σχεδιαστικές επιλογές της τελικής έκδοσης δεν είναι αυθαίρετες, αλλά προέκυψαν από συγκεκριμένες αποτυχίες ή περιορισμούς των προηγούμενων εκδόσεων σε συνδυασμό με υφιστάμενες προτάσεις της ερευνητικής βιβλιογραφίας, όπως αυτές αναφέρθηκαν στο κεφάλαιο 3.

Η δομή του τρέχοντος κεφαλαίου είναι η κάτωθι:

- Στην ενότητα 5.2 ορίζονται το τεχνικό περιβάλλον και τα εργαλεία υλοποποίησης.
- Στην ενότητα 5.3 ορίζεται το κοινό πλαίσιο αξιολόγησης που εφαρμόζεται και στις τρεις προσεγγίσεις, όπως το σημείο αναφοράς πρόβλεψης (*forecast origin*), οι επιλεγμένοι χρονικοί ορίζοντες, η διασφάλιση της ακεραιότητας των δεδομένων και οι μετρικές απόδοσης.
- Στην ενότητα 5.4 περιγράφεται η 1η προσέγγιση, η οποία χρησιμοποιεί Νευρωνικά Δίκτυα Γράφων (*Graph Neural Networks*) για την αποτύπωση συσχετίσεων μεταξύ των προϊόντων και πραγματοποιεί κυλιόμενες προβλέψεις 1, 15 και 90 ημερών.
- Στην ενότητα 5.5 περιγράφεται η 2η προσέγγιση, η οποία χρησιμοποιεί μοντέλα προβλέψεων « Εμποδίου δύο σταδίων» (t*wo-stage* Hurdle *models*) και `Sentence-Transformers` για την αποτύπωση συσχετίσεων μεταξύ των προϊόντων και πραγματοποιεί κυλιόμενες προβλέψεις μίας εβδομάδας.
- Στην ενότητα 5.6 περιγράφεται 3η προσέγγιση που είναι και η τελική. Σε αυτήν ενσωματώνονται μοντέλα προβλέψεων « Εμποδίου δύο σταδίων» (*two-stage* Hurdle *models*) και `Sentence-Transformers`. Η εκπαίδευση γίνεται σε ανεξάρτητα μεταξύ τους στιγμιότυπα του συνόλου εκπαίδευσης πλαισιωμένα από ένα μοντέλο «Μέτα-μάθησης» (*Meta-learner*). Επίσης, προϊόντα χωρίς απόθεμα λαμβάνονται υπόψη με μικρότερο συντελεστή βαρύτητας. Οι χρονικοί ορίζοντες πρόβλεψης είναι πολλαπλοί και οι προβλέψεις δεν είναι κυλιόμενες αλλά γίνονται απ’ ευθείας από το σημείο αναφοράς (*forecast origin*).

5.1.1 Η κοινή βάση του προβλήματος

Όλες οι προσεγγίσεις χρησιμοποιούν κοινό σύνολο δεδομένων. Επομένως, αντιμετωπίζουν τις ίδιες προκλήσεις που προκύπτουν από τα δομικά χαρακτηριστικά του εν λόγω συνόλου, όπως αυτά παρουσιάστηκαν αναλυτικά στο Κεφάλαιο 4. Εν συντομία:

- Αραιή (*intermittent*) ζήτηση. Το 96,5% των ειδών κατατάσσονται στην διαλείπουσα κατηγορία κατά Syntetos–Boylan (βλ. ενότητα 4.6). Όλα τα είδη παρουσιάζουν μηδενική ζήτηση κατά την πλειονότητα την ημερών.
- Υψηλή συγκέντρωση τζίρου από μικρό ποσοστό ειδών του συνολικού καταλόγου.
- Δυσανάλογος καταμερισμός αξίας αποθήκης στα διαθέσιμα προϊόντα.
- Έντονη τροποποίηση του καταλόγου προϊόντων ανά έτος.

Οι βασικές αρχές, πάνω στις οποίες αναπτύχθηκαν και οι 3 προσεγγίσεις είναι οι ακόλουθες:

- Μηδενική διαρροή δεδομένων από το σύνολο αξιολόγησης (*test set*) για χρήση κατά την εκπαίδευση.
- Επαρκής ακρίβεια πρόβλεψης στα είδη υψηλού τζίρου.
- Μείωση αξίας αποθήκης σπάνια πωλούμενων ειδών.
- Δυνατότητα αξιοποίησης αποτελεσμάτων ως είσοδο σε σύστημα προτάσεων παραγγελιών. Παράγονται αρχεία `csv`, τα οποία περιέχουν προβλέψεις ζήτησης και αποθέματα ασφαλείας ανά ορίζοντα για κάθε προϊόν ξεχωριστά, για κάθε ομάδα εναλλακτών ειδών και για κάθε κατηγορία είδους.

5.1.2 Σύντομη περιγραφή κάθε προσέγγισης

5.1.2.1 Προσέγγιση 1 – Ενοποιημένη ροή εκτέλεσης (*unified pipeline*) με Νευρωνικό Δίκτυο Γράφων (GNN)

Η πρώτη προσέγγιση επιχειρεί να εκμεταλλευτεί τις συσχετίσεις μεταξύ προϊόντων (κοινοί πελάτες, κοινή κατηγορία, κοινοί προμηθευτές, κατηγορία τιμής, συνδυασμένες αγορές, εναλλακτά είδη) μέσω ενός GNN, το οποίο αναπαριστά τις ανωτέρω συσχετίσεις σε μορφή διανυσμικών αναπαραστάσεων (embeddings). Τα embeddings αυτά συνδυάζονται με ένα πολύ πλούσιο σύνολο χαρακτηριστικών, όπως εποχικότητα, τάση αγοράς, κυλιόμενα στατιστικά και σημαντικές ημερομηνίες. Έπειτα, τα δεδομένα που προκύπτουν τροφοδοτούν πολλαπλά δενδρικά μοντέλα (`XGBoost`, `LightGBM`, `CatBoost`, `RandomForest`), τα οποία εκπαιδεύονται σε τρεις ορίζοντες πρόβλεψης, 1 ημέρας, 15 ημερών, 90 ημερών. Οι προβλέψεις είναι κυλιόμενες (*rolling predictions*) εντός του συνόλου αξιολόγησης. Η κύρια συνεισφορά της προσέγγισης είναι η καλή διαχείριση νέο-εισερχόμενων προϊόντων (*cold-starts*). Οι ημερήσιες προβλέψεις συναθροίζονται για εβδομαδιαίες, μηνιαίες, τριμηνιαίες και ετήσιες αναφορές με πολύ ικανοποιητικά αποτελέσματα.

5.1.2.2 Προσέγγιση 2 – Εβδομαδιαίες προβλέψεις μέσω μοντέλου Εμποδίου 2 σταδίων με διανυσματικές αναπαραστάσεις (*Embedding-based weekly two-stage* Hurdle *model*)

Η δεύτερη προσέγγιση μείωσε την αυξημένη πολυπλοκότητα ενός GNN χρησιμοποιώντας `Sentence-Transformers` για την παραγωγή διανυσματικών αναπαραστάσεων κειμένου (*text* embeddings) που αποτυπώνουν τις συσχετίσεις μεταξύ των προϊόντων. Στην προκειμένη περίπτωση, οι συσχετίσεις είναι πιο περιορισμένες και αφορούν μόνο περιγραφή είδους, κατηγορία και προμηθευτή. Η χρονική ανάλυση αλλάζει από ημερήσια σε εβδομαδιαία για τη μείωση της σποραδικότητας. Διατηρείται και εδώ η κυλιόμενη λογική των προβλέψεων. Η βασική μεθοδολογική καινοτομία είναι η εισαγωγή μοντέλου

προβλέψεων δύο σταδίων (*two-stage* Hurdle *models*). Το πρώτο μοντέλο είναι ένας ταξινομητής που προβλέπει την πιθανότητα να υπάρξει ζήτηση τη δεδομένη εβδομάδα. Το δεύτερο χρησιμοποιεί παλινδρόμηση και προβλέπει την ποσότητα ζήτησης, υπό τη συνθήκη ότι υπάρχει ζήτηση. Η προσέγγιση συνοδεύεται από πολύπλοκη και έντονη βαθμονόμηση (*calibration*) των αρχικών αποτελεσμάτων του στοχεύοντας σε αυξημένη ακρίβεια σε επίπεδο κατηγορίας προϊόντων.

5.1.2.3 Προσέγγιση 3 – Ροή εκτέλεσης πραγματικής πρόβλεψης διαλείπουσας ζήτησης (*Sparse true forecast pipeline*)

Η τρίτη και τελική προσέγγιση αλλάζει θεμελιωδώς τη φιλοσοφία πρόβλεψης. Αντί για κυλιόμενες (*rolling*) προβλέψεις που πραγματοποιούν προβλέψεις για κάθε νέα ημέρα/εβδομάδα, εφαρμόζεται μια στρατηγική απ' ευθείας πρόβλεψης σε πολλαπλούς χρονικούς ορίζοντες από το σημείο αναφοράς. Δηλαδή, από ένα σταθερό χρονικό σημείο, το οποίο είναι η πρώτη ημέρα του συνόλου αξιολόγησης (*forecast origin* – 01/01/2025), προβλέπεται απ' ευθείας η πιθανή μελλοντική ζήτηση για οκτώ χρονικούς ορίζοντες (15, 30, 45, 60, 75, 90, 180, 365 ημέρες) χωρίς να χρησιμοποιηθεί κανένα δεδομένο του συνόλου αξιολόγησης. Επίσης, γίνεται ανακατασκευή του αποθέματος ανά είδος και εφαρμόζεται λογοκρισία ζήτησης (*demand censoring*) στα προϊόντα χωρίς απόθεμα, ώστε να διορθωθούν οι πλασματικές μηδενικές πωλήσεις που οφείλονται σε ελλείψεις (*stock-outs*) και όχι σε πραγματική απουσία ζήτησης. Τα χαρακτηριστικά εκπαίδευσης υπολογίζονται από στατικά στιγμιότυπα προγενέστερων χρονικών στιγμών, τα οποία ορίζονται ως εξαμηνιαία σημεία αποκοπής (*semi-annual cutoffs*). Με τον τρόπο αυτό αποκλείεται πλήρως η πιθανότητα διαρροής μεταγενέστερων δεδομένων κατά την εκπαίδευση. Διατηρεί τη βασική ιδέα της αρχιτεκτονικής «*two-stage* Hurdle» από την Προσέγγιση 2, η οποία εμπλουτίζεται και εφαρμόζεται σε τέσσερα μοντέλα (RandomForest, ExtraTrees, HistGradientBoosting, CatBoost). Τα αποτελέσματα όλων συνδυάζονται σε τρεις διαφορετικές συνθέσεις (*ensembles*). Οι δύο πρώτες είναι σχετικά απλές. Στην πρώτη χρησιμοποιείται απλώς η διάμεσος όλων των αποτελεσμάτων ανά ορίζοντα, ενώ στη δεύτερη εισάγονται μεταβλητά βάρη αντιστρόφως ανάλογα του σφάλματος κάθε μοντέλου ανά ορίζοντα. Η τρίτη σύνθεση βασίζεται σε ένα μοντέλο μέτα-μάθησης (*meta-learner*). Χρησιμοποιώντας τα σφάλματα της εκπαίδευσης δημιουργεί νέες προβλέψεις σε αποκλειστικώς ιστορικά στιγμιότυπα και βάσει αυτών παράγει νέες προβλέψεις ανά μοντέλο και ανά ορίζοντα για κάθε στιγμιότυπο. Τέλος, συνδυάζονται οι νέες προβλέψεις σε ένα μοντέλο παλινδρόμησης και παράγονται τα τελικά βάρη της σύνθεσης.

5.1.3 Η εξέλιξη του συστήματος μέσα από τις 3 προσεγγίσεις

Η ανάγκη για πολλαπλές προσεγγίσεις δεν ήταν, προφανώς, αποφασισμένη εκ των προτέρων. Κατά συνέπεια, και η μετάβαση από τη μια προσέγγιση στην επόμενη δεν ήταν προκαθορισμένη αλλά προέκυψε λόγω ορισμένων μεθοδολογικών και σχεδιαστικών περιορισμών κατά την εξέλιξη κάθε μίας.

5.1.3.1 Από την Προσέγγιση 1 στην Προσέγγιση 2: Η πολυπλοκότητα του GNN και ο περιορισμός των κυλιόμενων ημερήσιων προβλέψεων

Η Προσέγγιση 1 παράγει κυλιόμενες ημερήσιες προβλέψεις, ώστε για να υπολογιστεί η εβδομαδιαία, μηνιαία κοκ. ζήτηση. Οι ημερήσιες προβλέψεις συναθροίζονται εκ των υστέρων, ώστε να προκύψουν τα τελικά αποτελέσματα. Αν και αυτό είναι άρτιο από τεχνικής και μαθηματικής άποψης και χωρίς διαρροή δεδομένων, όταν υλοποιείται σωστά, στην πράξη σημαίνει ότι κάθε νέα ημέρα απαιτεί εκ νέου υπολογισμό. Αυτό έχει ως αποτέλεσμα τη διαρκή ενημέρωση του συνόλου των χαρακτηριστικών. Αυτή η λογική είναι κατάλληλη για ακαδημαϊκή έρευνα και διαγωνισμούς, καθώς επίσης, και για διαχείριση συχνά πωλούμενων ειδών με διαρκή ενημέρωση.

Δεν είναι, όμως, κατάλληλη για ένα σύστημα παραγωγής που προτείνει παραγγελίες μεσαίου/μεγάλου ορίζοντα, π.χ. ετήσιες συμφωνίες με προμηθευτές, εποχικές αγορές. Σε αυτό το πλαίσιο, ο υπεύθυνος αγορών χρειάζεται να γνωρίζει σήμερα πόσες μονάδες αναμένεται να πουληθούν τους επόμενους 3 ή 6 μήνες, χωρίς να εξαρτάται από έναν μηχανισμό που απαιτεί καθημερινή ανανέωση των προβλέψεων. Επίσης, η διαχείριση ενός μεγάλου GNN είναι πολύ απαιτητική σε πόρους, όταν ενσωματωθεί σε ένα μοντέλο παραγωγής.

Η Προσέγγιση 2 επιχειρεί να μετριάσει αυτούς τους περιορισμούς με τους ακόλουθους τρόπους:

- Η ανάλυση μετακινείται από ημερήσια σε εβδομαδιαία βάση και, παράλληλα, βελτιώνεται η διαχείριση της σποραδικότητας μέσω της *two-stage* Hurdle αρχιτεκτονικής.
- Απλοποιείται το δίκτυο αναπαράστασης συσχετίσεων με την ενσωμάτωση Sentence Trasformers για τον υπολογισμό των embeddings.

5.1.3.2 Από την Προσέγγιση 2 στην Προσέγγιση 3: Ο κίνδυνος της υπερβολικής βαθμονόμησης

Η Προσέγγιση 2 πετυχαίνει εντυπωσιακά αριθμητικά αποτελέσματα, κυρίως, χάριν στην πολλαπλών επιπέδων βαθμονόμηση που πραγματοποιείται και στοχεύει σε ακρίβεια >92% σε επίπεδο κατηγορίας. Ωστόσο, μια ενδελεχής εξέταση κατά την ολοκλήρωση της ανάπτυξής της αποκάλυψε δύο σοβαρά προβλήματα:

- Η κυλιόμενη εβδομαδιαία πρόβλεψη παραμένει θεμελιωδώς ασύμβατη με τη λογική των απ' ευθείας προβλέψεων σε πολλαπλούς χρονικούς ορίζοντες. Για παράδειγμα, δεν είναι δυνατόν, να αξιολογηθεί η πρόβλεψη 180 ημερών χωρίς κυλιόμενη ανανέωση του συνόλου χαρακτηριστικών, αφού το μοντέλο εκπαιδεύεται σε εβδομαδιαίους στόχους.
- Η υπερβολική βαθμονόμηση, αν και εντυπωσιακή στα αποτελέσματα των μετρικών απόδοσης, εισάγει σημαντική υπερ-προσαρμογή (*over-fitting*) και τροποποιήσεις που βασίζονται στα αποτελέσματα. Ως εκ τούτου, η ικανότητα γενίκευσης του τελικού μοντέλου είναι απογοητευτική. Με άλλα λόγια, στο κυνήγι της βελτιστοποίησης των αποτελεσμάτων έπαψαν να είναι αξιόπιστα.

Οπότε, η Προσέγγιση 3 σχεδιάστηκε «***ex ante***» με σκοπό να άρει αυτούς τους δύο περιορισμούς. Με άλλα λόγια, στην τελική προσέγγιση ακολουθείται σαφής σχεδιαστική στρατηγική, ώστε να είναι στιβαρή (*robust by design*) και να εξασφαλίζονται οι ακόλουθες προϋποθέσεις:

- Απ' ευθείας πρόβλεψη σε πολλαπλούς χρονικούς ορίζοντες από το σημείο αναφοράς.
- Πειθαρχία στη βαθμονόμηση, ώστε να εξασφαλίζεται ικανοποιητική ικανότητα γενίκευσης.

5.2 Τεχνικό Περιβάλλον και Εργαλεία Υλοποίησης

Η αρχιτεκτονική του συστήματος πρόβλεψης ζήτησης υλοποιήθηκε εξ' ολοκλήρου με τη χρήση σύγχρονων τεχνολογιών ανοικτού λογισμικού εστιάζοντας στην υπολογιστική αποδοτικότητα, τη σπονδυλωτή (*modular*) δομή και τη δυνατότητα παραγωγικής κλιμάκωσης (*production readiness*). Δεδομένου του μεγάλου όγκου του συνόλου δεδομένων, το οποίο περιλαμβάνει 1.617.633 εγγραφές συναλλαγών για 79.363 μοναδικά SKU, η επιλογή των εργαλείων έγινε με γνώμονα τη βέλτιστη διαχείριση μνήμης και την ταχύτητα εκτέλεσης όλων των απαιτούμενων διεργασιών.

5.2.1 Γλώσσα Προγραμματισμού και Βασικό Οικοσύστημα

Ως κεντρική γλώσσα προγραμματισμού για την ανάπτυξη και την εκτέλεση του pipeline επιλέχθηκε η `Python` (έκδοση `3.10+`) [146]. Η επιλογή της υπαγορεύτηκε από την καθολική της κυριαρχία στα πεδία

της Μηχανικής Μάθησης και της Επιστήμης Δεδομένων, καθώς, και από τη διαθεσιμότητα εξειδικευμένων βιβλιοθηκών υψηλής απόδοσης.

Για τη διασφάλιση της ποιότητας του πηγαίου κώδικα (όπως αυτός αναπτύχθηκε στα αρχεία `unified_forecasting_v6.ipynb`, `08_two_stage_forecasting.ipynb` και `sparse_true_forecast_pipeline.py`) και της μακροπρόθεσμης δυνατότητας συντήρησής του, αξιοποιήθηκαν προηγμένες δυνατότητες στατικής τυποποίησης μέσω του ενσωματωμένου module `typing` και της δήλωσης `from __future__ import annotations`. Επιπλέον, χρησιμοποιήθηκαν αντικειμενοστραφείς δομές δεδομένων (`dataclasses`) για την καθαρή και ασφαλή αναπαράσταση των παραμέτρων του pipeline, των αποτελεσμάτων αξιολόγησης και των ορισμάτων εισόδου.

5.2.2 Διαχείριση και Μετασχηματισμός Δεδομένων Μεγάλης Κλίμακας

Η επεξεργασία του ιστορικού συναλλαγών και η εξαγωγή των χαρακτηριστικών (*feature engineering*) βασίστηκαν σε δύο θεμελιώδεις βιβλιοθήκες του οικοσυστήματος της `Python`:

- `NumPy`: Χρησιμοποιήθηκε για τη διαχείριση πολυδιάστατων πινάκων, την εκτέλεση γρήγορων διανυσματικών πράξεων (*vectorized operations*) και τον υπολογισμό των μαθηματικών συναρτήσεων σφάλματος [147].
- `Pandas`: Αποτέλεσε τη βάση για τον χειρισμό των δομημένων δεδομένων μέσω των αντικειμένων `DataFrame` και `Series` [148]. Μέσω του `Pandas` πραγματοποιήθηκε η διαλογή, ο καθαρισμός, ο χρονικός συγχρονισμός και η ομαδοποίηση των πωλήσεων. Ιδιαίτερα κρίσιμη ήταν η συμβολή της βιβλιοθήκης στην υλοποίηση του αλγορίθμου Λογοκρισίας Ζήτησης (*Demand Censoring*) για τη διόρθωση των ελλείψεων αποθεμάτων, καθώς, και στη διαστρωματωμένη κλιμάκωση (*Stratified Scaling*) για την προστασία των προϊόντων υψηλού τζίρου (Top-20 και Pareto-A) από πιθανή συστηματική υπο-πρόβλεψη.

5.2.3 Αρχιτεκτονικές Μηχανικής Μάθησης και Πλαίσια Μοντέλων Συνόλου (Ensemble Frameworks)

Για την υλοποίηση των μοντέλων Εμποδίου δύο σταδίων (*two-stage* Hurdle *models*) και της τελικής αρχιτεκτονικής συσσωρευτικής γενίκευσης (*stacked generalization*) [58], επιστρατεύτηκε ένας συνδυασμός κορυφαίων frameworks δενδρικών δομών:

- `Scikit-Learn (sklearn)`: Παρείχε τις βασικές δομές για τον διαχωρισμό εκπαίδευσης/αξιολόγησης χωρίς διαρροή πληροφορίας (*no data leakage*), τις μεθοδολογίες των βασικών μοντέλων αναφοράς (*baselines*), καθώς, και ισχυρές ensemble υλοποιήσεις `HistGradientBoostingClassifier`, `HistGradientBoostingRegressor`, `RandomForestClassifier` και `ExtraTreesClassifier` [57].
- `LightGBM (lgbm)`: Επιλέχθηκε για την εξαιρετική του ταχύτητα και τις χαμηλές απαιτήσεις μνήμης, χάρη στην τεχνική εκπαίδευσης βάσει φύλλων (*leaf-wise growth*) [55]. Αποδείχθηκε ιδιαίτερα αποδοτικό στην πιθανοκρατική πρόβλεψη μέσω των `Quantile losses` (ειδικότερα για τις άλφα τιμές `Q0.1` και `Q0.9`).
- `CatBoost`: Ενσωματώθηκε για τη μοναδική του ικανότητα να διαχειρίζεται εγγενώς κατηγορικά χαρακτηριστικά (όπως η κατηγορία προϊόντος ή ο προμηθευτής) προσφέροντας υψηλή στιβαρότητα (*robustness*) έναντι της υπερ-προσαρμογής (*overfitting*) [56].
- `XGBoost`: Χρησιμοποιήθηκε ως βασικός πυλώνας για τη μοντελοποίηση των μη γραμμικών σχέσεων [54] αξιοποιώντας εξειδικευμένες παραμετροποιήσεις, όπως η συνάρτηση απώλειας

`Tweedie loss` και τα ασύμμετρα βάρη δειγμάτων (`spike-aware asymmetric sample weights`), για τη σωστή διαχείριση της μεγάλης αραιότητας των δεδομένων.

5.2.4 Εξειδικευμένα Πλαίσια Βαθιάς Μάθησης και Εξαγωγής Γνώσης

Στο πλαίσιο της παρούσας έρευνας και κατά τη διάρκεια ανάπτυξης και αξιολόγησης πειραματικών αρχιτεκτονικών (πριν την οριστικοποίηση του τελικού μοντέλου) αξιοποιήθηκαν, επιπλέον, εξειδικευμένα εργαλεία για την αντιμετώπιση στοχευμένων προβλημάτων:

- `PyTorch & PyTorch Geometric:` Χρησιμοποιήθηκαν για την ανάπτυξη πειραματικών μοντέλων Μηχανικής Μάθησης σε Γράφους (Multi-Task Graph Neural Networks), στοχεύοντας στην παραγωγή structural node embeddings από τις σχέσεις *co-purchase* και ομοκατηγορίας των SKU [149], **Error! Reference source not found.**.
- `Sentence-Transformers:` Το συγκεκριμένο πλαίσιο εργασίας (framework) ενσωματώθηκε οργανικά στον κώδικα για την εξαγωγή σημασιολογικών αναπαραστάσεων (*text* embeddings) από τις λεκτικές περιγραφές του καταλόγου ειδών στοχεύοντας στην επίλυση του προβλήματος ψυχρής εκκίνησης (*cold start*) σε νεοεισαχθέντα προϊόντα. Ενώ σε πρώιμα πειραματικά στάδια διερευνήθηκε η χρήση του πολυπλοκότερου μοντέλου `BGE-M3`, στην τελική αρχιτεκτονική («`Sparse True Forecast Pipeline`») υιοθετήθηκε το μοντέλο `MiniLM` [151]. Η επιλογή αυτή εξασφάλισε τη βέλτιστη ισορροπία μεταξύ της ποιότητας των διανυσματικών αναπαραστάσεων και της ταχύτητας εξαγωγής (*inference speed*) διατηρώντας, ταυτόχρονα, εξαιρετικά χαμηλές απαιτήσεις σε μνήμη RAM, γεγονός απαραίτητο για την παραγωγική λειτουργία του συστήματος.

5.2.5 Υπολογιστική Επιτάχυνση και Βελτιστοποίηση Υπερπαραμέτρων

- `Optuna:` Η αυτοματοποιημένη αναζήτηση υπερπαραμέτρων πραγματοποιήθηκε μέσω του `Optuna`, το οποίο εφαρμόζει αλγορίθμους Bayesian βελτιστοποίησης [113]. Επιτρέπει την αποδοτική εξερεύνηση του χώρου των παραμέτρων (π.χ. ρυθμός μάθησης, βάθος δένδρων, συντελεστές `L1/L2 regularisation`) μειώνοντας δραστικά τον απαιτούμενο χρόνο σε σχέση με το κλασική αναζήτηση πλέγματος (*grid search*).
- `Numba:` Λόγω των σύνθετων προσαρμοσμένων (*custom*) μετρικών που απαιτεί η επιχειρησιακή αξιολόγηση (όπως η `NDCG@20` για την ιεράρχηση των Top-20 SKU και οι εξειδικευμένοι δείκτες `WAPE` ανά βαθμίδα τζίρου), χρησιμοποιήθηκε ο διακοσμητής (*decorator*) `@jit` της `Numba` [152]. Η `Just-In-Time (JIT)` μεταγλώττιση σε καθαρό κώδικα μηχανής παρέκαμψε τους περιορισμούς ταχύτητας της Python επιτρέποντας την εκτέλεση εκατομμυρίων επαναληπτικών υπολογισμών σε ελάχιστα δευτερόλεπτα.

5.2.6 Διαχείριση Κατάστασης (State Management) και Αναπαραγωγιμότητα (Reproducibility)

Για τη διασφάλιση της επιστημονικής αναπαραγωγιμότητας και της δυνατότητας τμηματικής εκτέλεσης του pipeline, εφαρμόστηκε ένας αυστηρός μηχανισμός αποθήκευσης κατάστασης:

- `Pickle & Dill:` Τα εκπαιδευμένα μοντέλα, οι μάσκες δεδομένων (`train_mask`, `val_mask`), τα τελικά embeddings και τα αποτελέσματα του αναδρομικού ελέγχου (*backtest predictions*) αποθηκεύονται σε σειριοποιημένη μορφή σε αρχεία της μορφής `xxxxxxx.pkl` επιτρέποντας την άμεση φόρτωση δεδομένων κατά την εκτέλεση του pipeline χωρίς την ανάγκη επανεκπαίδευσης από μηδενική βάση.

- `Hashlib` & `JSON`**:** Χρησιμοποιήθηκαν για τη δημιουργία μοναδικών αναγνωριστικών (hashes) των παραμέτρων και των δεδομένων εισόδου διασφαλίζοντας την ακεραιότητα της διαδικασίας και αποτρέποντας την ακούσια αλλοίωση των αποτελεσμάτων.

5.3 Πλαίσιο Αξιολόγησης

Στην τρέχουσα ενότητα ορίζεται ένα ενιαίο πλαίσιο αξιολόγησης όλων των προσεγγίσεων, ώστε να διασφαλίζεται η αξιόπιστη σύγκριση των αποτελεσμάτων. Τα σημεία αναφοράς είναι τα κάτωθι:

- Το χρονικό σημείο αναφοράς της πρόβλεψης (*forecast origin*).
- Η διαίρεση του συνόλου δεδομένων σε σύνολα εκπαίδευσης και ελέγχου (*train/test split*).
- Οι χρονικοί ορίζοντες πρόβλεψης (*forecast horizons*).
- Οι μετρικές απόδοσης και οι μαθηματικοί τους ορισμοί.
- Η αξιολόγηση των προϊόντων ανά βαθμίδα τζίρου (Top-20 / *Pareto-A* / Long-tail).
- Οι αρχές ακεραιότητας δεδομένων (*no data leakage*).

Όταν εισάγονται επιπλέον ή διαφοροποιημένες μετρικές σε κάποια προσέγγιση, όπως η σύμμορφη κάλυψη ζήτησης (*conformal demand coverage*) στην Προσέγγιση 1 ή το ποσοστό κάλυψης ζήτησης ανά κατηγορία (*category fill rate*) στην Προσέγγιση 2, αυτές παρουσιάζονται αυτοτελώς μέσα στην αντίστοιχη ενότητα.

5.3.1 Σημείο αναφοράς της πρόβλεψης (*forecast origin*) και διαίρεση συνόλων εκπαίδευσης/αξιολόγησης (*train/test set*)

Ως σημείο αναφοράς της πρόβλεψης ορίζεται η 01/01/2025. Αυτό σημαίνει:

- Όλα τα δεδομένα έως 31/12/2024 είναι διαθέσιμα για εκπαίδευση, εξαγωγή χαρακτηριστικών, βαθμονόμηση και επιλογή υπερπαραμέτρων.
- Όλα τα δεδομένα από 01/01/2025 έως 31/12/2025 χρησιμοποιούνται αποκλειστικά για αξιολόγηση. **Ποτέ** για εκπαίδευση ή ρύθμιση παραμέτρων.

Η επιλογή του σταθερού σημείου αναφοράς της πρόβλεψης έχει δύο πλεονεκτήματα:

- Αντιπροσωπευτικότητα, γιατί το 2025 καλύπτει έναν πλήρη ετήσιο κύκλο, περιλαμβανομένων της θερινής αιχμής, της αυγουστιάτικης ύφεσης και της φθινοπωρινής κορύφωσης (βλ. ενότητα 4.2).
- Συγκρισιμότητα, γιατί χρησιμοποιείται κοινό διάστημα αξιολόγησης και στις τρεις προσεγγίσεις, ώστε οι μετρικές να είναι άμεσα συγκρίσιμες.

Όσον αφορά τον διαχωρισμό του συνόλου δεδομένων (*dataset*) σε σύνολα εκπαίδευσης (*train*) και αξιολόγησης (*test*), παρουσιάζεται στον Πίνακα 5.1:

Πίνακας 5.1: Διαχωρισμός συνόλων εκπαίδευσης/αξιολόγησης

Σύνολο	**Περίοδος**	**Εγγραφές πωλήσεων (`type=1`)**	**Είδη με τουλάχιστον μία πώληση**
Εκπαίδευσης	01/01/2021 – 31/12/2024	~930.000	~62.000
Αξιολόγησης / Ελέγχου	01/01/2025 – 31/12/2025	~232.000	~44.000

Επισημαίνεται ότι περίπου 42,7% των αντικειμένων στο σύνολο εκπαίδευσης δεν εμφανίζονται στο σύνολο αξιολόγησης (βλ. ενότητα 4.7).

5.3.2 Χρονικοί ορίζοντες πρόβλεψης

Ο κάθε ορίζοντας h ορίζεται ως **πλήθος ημερών προς το μέλλον από το *forecast origin***. Για ένα δεδομένο κωδικό είδους (SKU), η πραγματική ζήτηση του ορίζοντα h είναι το άθροισμα:

$$Actual_h = \sum_{d=origin+1}^{origin+h} q\,ty_{SKU,d} \tag{5.1}$$

Κάθε μία από τις σχεδιαστικές προσεγγίσεις που ακολουθούνται υιοθετεί διαφορετικό σύνολο χρονικών οριζόντων.

Πίνακας 5.2: Ορίζοντες πρόβλεψης ανά σχεδιαστική προσέγγιση

Προσέγγιση	**Ορίζοντες**	**Φύση πρόβλεψης**
1 (GNN)	1d, 15d, 90d	Κυλιόμενες απ' ευθείας 1, 15 και 90 ημερών. Επίσης αθροιστικές προβλέψεις 1 ημέρας σε σύνολο εβδομάδας, μήνα, τριμήνου και έτους
2 (*Weekly two-stage*)	Εβδομαδιαίος	Κυλιόμενη εβδομαδιαία
3 (*True forecast pipeline*)	15, 30, 45, 60, 75, 90, 180, 365	Απ' ευθείας προβλέψεις πολλαπλών οριζόντων αποκλειστικά από το σημείο αναφοράς

Η Προσέγγιση 3, που αποτελεί τον άξονα αναφοράς της εργασίας, παράγει **οκτώ ορίζοντες** ταυτόχρονα, καθένας με το δικό του εξειδικευμένο μοντέλο. Έτσι, καλύπτονται όλες οι πιθανές πρακτικές πολιτικής αναπλήρωσης: άμεση (15d), κανονική (30–90d), εποχική (180d) και ετήσιες συμφωνίες προμηθευτών (365d).

5.3.3 Μετρικές απόδοσης

Για κάθε συνδυασμό μοντέλου πρόβλεψης – χρονικού ορίζοντα υπολογίζεται ένα σύνολο μετρικών απόδοσης. Η κύρια λογική πίσω από τη χρήση πλήθους μετρικών είναι ότι καμία μεμονωμένη μετρική δεν είναι επαρκής για πλήρη κάλυψη προβλημάτων διαλείπουσας ζήτησης. Για παράδειγμα, το «MAPE» (Mean Absolute Percentage Error – Μέσο Απόλυτο Ποσοστιαίο Σφάλμα) δεν ορίζεται, όταν η πραγματική ζήτηση είναι μηδενική, το «MAE» (Mean Absolute Error – Μέσο Απόλυτο Σφάλμα) αγνοεί τη σχετική κλίμακα και το «RMSE» (Root Mean Squared Error – Ρίζα Μέσου Τετραγωνικού Σφάλματος) δίνει υπέρμετρο βάρος σε ακραίες τιμές (*outliers*) [20]. Ως εκ τούτου, χρησιμοποιείται ένα σύνολο μετρικών, ώστε να καλύπτονται όλες οι πιθανές ανάγκες.

Συμβολίζουμε με y_i την πραγματική τιμή του προϊόντος i στον ορίζοντα h, και με $\hat{y}_i$ την αντίστοιχη πρόβλεψη του μοντέλου. Ακολουθούν οι ορισμοί των μετρικών απόδοσης που χρησιμοποιούνται στην παρούσα Δ.Ε.

5.3.3.1 Σταθμισμένο Απόλυτο Ποσοστιαίο Σφάλμα (Weighted Absolute Percentage Error – WAPE)

$$WAPE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{\sum_{i=1}^{N}|y_i| + \varepsilon} \quad (5.2)$$

Η WAPE αποτελεί το σταθμισμένο απόλυτο ποσοστιαίο σφάλμα και είναι η κύρια μετρική ακρίβειας της εργασίας. Σε αντίθεση με τη MAPE, η WAPE ορίζεται ακόμη και όταν μεμονωμένα $y_i = 0$ και αποδίδει μεγαλύτερο βάρος σε SKUs με υψηλότερη πραγματική ζήτηση [14]. Επίσης, είναι αδιάστατη μετρική και ισχύει ο κανόνας της ελαχιστοποίησης, δηλαδή όσο μικρότερη τιμή, τόσο καλύτερη η απόδοση.

Η σταθερά $\varepsilon = 10^{-9}$ εξασφαλίζει αριθμητική ευστάθεια, όταν το σύνολο των προϊόντων έχει μηδενική ζήτηση, ώστε να μην μηδενίζεται ο παρονομαστής.

5.3.3.2 Λόγος Μεροληψίας Πρόβλεψης — Ratio (*Forecast Bias*)

$$Ratio = \frac{\sum_{i=1}^{N}\hat{y}_i}{\sum_{i=1}^{N}y_i + \varepsilon} \quad (5.3)$$

Το Ratio μετρά τη **συστηματική μεροληψία** του μοντέλου:

- Ratio = 1,00: τέλεια ισορροπία σε συνολικό άθροισμα
- Ratio < 1,00: υπο-πρόβλεψη (under-forecast)
- Ratio > 1,00: υπερ-πρόβλεψη (over-forecast)

Σε σύστημα διαχείρισης αποθεμάτων, η υπο-πρόβλεψη οδηγεί σε πιθανές ελλείψεις προϊόντων και απώλεια πωλήσεων, ενώ η υπερ-πρόβλεψη οδηγεί σε αυξημένο κόστος αποθήκευσης και νεκρό απόθεμα. Επιχειρησιακός στόχος του συστήματος είναι ο λόγος μεροληψίας (*ratio*) να κυμαίνεται στο διάστημα [0,95, 1,05] σε καθολικό επίπεδο, και στο διάστημα [0,85, 1,15] ανά βαθμίδα τζίρου (όπως για παράδειγμα στα είδη Top-20).

Επισημαίνεται ότι το ratio δείχνει αποκλειστικά αθροιστική μεροληψία ενός συνόλου προϊόντων. Είναι πιθανό το ratio να βρίσκεται πολύ κοντά στο 1 αλλά το επιμέρους σφάλμα κάθε προϊόντος να είναι υψηλό και να αλληλοεξουδετερώνονται.

5.3.3.3 Συνδυαστική Μετρική Ακρίβειας – Μεροληψίας (Combined Accuracy–Bias Score)

$$Score = WAPE \cdot (1 + max\,(0,\, 1 - Ratio)) \quad (5.4)$$

Το **Score** συνδυάζει την ακρίβεια (μέσω της μετρικής WAPE) με την επιβολή ποινής στη συστηματική υπο-πρόβλεψη. Συγκεκριμένα, η υπερ-πρόβλεψη δεν τιμωρείται διπλά (καθώς η απόκλιση αντανακλάται ήδη στον υπολογισμό της WAPE), ενώ αντίθετα, η υπο-πρόβλεψη επιδεινώνει πολλαπλασιαστικά την τελική βαθμολογία.

Πρόκειται για μια ειδικά κατασκευασμένη μετρική, η οποία αναπτύχθηκε για να εξυπηρετήσει τον θεμελιώδη επιχειρησιακό κανόνα: *«stockouts are costlier than overstock for critical SKUs»* [153]. Δηλαδή, αναγνωρίζει μαθηματικά ότι οι ελλείψεις κρίσιμων ειδών έχουν μεγαλύτερο κόστος ευκαιρίας από τη διατήρηση υπερβολικού αποθέματος. Όπως και στη WAPE, ισχύει ο κανόνας της ελαχιστοποίησης (όσο μικρότερη η τιμή του Score, τόσο καλύτερη η απόδοση του μοντέλου).

5.3.3.4 Ποσοστό Θετικής Αποδοχής (HitRate30)

$$HitRate30 \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\frac{|\,y_i - \hat{y}_i\,|}{max\,(y_i, 1)} \leq 0{,}3\right] \tag{5.5}$$

Το HitRate30 εκφράζει το ποσοστό των SKUs, για τα οποία η παραγόμενη πρόβλεψη βρίσκεται εντός αποδεκτού εύρους απόκλισης, οριζόμενου στο ±30% της πραγματικής ζήτησης. Στον παρονομαστή του μαθηματικού τύπου εισάγεται η συνθήκη $\max(y_i, 1)$ για την αποφυγή απειρισμού σε περιπτώσεις μηδενικής πραγματικής ζήτησης.

Η εν λόγω μετρική αποτυπώνει με διαισθητικό τρόπο τον αριθμό των κωδικών που προβλέφθηκαν ικανοποιητικά καθιστώντας, έτσι, την ιδιαίτερα χρήσιμη για επιχειρηματικές αναφορές και αξιολόγηση από μη τεχνικούς χρήστες [154].

5.3.3.5 Αξιολόγηση ταξινόμησης (F1-Score)

Για αρχιτεκτονικές που διαχειρίζονται τη διαλείπουσα ζήτηση μέσω μοντέλων δύο σταδίων (Hurdle *models*), απαιτείται η ρητή δυαδική αξιολόγηση των περιόδων σε «ενεργές» (θετική ζήτηση) και «ανενεργές» (μηδενική ζήτηση). Για τους δυαδικούς ταξινομητές, χρησιμοποιείται το F1-Score, ο αρμονικός μέσος όρος Ακρίβειας (*Precision*) και Ανάκλησης (*Recall*) που ορίζεται ως εξής:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.6}$$

Όπου

- **Precision (Ακρίβεια):** Το ποσοστό των ορθών θετικών προβλέψεων επί του συνόλου των θετικών προβλέψεων ($\frac{TP}{TP+FP}$). Απαντά στο ερώτημα: «Όταν το μοντέλο προέβλεψε ζήτηση, πόσες φορές είχε δίκιο;»
- **Recall (Ανάκληση):** Το ποσοστό των ορθών θετικών προβλέψεων επί του συνόλου των πραγματικών θετικών περιπτώσεων ($\frac{TP}{TP+FN}$). Απαντά στο ερώτημα: «Από όλες τις μέρες που όντως υπήρξε ζήτηση, πόσες κατάφερε να εντοπίσει το μοντέλο;»

Λόγω της ακραίας ανισορροπίας κλάσεων (*class imbalance*, με τα μηδενικά δείγματα να προσεγγίζουν το 95%), η κλασική μετρική της Ακρίβειας (Accuracy) θα απέδιδε παραπλανητικά υψηλά ποσοστά. Το F1-Score διασφαλίζει ότι το σύστημα αξιολογείται αυστηρά ως προς την ικανότητά του να εντοπίζει τα επιχειρησιακά κρίσιμα γεγονότα εκδήλωσης ζήτησης.

5.3.3.6 Μέσο Απόλυτο Κλιμακωτό Σφάλμα (Mean Absolute Scaled Error – MASE)

$$MASE \;=\; \frac{\sum_{i=1}^{N} |\,y_i - \hat{y}_i\,|}{\sum_{i=1}^{N} |\,y_i - \hat{y}_i^{naive}\,|} \tag{5.7}$$

όπου $\hat{y}_i^{naive}$ είναι η αφελής (naive) πρόβλεψη.

Στο παρόν σύστημα, ως αφελής πρόβλεψη χρησιμοποιείται η ιστορική ζήτηση του ίδιου SKU κατά το αντίστοιχο προηγούμενο χρονικό διάστημα (π.χ., για ορίζοντα $h = 30$, υπολογίζεται η μέση ημερήσια ζήτηση των τελευταίων 30 ημερών πριν το σημείο αποκοπής και πολλαπλασιάζεται με το h).

Η μετρική MASE εισήχθη από τους Hyndman και Koehler [20] ως μια στιβαρή εναλλακτική σε περιπτώσεις διαλείπουσας ζήτησης, κατά τις οποίες αποτυγχάνει το MAPE (λόγω μηδενικών παρονομαστών). Η ερμηνεία της ακολουθεί τους εξής κανόνες:

- MASE < 1**:** Το μοντέλο υπερτερεί της αφελούς πρόβλεψης.
- MASE ≈ 1: Το μοντέλο είναι ισοδύναμο της αφελούς πρόβλεψης.
- MASE > 1: Η αφελής πρόβλεψη είναι καλύτερη (το μοντέλο εισάγει θόρυβο).

Αιτιολόγηση Απόρριψης του RMSSE

Σημειώνεται ότι δεν χρησιμοποιείται το RMSSE (Root Mean Squared Scaled Error – Ρίζα Μέσου Τετραγωνικού Κλιμακωτού Σφάλματος), παρότι αποτελεί την επίσημη μετρική αξιολόγησης του διαγωνισμού M5 [21]. Η απόφαση αυτή στηρίζεται σε δύο θεμελιώδεις άξονες:

- **Ευαισθησία σε ακραίες τιμές (*Outliers*).** Στα ανταλλακτικά αυτοκινήτων καταγράφονται συχνά ακραίες τιμές. Είναι αρκετά πιθανό ένα προϊόν με τυπική ζήτηση 1–2 τεμαχίων τον μήνα να δεχτεί αιφνιδίως μια παραγγελία πολλών τεμαχίων (π.χ., από πελάτη με μεγάλο στόλο οχημάτων). Αυτό το γεγονός εισάγει μη προβλέψιμο θόρυβο. Η μετρική MASE τιμωρεί το σφάλμα γραμμικά (ένα λάθος 48 τεμαχίων κοστίζει 48 μονάδες), ενώ το RMSSE το τιμωρεί στο τετράγωνο ($48^2 = 2.304$ μονάδες). Ως εκ τούτου, ένα μοντέλο που βελτιστοποιείται βάσει του RMSSE τείνει να προσαρμόζεται υπερβολικά σε αυτές τις ακραίες τιμές με αποτέλεσμα να καταλήγει σε υπερ-προσαρμογή (*overfit*) πάνω σε τυχαίο θόρυβο [14].
- **Επιχειρηματική Ερμηνεία.** Στη διαχείριση αποθεμάτων, τα κόστη αυξάνονται γραμμικά σε σχέση με την απόκλιση της πρόβλεψης (π.χ., 10 τεμάχια νεκρού αποθέματος ισοδυναμούν με 10 φορές το μοναδιαίο κόστος αποθήκευσης, όπως ακριβώς αποτυπώνεται στο MASE, και όχι 100 φορές, όπως θα υποδείκνυε το RMSSE). Συμπερασματικά, οι μετρικές MASE και WAPE «μιλούν τη γλώσσα των επιχειρήσεων», σε αντίθεση με το RMSSE που εξυπηρετεί κυρίως ακαδημαϊκές στατιστικές ερμηνείες.

Αυτή η επιλογή (MASE/WAPE) ευθυγραμμίζεται πλήρως με τις προτάσεις των Syntetos *et al.* [155] για την πρόβλεψη αραιής ζήτησης σε αγορές ανταλλακτικών, καθώς, και του Kolassa [14] για τις σποραδικές πωλήσεις λιανικής, όπου το εκάστοτε σκορ σχετίζεται άμεσα με το πραγματικό κόστος αποθήκευσης ή τις απωλεσθείσες πωλήσεις της επιχείρησης.

5.3.3.7 Μετρικές Αραιής Ζήτησης (Intermittent Demand Metrics)

Για τα προϊόντα που παρουσιάζουν υψηλό βαθμό μηδενικής ζήτησης, υιοθετούνται τρεις συμπληρωματικές μετρικές, εναρμονισμένες με τις προσεγγίσεις της διεθνούς βιβλιογραφίας [1], [2]:

- Σχετική Ακρίβεια Μέσου Απόλυτου Σφάλματος (sMAE_Accuracy):

$$sMAE\_Accuracy = 1 - \frac{MAE}{MAE_{naive}} \tag{5.8}$$

όπου MAE_{naive} είναι το σφάλμα της αφελούς πρόβλεψης όταν χρησιμοποιείται η μέση ιστορική τιμή, δηλαδή

$$MAE_{naive} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \bar{y}| \tag{5.9}$$

Η μετρική λαμβάνει τιμές στο διάστημα [0,1], όπου οι υψηλότερες τιμές υποδηλώνουν καλύτερη απόδοση του συστήματος έναντι της απλής μέσης τιμής.

- Δυαδικό Ποσοστό Επιτυχίας (HitRate_Binary):

Εκφράζει το ποσοστό της ορθής ταξινόμησης μεταξύ περιόδων με ζήτηση και περιόδων χωρίς ζήτηση (δηλαδή, $actual > 0$ έναντι $pred > 0$). Η συγκεκριμένη μετρική αντανακλά ευθέως την ικανότητα του μοντέλου να διακρίνει τα μοτίβα «πώλησης/μη πώλησης».

- Υπό Συνθήκη WAPE (Conditional WAPE – Cond_WAPE):

  Πρόκειται για τη μετρική WAPE υπολογισμένη αποκλειστικά στα είδη (και τις χρονικές περιόδους) που παρουσίασαν μη μηδενική πραγματική ζήτηση. Αντανακλά την ακρίβεια του συστήματος στο «θετικό σκέλος» της ζήτησης. Ουσιαστικά, απομονώνει την απόδοση της παλινδρόμησης από την επίδοση του ταξινομητή, στοιχείο θεμελιώδες για την αξιολόγηση ενός μοντέλου πρόβλεψης δύο σταδίων.

5.3.3.8 Κάλυψη και Εύρος Διαστήματος Πρόβλεψης (Prediction Interval Coverage & Width)

Για τις μεθοδολογικές προσεγγίσεις που εξάγουν διαστήματα πρόβλεψης, δηλαδή η Προσέγγιση 1 μέσω σύμμορφης πρόβλεψης (*conformal prediction*) και η Προσέγγιση 3 μέσω ποσοστημορίων των καταλοίπων εκπαίδευσης (*residual-quantiles*), υπολογίζονται δύο επιπλέον μετρικές.

Αρχικά, η εμπειρική κάλυψη του διαστήματος ορίζεται ως:

$$PI\_Coverage \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\, q_i^{low} \le y_i \le q_i^{high} \,\right] \tag{5.10}$$

Βασικός στόχος είναι η εμπειρική κάλυψη (empirical coverage) να προσεγγίζει το εκάστοτε ονομαστικό επίπεδο. Για παράδειγμα, εάν επιλεγούν τα ποσοστημόρια (*quantiles*) 0,1 και 0,9, τότε το επιθυμητό ποσοστό κάλυψης ανέρχεται στο 80%. Τα ποσοστημόρια $Q_{0.1}$ και $Q_{0.9}$ αντιπροσωπεύουν το 10% και το 90% της αθροιστικής κατανομής της ζήτησης ενός προϊόντος. Ουσιαστικά, αποτυπώνουν το απαισιόδοξο και το έντονα αισιόδοξο σενάριο ζήτησης, αντίστοιχα.

Συμπληρωματικά, το σχετικό εύρος του διαστήματος υπολογίζεται ως εξής:

$$PI\_Width = \frac{q_{high} - q_{low}}{max(\hat{y}, 1)} \tag{5.11}$$

Η εν λόγω μετρική αποτιμά το εύρος του παραγόμενου διαστήματος κανονικοποιημένο ως προς τη σημειακή πρόβλεψη (y). Ο παρονομαστής $\max(\hat{y}, 1)$ διασφαλίζει την αποφυγή απειρισμού σε περιπτώσεις, όπου η παραγόμενη σημειακή πρόβλεψη είναι μηδενική. Επιχειρησιακά, ένα στενότερο εύρος (μικρότερο `PI_Width`), ενώ, παράλληλα, διατηρείται η επιθυμητή κάλυψη, υποδηλώνει μεγαλύτερη βεβαιότητα του μοντέλου. Αυτό, κατά συνέπεια, μεταφράζεται άμεσα σε μικρότερες απαιτήσεις για απόθεμα ασφαλείας.

5.3.4 Αξιολόγηση ανά βαθμίδα τζίρου

Η καθολική τιμή μιας μετρικής ενδέχεται να αποκρύπτει σημαντικές διαφορές επιδόσεων μεταξύ διαφορετικών τμήματων του συνολικού καταλόγου προϊόντων. Δεδομένου ότι η κατανομή τζίρου ακολουθεί μορφή «βαριάς ουράς» (βλ. ενότητα 4.3, Gini = 0,742), οι μετρικές απόδοσης αναφέρονται συστηματικά σε τρία επίπεδα βάσει βαθμίδας τζίρου των ειδών.

- Top-20 (βάσει τζίρου). Περιλαμβάνει τα 20 SKUs με τον υψηλότερο ιστορικό τζίρο. Πρόκειται για κωδικούς επιχειρησιακά κρίσιμους, στους οποίους ακόμα και μια μικρή βελτίωση της πρόβλεψης έχει σημαντική αξία. Τα είδη αυτά αντιπροσωπεύουν περίπου το 7% του συνολικού τζίρου της επιχείρησης.

- Pareto-A. Αποτελείται από τα SKUs που αθροιστικά αντιπροσωπεύουν περίπου το 80% του συνολικού τζίρου (~16.400 SKUs στο παρόν σύνολο δεδομένων). Η συγκεκριμένη ομαδοποίηση αποτελεί συνήθη πρακτική στη βιβλιογραφία ακολουθώντας τη λογική της ταξινόμησης ABC [11], [144].
- Long-tail. Περιλαμβάνει τα υπόλοιπα ~56.000 SKUs, τα οποία χαρακτηρίζονται από αραιή ζήτηση και οριακή συνεισφορά στον τζίρο. Σε αυτό το τμήμα η πρόβλεψη είναι δομικά πολύ πιο δύσκολη, ωστόσο αποτελεί το τμήμα που δεσμεύει τη μεγαλύτερη αξία αποθήκης (βλ. ενότητα 4.5).

Οι κύριες μετρικές του συστήματος (WAPE, Ratio, MASE) υπολογίζονται και αναφέρονται χωριστά για κάθε μία από τις παραπάνω βαθμίδες. Αυτή η πρακτική θεωρείται θεμελιώδης στις σύγχρονες μελέτες πρόβλεψης διαλείπουσας ζήτησης [3], καθώς αποτρέπει το φαινόμενο της αλλοίωσης των αποτελεσμάτων. Ειδικότερα, εμποδίζει τον τεράστιο όγκο της μακράς ουράς (*long-tail*) να επισκιάζει τα όποια σφάλματα πρόβλεψης στα κορυφαία SKUs, και αντίστροφα.

5.4 Ενοποιημένη ροή εκτέλεσης (*unified pipeline*) με Νευρωνικό Δίκτυο Γράφων (GNN)

Η πρώτη ολοκληρωμένη μεθοδολογική προσέγγιση της παρούσας ΔΕ υλοποιεί μια ενοποιημένη ροή εκτέλεσης (*unified pipeline*) ημερήσιας πρόβλεψης ζήτησης, η οποία ενισχύεται μέσω της ενσωμάτωσης δομών γράφου. Η κεντρική ερευνητική υπόθεση στηρίζεται στον εμπλουτισμό των παραδοσιακών μεταβλητών που προκύπτουν από τη μηχανική χαρακτηριστικών (*feature engineering*) με τοπολογικές και δομικές πληροφορίες, οι οποίες κωδικοποιούν σχέσεις συνδυασμένων αγορών (*co-purchase*), κοινής κατηγοριοποίησης, κοινού προμηθευτή και χωρικής εγγύτητας των πελατών.

Ο εμπλουτισμός αυτός επιτυγχάνεται μέσω ενός Νευρωνικού Δικτύου Γράφων Πολλαπλών Έργων (*Multi-Task Learning Graph Neural Network* – MTL-GNN), το οποίο παράγει διανυσματικές αναπαραστάσεις κόμβων (*node* embeddings) ανά κωδικό είδους (SKU). Αυτές οι ενσωματώσεις (embeddings) τροφοδοτούνται ως πρόσθετα χαρακτηριστικά εισόδου σε μια υβριδική παραλλαγή τεσσάρων οικογενειών δενδρικών μοντέλων (XGBoost, LightGBM, CatBoost και RandomForest). Η στρατηγική πρόβλεψης είναι κυλιόμενη (*rolling forecast*) για τρεις χρονικούς ορίζοντες (1, 15 και 90 ημερών), με βασικό άξονα τον κυλιόμενο ημερήσιο ορίζοντα. Βάσει των ημερήσιων αποτελεσμάτων και με τη χρήση μεθόδων χρονικής συνάθροισης (*temporal aggregation*), προκύπτουν οι εκτιμήσεις και για τους περισσότερους χρονικούς ορίζοντες.

Παρά την τεχνική της αρτιότητα, η προσέγγιση αυτή λειτούργησε ως μεθοδολογικός προάγγελος για την τελική Προσέγγιση 3 αναδεικνύοντας κρίσιμους περιορισμούς στην ικανότητα γενίκευσης σε μέσους και μακρούς ορίζοντες, καθώς, και στην αποτελεσματική διαχείριση της λογοκριμένης ζήτησης (*censored demand*) που προκύπτει από τις ελλείψεις αποθεμάτων.

5.4.1 Σκεπτικό και κίνητρο

Η δευτερογενής αγορά των ανταλλακτικών αυτοκινήτων (*automotive aftermarket spare parts*) χαρακτηρίζεται από επιχειρήσεις που διαχειρίζονται έναν εξαιρετικά διευρυμένο κατάλογο προϊόντων, όπου η πλειονότητα των κωδικών παρουσιάζει έντονα διαλείπουσα και αραιή ζήτηση. Η Προσέγγιση 1 σχεδιάστηκε με τη βασική παραδοχή ότι η πληροφορία που ελλείπει από το ιστορικό πωλήσεων ενός μεμονωμένου κωδικού, ιδιαίτερα σε περιπτώσεις νέων προϊόντων χωρίς πρότερο ιστορικό πωλήσεων (πρόβλημα cold-start), μπορεί να ανακτηθεί από τη συμπεριφορά παρόμοιων ή συσχετιζόμενων προϊόντων εντός του δικτύου.

Η ενσωμάτωση ενός δικτύου MTL-GNN είχε ως στόχο:

- **Την αξιοποίηση της τοπολογικής δομής του γράφου,** δηλαδή την κωδικοποίηση και ενσωμάτωση μη γραμμικών σχέσεων, όπως οι συνδυασμένες αγορές προϊόντων από τους ίδιους πελάτες (co-purchase patterns), η κοινή προϊοντική κατηγορία και οι κοινοί προμηθευτές.
- **Την εξειδίκευση ανά χρονικό ορίζοντα,** δηλαδή την παραγωγή διαφοροποιημένων διανυσμάτων αναπαράστασης (embeddings) για τις βραχυπρόθεσμες (1 ημέρα) και τις μεσοπρόθεσμες (15 και 90 ημέρες) ανάγκες πρόβλεψης, μέσω ανεξάρτητων διαδικασιών εκπαίδευσης για κάθε έργο (task).
- **Την αντιμετώπιση του προβλήματος Cold-Start,** δηλαδή την παροχή αξιόπιστων αναπαραστάσεων για νέα προϊόντα που στερούνται ιστορικού πωλήσεων, μέσω ενός ιεραρχικού μηχανισμού αναπλήρωσης (*fallback chain*), ο οποίος βασίζεται στα στατικά χαρακτηριστικά του προμηθευτή, του επιπέδου τιμής του είδους και της κατηγορίας του.

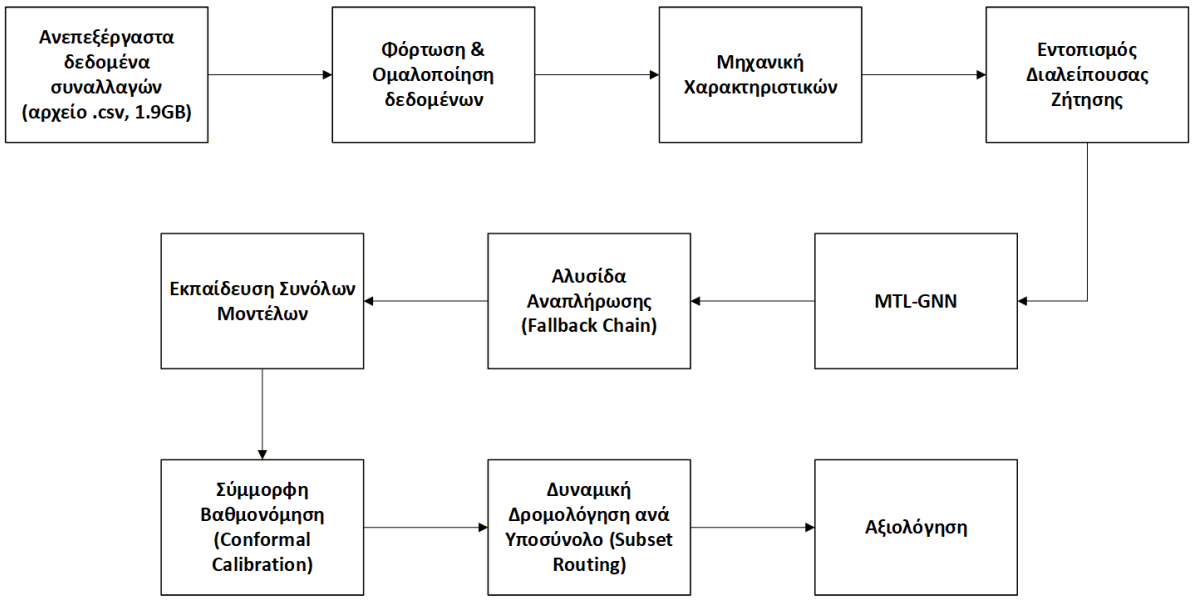
5.4.2 Αρχιτεκτονική και ροή δεδομένων

Η ροή εκτέλεσης (pipeline) του συστήματος δομείται σε εννέα διαδοχικά στάδια (Σχήμα 5.1), με πρωταρχικό μέλημα τη διασφάλιση της ακεραιότητας των δεδομένων και την απόλυτη αποτροπή διαρροής πληροφορίας (data leakage) κατά την αξιολόγηση.

- **Φόρτωση και ομαλοποίηση (*Data loading and normalization*):** Πραγματοποιείται διαχωρισμός των καθαρών πωλήσεων από τις λοιπές συναλλαγές (π.χ., επιστροφές, απογραφές) και δημιουργείται μια αραιή αναπαράσταση επιπέδου `SKU-Day` (συμπίεση όγκου ~98%), όπου αθροίζονται όλες οι συναλλαγές πωλήσεων που λαμβάνουν χώρα ανά κωδικό είδους σε ημερήσια βάση. Με τον τρόπο αυτό αποφεύγεται η παραδοσιακή συμπλήρωση με μηδενικά (*zero-filling*), η οποία σύμφωνα με τη βιβλιογραφία εισάγει συστηματική μεροληψία υποεκτίμησης (*under-forecast bias*) σε δεδομένα αραιής ζήτησης [1][3].
- **Μηχανική χαρακτηριστικών (*Feature engineering*):** Παράγεται ένα πλούσιο σύνολο 40 χαρακτηριστικών, το οποίο περιλαμβάνει ημερολογιακές μεταβλητές, υστερήσεις (*lags*), εποχικότητα βάσει μετασχηματισμού Fourier και επιδράσεις εορτών (*holidays*).
- **Εντοπισμός διαλείπουσας ζήτησης (*Intermittent demand tagging*):** Τα προϊόντα (SKUs) αξιολογούνται και κατατάσσονται βάσει του συντελεστή μεταβλητότητας (CV > 0,7), επιτρέποντας τη μετέπειτα εφαρμογή ασύμμετρης στάθμισης κατά την εκπαίδευση.
- **Εκπαίδευση Νευρωνικού Δικτύου (MTL-GNN):** Κατασκευάζεται ένας γράφος προϊόντων (12 χαρακτηριστικά ανά κόμβο) και εκπαιδεύεται με διπλό στόχο. Αρχικά, για την εκμάθηση αναπαραστάσεων μέσω μιας συνάρτησης αντιθετικής απώλειας (*contrastive loss*) και, εν συνεχεία, με τη χρήση μιας βοηθητικής παλινδρόμησης ζήτησης (*auxiliary demand regression*).
- **Αλυσίδα αναπλήρωσης (*Fallback chain*):** Εφαρμόζεται ένας ιεραρχικός μηχανισμός απόδοσης διανυσμάτων (embeddings) σε νέα προϊόντα χωρίς ιστορικό (*cold-start*), αξιοποιώντας αναπαραστάσεις που αντλούνται από τα στατικά χαρακτηριστικά του προμηθευτή (π.χ., `Supplier SVD embeddings`).
- **Εκπαίδευση συνόλων μοντέλων (*Tree ensembles training*):** Εκπαιδεύονται παράλληλα διαφορετικές οικογένειες δενδρικών μοντέλων (XGBoost, LightGBM, CatBoost, Random Forests). Για κάθε αλγόριθμο αναπτύσσονται δύο διακριτές εκδοχές, μια γραμμή βάσης (*baseline*) και μια υβριδική (*hybrid*), η οποία ενσωματώνει τα διανύσματα χαρακτηριστικών του γράφου. Στη συνέχεια, τα επιμέρους μοντέλα συνδυάζονται στα αντίστοιχα σύνολα (*baseline* και *hybrid ensembles*). Κατά την εκπαίδευση, εφαρμόζεται η συνάρτηση απώλειας `Tweedie` [73] επί των αρχικών, μη μετασχηματισμένων στόχων (στα XGBoost και LightGBM)

και η συνάρτηση μέσου τετραγωνικού σφάλματος (MSE) επί λογαριθμικά μετασχηματισμένων στόχων (στα CatBoost και Random Forests).

- **Σύμμορφη βαθμονόμηση (*Conformal calibration*):** Εφαρμόζονται τεχνικές σύμμορφης πρόβλεψης για την παραγωγή διαστημάτων εμπιστοσύνης διασφαλίζοντας ονομαστική κάλυψη 85% για τα παραγόμενα διαστήματα πρόβλεψης.
- **Δυναμική δρομολόγηση (*Subset routing*):** Υλοποιείται ένας μηχανισμός επιλογής της βέλτιστης παραλλαγής μεταξύ του μοντέλου βάσης και του υβριδικού μοντέλου, διαφοροποιημένος ανάλογα με το τμήμα του καταλόγου στο οποίο ανήκει το εκάστοτε προϊόν (π.χ., είδη κορυφαίου τζίρου, είδη χαμηλής κυκλοφορίας, νέα προϊόντα).
- **Αξιολόγηση και συνάθροιση (*Evaluation and aggregation*):** Ολοκληρώνεται η διαδικασία με τη συγκριτική ανάλυση των αποτελεσμάτων και την εφαρμογή στρατηγικών χρονικής συνάθροισης (*temporal aggregation*) για τη μετάβαση από τον ημερήσιο σε μεγαλύτερους χρονικούς ορίζοντες.



Σχήμα 5.1: Ροή δεδομένων συστήματος πρόβλεψης

5.4.3 Μηχανική χαρακτηριστικών και ορισμός στόχων

Μια κρίσιμη σχεδιαστική επιλογή για τη διαμόρφωση του πίνακα εκπαίδευσης της Προσέγγισης 1 ήταν η υιοθέτηση μιας αμιγώς αραιής (*sparse*) αναπαράστασης σε επίπεδο `SKU-Day` αποφεύγοντας την τεχνητή πλήρωση των ημερών μηδενικής πώλησης (*zero-filling*). Η στρατηγική αυτή απέτρεψε τη συστηματική μεροληψία υπο-πρόβλεψης (*under-forecast bias*), η οποία εισάγεται νομοτελειακά, όταν οι μηδενικές παρατηρήσεις κυριαρχούν συντριπτικά στις συναρτήσεις κόστους (*cost functions*) και «ωθούν» τους αλγορίθμους να συγκλίνουν προς το μηδέν.

Ως μεταβλητές στόχοι (`targets`) του συστήματος ορίστηκαν οι πραγματικές ημερήσιες ποσότητες ζήτησης ανά κωδικό προϊόντος. Τα μοντέλα εκπαιδεύονται για την απευθείας εκτίμηση της ζήτησης αυτής στους τρεις επιλεγμένους χρονικούς ορίζοντες ($h = 1,15$ και 90 ημέρες), είτε επί των αρχικών φυσικών τιμών είτε επί λογαριθμικά μετασχηματισμένων στόχων, ανάλογα με τη συνάρτηση απώλειας του εκάστοτε αλγορίθμου.

Για τον εμπλουτισμό της πληροφορίας, παράγεται ένα διευρυμένο σύνολο 40 χαρακτηριστικών εισόδου (features), τα οποία κατηγοριοποιούνται σε τέσσερις βασικούς άξονες:

- **Ημερολογιακά χαρακτηριστικά (*Calendar features*).** Εφαρμόζεται κυκλική (τριγωνομετρική) κωδικοποίηση μέσω μετασχηματισμού ημιτόνου και συνημιτόνου για τις μεταβλητές του μήνα και της ημέρας της εβδομάδας διασφαλίζοντας ότι η χρονική συνέχεια και η κυκλικότητα (π.χ. η μετάβαση από τον Δεκέμβριο στον Ιανουάριο) αποτυπώνονται ορθά από τα μοντέλα. Επιπλέον, ενσωματώνονται δυαδικοί δείκτες (*flags*) για τις επίσημες ελληνικές αργίες (π.χ. Πάσχα, Χριστούγεννα), οι οποίες επηρεάζουν άμεσα τη λειτουργία και την καταναλωτική κίνηση της αγοράς.
- **Υστερήσεις και κυλιόμενα στατιστικά (*Lags and rolling statistics*).** Υπολογίζονται χρονικές υστερήσεις (*lags*) της ιστορικής ζήτησης που εκτείνονται έως και τις 365 ημέρες στο παρελθόν, με σκοπό τη σύλληψη της ετήσιας υστέρησης. Παράλληλα, εισάγονται κυλιόμενοι μέσοι όροι (*rolling means*) χρονικού παραθύρου 7, 14 και 30 ημερών για την αποτύπωση των βραχυπρόθεσμων και μεσοπρόθεσμων τάσεων (trends) της αγοράς.
- **Εποχικότητα μέσω αρμονικών Fourier (*Fourier seasonality*).** Συντίθενται αρμονικές συναρτήσεις Fourier σε ετήσια, εξαμηνιαία και τριμηνιαία βάση. Η προσέγγιση αυτή επιτρέπει στα δενδρικά μοντέλα να προσαρμόζονται ομαλά στα μακροπρόθεσμα κυκλικά μοτίβα ζήτησης, τα οποία σχετίζονται άμεσα με τα πρότυπα συντήρησης οχημάτων στην Ελλάδα (π.χ. αυξημένη προετοιμασία και έλεγχος οχημάτων πριν από τη θερινή ταξιδιωτική περίοδο και τη χειμερινή περίοδο).
- **Δείκτες διαλείπουσας ζήτησης (*Intermittent demand indicators*).** Ενσωματώνονται οι στατιστικοί δείκτες ADI (Μέσο Διάστημα μεταξύ Ζήτησης) και CV^2 (Τετράγωνο του Συντελεστή Μεταβλητότητας). Υπολογίζονται δυναμικά ανά SKU, ώστε να παρέχουν στα μοντέλα τη δομική πληροφορία του βαθμού σποραδικότητας του κάθε κωδικού (βάσει της ταξινόμησης Syntetos–Boylan [5]).

Ιδιαίτερη έμφαση δίνεται στην εισαγωγή του εξειδικευμένου χαρακτηριστικού κατηγορίας ποιότητας (`Quality_Segment`). Το εν λόγω χαρακτηριστικό απομονώνει και επισημαίνει το κορυφαίο 30% των προϊόντων υψηλής αξίας ή αυξημένης κίνησης ("*premium*" κωδικοί) εντός της ίδιας ομάδας εναλλακτών ειδών, η οποία ορίζεται από το κοινό Μητρικό Κωδικό Προϊόντος (`Product_Master_ID`). Η πληροφορία αυτή είναι κρίσιμη, καθώς αποτρέπει τα μοντέλα από το να εξομοιώσουν λανθασμένα τα επώνυμα ανταλλακτικά υψηλής κυκλοφορίας με δευτερεύοντες, φθηνότερους ή σπάνια κινούμενους εναλλακτικούς κωδικούς που ανήκουν στην ίδια ακριβώς τεχνική ομάδα.

5.4.4 Αρχιτεκτονική του Νευρωνικού Δικτύου Γράφων Πολλαπλών Έργων (MTL-GNN)

Ο πυρήνας της πρώτης μεθοδολογικής προσέγγισης εντοπίζεται στην αρχιτεκτονική `MultiTaskGNN`, η οποία υλοποιήθηκε στο υπολογιστικό περιβάλλον του `PyTorch Geometric` [150]. Η συγκεκριμένη δομή αναπτύχθηκε με σκοπό την εκμάθηση συμπυκνωμένων διανυσματικών αναπαραστάσεων των προϊόντων, αξιοποιώντας την τοπολογική πληροφορία του γράφου σε δύο διακριτές, ανταγωνιστικές σχέσεις (θετική και αρνητική). Η παραγωγή των τελικών διανυσμάτων διαφοροποιείται ανά χρονικό ορίζοντα πρόβλεψης μέσω εξειδικευμένων κεφαλών εξόδου (*prediction heads*). Ο σχεδιασμός του δικτύου αντλεί αρχές από τα Σχεσιακά Συνελικτικά Δίκτυα Γράφων (Relational Graph Convolutional Networks – *R-GCN*) [156] και το θεωρητικό πλαίσιο της μάθησης πολλαπλών έργων (*multi-task learning*) [88], [157].

5.4.4.1 Κατασκευή του γράφου προϊόντων

Η κατασκευή του γράφου υλοποιείται από την κλάση `GraphBuilder`. Ως μοναδική είσοδος χρησιμοποιείται το υποσύνολο `df_train_raw` που περιλαμβάνει αποκλειστικά εγγραφές καθαρών πωλήσεων. Έτσι, διασφαλίζεται μια διαδικασία απαλλαγμένη από διαρροή πληροφορίας (*leakage-free graph construction*) προς το σύνολο αξιολόγησης.

Κόμβοι (Nodes)

Ορίζεται ένας κόμβος για κάθε μοναδικό κωδικό προϊόντος (`Product_Code`) που εντοπίζεται στο σύνολο εκπαίδευσης, οδηγώντας σε ένα σταθερό πλήθος $\boldsymbol{N = 64.342}$ κόμβων. Σε κάθε κόμβο αντιστοιχίζεται ένα διάνυσμα τριών αρχικών στατικών χαρακτηριστικών:

- `Quality_Segment` (Δυαδικός Δείκτης Ποιότητας): Λαμβάνει την τιμή 1 για τους premium κωδικούς (κορυφαίο 30% του `Product_Master ID`) και 0 σε διαφορετική περίπτωση.
- Κανονικοποιημένη Λογαριθμική Δημοτικότητα: Υπολογίζεται ως $\log(1 + p_i) \,/\, \log(1 + p_{max})$, όπου p_i εκφράζει το πλήθος των εμφανίσεων (συναλλαγών) του κωδικού i στο σύνολο εκπαίδευσης.
- Τυποποιημένη Τιμή (Z-score): Αντιπροσωπεύει τη στατιστική τυποποίηση της διάμεσης τιμής μονάδας του προϊόντος σε σχέση με τη συγκεκριμένη προϊοντική ιεραρχία στην οποία ανήκει (`Product_Hierarchy2_Level_3_Code` – κωδικός κατηγορίας), με επιβολή ορίων αποκοπής (clipping) στο διάστημα $[-3, +3]$ για τον περιορισμό των ακραίων τιμών.

Θετικές ακμές (E_{pos})

Αντιπροσωπεύουν σχέσεις συμπληρωματικότητας και συνδυαστικής ζήτησης. Προκύπτουν από τη συνένωση δύο ανεξάρτητων πηγών πληροφορίας:

- Ταυτόχρονες πωλήσεις μέσω παραστατικών (*Invoice-based co-occurrence*): Κατασκευάζεται ένας αραιός πίνακας σχέσεων παραστατικού-προϊόντος $\mathbf{M} \in \{0,1\}^{I \times N}$. Το εσωτερικό γινόμενο $\mathbf{M}^T\mathbf{M}$ παράγει τον πίνακα ταυτόχρονων εμφανίσεων μεταξύ των προϊόντων. Οι διαγώνιοι όροι μηδενίζονται, ενώ διατηρούνται μόνο οι ακμές με βάρος $w > 1$ για το φιλτράρισμα τυχαίων, μεμονωμένων συνδυασμών.
- Χρονικά παράθυρα αγορών ανά πελάτη (*Customer-window co-occurrence*): Για κάθε πελάτη ξεχωριστά, οι αγορές που πραγματοποιούνται εντός κυλιόμενου παραθύρου 2 ημερών συνδέονται μεταξύ τους. Για την αποφυγή υπολογιστικής συμφόρησης και την εξάλειψη του στατιστικού θορύβου, εξαιρούνται από τη διαδικασία πελάτες με περισσότερες από 400 ετήσιες αγορές (κυρίως μεγάλοι λογαριασμοί χονδρικής με μη-εστιασμένη αγοραστική συμπεριφορά).

Οι δύο παραπάνω πηγές ενοποιούνται αθροιστικά και το τελικό σύνολο περιορίζεται στις top-K (το K ορίζεται από το χρήστη στο αρχείο ρυθμίσεων, στις δοκιμές που έγιναν, χρησιμοποιήθηκαν τιμές από 5 έως 10) ισχυρότερες ακμές ανά κόμβο, αποδίδοντας συνολικά **407.770 θετικές ακμές** (αμφίδρομης κατεύθυνσης με `K=8`).

Αρνητικές ακμές (E_{neg})

Κωδικοποιούν σχέσεις υποκατάστασης (*substitutability*). Στον κλάδο των ανταλλακτικών αυτοκινήτων, προϊόντα που μοιράζονται το ίδιο `Product_Master_ID` αποτελούν εναλλακτικές κατασκευαστικές εκδοχές (διαφορετικοί κατασκευαστές) του ίδιου ακριβώς εξαρτήματος. Κατά συνέπεια, η ζήτηση λειτουργεί ανταγωνιστικά, γιατί σε κάθε αγορά ο πελάτης επιλέγει μία εκ των εναλλακτικών. Άρα, επιβάλλεται μια διακριτή και απωθητική αναπαράσταση των θέσεών τους στον χώρο των embeddings. Το στάδιο αυτό παράγει **234.740 αρνητικές ακμές** (αμφίδρομης κατεύθυνσης).

Σημειώνεται ότι ο λόγος πυκνότητας των θετικών προς τις αρνητικές ακμές ($E_{pos}/E_{neg} \approx 1{,}74$) εναρμονίζεται με τις βέλτιστες πρακτικές των Schlichtkrull *et al.* [156] και Yang *et al.* [158], οι οποίοι προκρίνουν μια ελαφρώς συμπαγέστερη θετική δομή για τη διασφάλιση της σταθερότητας κατά τη βελτιστοποίηση της αντιθετικής συνάρτησης απώλειας.

5.4.4.2 Μαθηματική αρχιτεκτονική του δικτύου

Το μοντέλο `MultiTaskGNN` ακολουθεί μια εξειδικευμένη διάταξη «Ύστερης Συγχώνευσης» (*Late Fusion)* δύο παράλληλων καναλιών, η οποία δομείται ως εξής:

- **Πίνακας αναζήτησης ενσωματώσεων (Embeddings *lookup table*):** Ορίζεται ένας εκπαιδεύσιμος πίνακας $\mathbf{E} \in \mathbb{R}^{N\times 32}$, ο οποίος παρέχει τη βασική αναπαράσταση κάθε SKU. Αρχικοποιείται τυχαία κατά Xavier (Xavier *Initialization*) [159] για τη διασφάλιση σταθερής διακύμανσης των κλίσεων και την αποτροπή του φαινομένου της εξαφάνισης της κλίσης (*vanishing gradients*) κατά τα πρώτα στάδια της εκπαίδευσης.
- **Προβολή και σύνθεση χαρακτηριστικών *(Feature projection*):** Τα αρχικά χαρακτηριστικά των 3 διαστάσεων προβάλλονται γραμμικά στον χώρο των 32 διαστάσεων μέσω ενός επιπέδου $\mathbf{W}_{feat} \in \mathbb{R}^{3\times 32}$ και **προστίθενται** αθροιστικά στον πίνακα **E**. Η επιλογή της προσθέσεως (*addition*) έναντι της συνένωσης (*concatenation*) αποτρέπει την άσκοπη αύξηση των διαστάσεων.
- **Παράλληλα επίπεδα συνέλιξης γράφου (*Parallel two-relation* GCN):** Αντί για την κλασική R-GCN αρχιτεκτονική, το δίκτυο χρησιμοποιεί δύο ανεξάρτητα, παράλληλα κανάλια δύο επιπέδων `GCNConv` [86] με ενδιάμεση ενεργοποίηση `ReLU` και `dropout = 0.15`. Για κάθε σχέση $r \in \{pos, neg\}$ και για κάθε κόμβο i, η διάδοση μηνύματος (*message passing*) διαμορφώνεται ως:

$$\mathbf{h}_i^{(\ell+1,r)} = \sigma\left(\sum_{j\in\mathcal{N}^{(r)}(i)} \frac{w_{ij}^{(r)}}{\sqrt{\hat{d}_i^{(r)}\hat{d}_j^{(r)}}} \cdot \mathbf{h}_j^{(\ell,r)}\mathbf{W}^{(\ell,r)}\right) \tag{5.12}$$

  όπου $\hat{d}$ αντιπροσωπεύει τον κανονικοποιημένο βαθμό του κόμβου με την προσθήκη αυτοβρόχων(*self-loops*).
- **Επίπεδο σύνθεσης (*Combine layer*):** Τα δύο παραγόμενα διανύσματα ($32 + 32 = 64$ διαστάσεων) συνενώνονται και τροφοδοτούνται σε ένα κοινό επίπεδο `LayerNorm`, παράγοντας το κοινό διάνυσμα αναπαράστασης (*shared* embedding):

$$\mathbf{h}_i^{shared} = \mathrm{LayerNorm}\left(\mathrm{Linear}_{64\to 32}\left(\left[\mathbf{h}_i^{(pos)} \,\|\, \mathbf{h}_i^{(neg)}\right]\right)\right) \tag{5.13}$$

- **Εξειδικευμένες κεφαλές ανά ορίζοντα (Horizon-specific heads):** Για κάθε χρονικό ορίζοντα $h \in \{1{,}15{,}90\}$, το `shared_embedding` (κοινό διάνυσμα αναπαράστασης) διέρχεται από ένα αυτόνομο δίκτυο `MLP` *(Multi-Layer Perceptron)* δύο επιπέδων (`Linear → ReLU → Dropout → Linear`), παράγοντας τα τελικά διανύσματα $\mathbf{h}_{i,h} \in \mathbb{R}^{32}$. Με τον τρόπο αυτό, ενώ το κεντρικό διάνυσμα αναπαράστασης διατηρεί τις καθολικές τοπολογικές ιδιότητες, οι επιμέρους κεφαλές εξειδικεύουν την πληροφορία ανάλογα με τη χρονική κλίμακα (π.χ., βραχυπρόθεσμα αγοραστικά πρότυπα για $h = 1$, μακροπρόθεσμη εποχικότητα για $h = 90$).

5.4.4.3 Πολυ-λειτουργική συνάρτηση κόστους και εκπαίδευση

Η εκπαίδευση του δικτύου πραγματοποιείται με τη μέθοδο της μάθησης πολλαπλών έργων (m*ulti-task learning*). Για να καταστεί δυνατή η οπισθοδιάδοση (back-propagation) του σφάλματος

παλινδρόμησης, κατά τη φάση της προ-εκπαίδευσης (*pre-training*) τοποθετείται προσωρινά πάνω από κάθε `horizon_head` ένα γραμμικό επίπεδο προβολής σε βαθμωτή τιμή, το οποίο αναλαμβάνει την πρόβλεψη του αθροίσματος της μελλοντικής ζήτησης. Η καθολική συνάρτηση κόστους $\mathcal{L}_{total}$ διαμορφώνεται ως:

$$\mathcal{L}_{total} = \mathcal{L}_{contrastive} + \gamma \cdot \mathcal{L}_{aux} + \lambda \cdot \|\mathbf{\Theta}\|_2^2 \tag{5.14}$$

Η αντιθετική απώλεια ($\mathcal{L}_{contrastive}$) βασίζεται σε συνάρτηση *softplus* [160], [161] και ορίζεται ως:

$$\mathcal{L}_{contrastive} = \mathbb{E}_{(i,j)\in E_{pos}}\left[softplus\left(-\langle \hat{\boldsymbol{h}}_i, \hat{\boldsymbol{h}}_j \rangle\right)\right] + \mathbb{E}_{(i,j)\in E_{neg}}\left[softplus\left(+\langle \hat{\boldsymbol{h}}_i, \hat{\boldsymbol{h}}_j \rangle\right)\right] \tag{5.15}$$

όπου $\hat{\mathbf{h}}_i = \mathbf{h}_i / \|\mathbf{h}_i\|_2$ εκφράζει τα L2-κανονικοποιημένα embeddings. Η βοηθητική συνάρτησης απώλειας παλινδρόμησης ($\mathcal{L}_{aux}$) υπολογίζεται ως το μέσο τετραγωνικό σφάλμα (MSE) των προσωρινών κεφαλών για το σύνολο των οριζόντων.

Για τον περιορισμό του υπολογιστικού φόρτου, σε κάθε εποχή εκπαίδευσης εφαρμόζεται υπο-δειγματοληψία (*mini-batch sampling*) 20.000 θετικών και 20.000 αρνητικών ακμών. Η συγκεκριμένη στρατηγική επιλογής περιορισμένου πλήθους αρνητικών παραδειγμάτων αντλεί την έμπνευσή της από την τεχνική της Αρνητικής Δειγματοληψίας (*Negative Sampling*) που καθιέρωσαν οι Mikolov *et al.* [160] επιτρέποντας την αποδοτική εκπαίδευση αναπαραστάσεων χωρίς τον εξαντλητικό υπολογισμό ολόκληρου του γράφου. Η βελτιστοποίηση εκτελείται μέσω του αλγορίθμου `AdamW` ($lr = 3 \cdot 10^{-3}$, $weight\_decay = 10^{-4}$) για 120 εποχές σε περιβάλλον CUDA.

5.4.4.4 Ιεραρχική αλυσίδα αναπλήρωσης νέων προϊόντων (*Cold-start fallback chain*)

Από το σύνολο των 64.342 κωδικών του γράφου, ένα ποσοστό ύψους 7,28% των προϊόντων που εμφανίζονται στο σύνολο αξιολόγησης στερείται πλήρως ιστορικού κωδικοποίησης στο σύνολο εκπαίδευσης (`Product_Idx == -1`). Για την αντιμετώπιση αυτού του προβλήματος (cold-start), αναπτύχθηκε ένας πολυεπίπεδος ιεραρχικός μηχανισμός μέσω της συνάρτησης `get_embeddings_with_fallback()`, ο οποίος αποδίδει διανύσματα με βάση το επίπεδο διαθέσιμης στατικής πληροφορίας:

$$\mathbf{h}i^{cold} = \{\boldsymbol{\mu}supplier, \quad \text{εάν ο Προμηθευτής είναι γνωστός}\backslash$$
$$\boldsymbol{\mu}cat \times qual, \&\text{εάν υπάρχει κοινή Κατηγορία και Segment Ποιότητας}\backslash$$
$$\boldsymbol{\mu}cat, \&\text{εάν υπάρχει μόνο κοινή Κατηγορία}\backslash \tag{5.16}$$
$$\boldsymbol{\mu}_{global}, \&\text{σε κάθε άλλη περίπτωση (Καθολικό Κεντροειδές)}$$

Όπου:

$\boldsymbol{\mu}_{supplier}$: Αντιπροσωπεύει το μέσο διάνυσμα (centroid) των embeddings όλων των γνωστών προϊόντων του ίδιου προμηθευτή.

$\boldsymbol{\mu}_{cat\times qual}$: Αντιπροσωπεύει το μέσο διάνυσμα των κωδικών που μοιράζονται την ίδια προϊοντική κατηγορία και το ίδιο segment ποιότητας.

$\boldsymbol{\mu}_{cat}$: Αντιπροσωπεύει το μέσο διάνυσμα της κατηγορίας.

$\boldsymbol{\mu}_{global}$: Αντιπροσωπεύει τον μέσο όρο του συνόλου των εκπαιδευμένων embeddings του γράφου.

Αυτή η αυστηρά δομημένη αλυσίδα αναπλήρωσης αποτελεί ένα από τα σημαντικότερα πλεονεκτήματα της Προσέγγισης 1, καθώς εξασφαλίζει σταθερή επιχειρησιακή λειτουργία και παρέχει αξιόλογη

βελτίωση της ακρίβειας πρόβλεψης αποκλειστικά για την κατηγορία των νέο-εισαχθέντων προϊόντων (cold-start SKUs).

5.4.5 Μοντελοποίηση και στρατηγικές διαχείρισης σφάλματος

Στο στάδιο της μοντελοποίησης, η ενοποιημένη ροή εκτέλεσης (*pipeline*) του συστήματος εκπαιδεύει συνολικά 24 διακριτά μοντέλα ανά πείραμα. Για καθέναν από τους τρεις ορίζοντες πρόβλεψης ($h \in \{1,15,90\}$ ημέρες), επιστρατεύονται τέσσερις διαφορετικές οικογένειες μοντέλων δέντρων αποφάσεων. Αυτές εκπαιδεύονται τόσο στη βασική (*baseline*) όσο και στην υβριδική (*hybrid*) παραλλαγή αρχιτεκτονικής. Η συγκεκριμένη στρατηγική επιτρέπει την αυστηρή και συστηματική αξιολόγηση της προστιθέμενης αξίας των GNN embeddings σε διαφορετικές χρονικές κλίμακες.

5.4.5.1 Αρχιτεκτονική και διαφοροποίηση μοντέλων

Η επιλογή των τεσσάρων διαφορετικών αλγορίθμων δεν είναι τυχαία, αλλά εδράζεται στη θεωρία της αντιστάθμισης μεροληψίας-διακύμανσης (*bias-variance trade-off*) και στη στόχευση προς τη μέγιστη δυνατή πολυμορφία του συνόλου (*ensemble diversity*) [38], [123]. Αναλυτικότερα:

- `XGBoost` **&** `LightGBM`**:** Επιλέγονται ως οι πλέον ισχυροί εκπρόσωποι της ενισχυμένης μάθησης με κλίση (*gradient boosting*). Το μοντέλο `XGBoost` παρέχει εξαιρετική στιβαρότητα μέσω της ενσωματωμένης `L1/L2 κανονικοποίησης (regularization)`, ενώ το `LightGBM` υπερέχει σε υπολογιστική ταχύτητα και διαχείριση μεγάλου όγκου δεδομένων, χάρη στη στρατηγική ανάπτυξης δέντρων ανά φύλλο (*leaf-wise growth*).
- `CatBoost`**:** Ενσωματώνεται στο σύστημα λόγω της εγγενούς ικανότητάς του να διαχειρίζεται αποδοτικά τα κατηγορικά δεδομένα, καθώς και της αρχιτεκτονικής της διατεταγμένης ενίσχυσης (*ordered boosting*), η οποία εξαλείφει σε μεγάλο βαθμό τη διαρροή πληροφορίας στόχου (*target leakage*).
- `Random Forest`**:** Λειτουργώντας ως μέθοδος συνόλου (*bagging*), προσφέρει μια εντελώς διαφορετική προσέγγιση στη μείωση της διασποράς και αποτελεί τον πιο σταθερό εκτιμητή σε περιβάλλοντα με ακραίες τιμές (*outliers*) και έντονο στατιστικό θόρυβο.

5.4.5.2 Διαχείριση μηδενικής ζήτησης μέσω κατανομής Tweedie

Η ζήτηση στην αγορά ανταλλακτικών αυτοκινήτων (*automotive aftermarket*) χαρακτηρίζεται από υπερβολική σποραδικότητα και παρουσιάζει έντονη συγκέντρωση μάζας στο μηδέν (*zero-inflated*). Για την αντιμετώπιση αυτής της ιδιαιτερότητας υιοθετείται η οικογένεια κατανομών `Tweedie`, η οποία συνιστά ένα μίγμα κατανομών Poisson και Gamma (*Compound Poisson-Gamma*). Ο συγκεκριμένος σχεδιασμός επιτρέπει τη μοντελοποίηση τόσο των διακριτών συμβάντων μηδενικής ζήτησης όσο και των συνεχών θετικών ποσοτήτων [73], [76].

Στη μηχανική μάθηση, ο αλγόριθμος βελτιστοποιείται ελαχιστοποιώντας τη Συνάρτηση Απόκλισης Tweedie (Tweedie Deviance Loss), η οποία για την παράμετρο διασποράς $p \in (1,2)$ ορίζεται ως:

$$d(y,\mu) = 2\left(\frac{y^{2-p}}{(1-p)(2-p)} - \frac{y\mu^{1-p}}{1-p} + \frac{\mu^{2-p}}{2-p}\right) \tag{5.17}$$

Στο παρόν σύστημα, η παράμετρος ρυθμίστηκε στην τιμή $p = 1.5$. Κρίσιμη μεθοδολογική επιλογή αποτέλεσε η εφαρμογή της συνάρτησης κόστους Tweedie **απευθείας επί των αρχικών, μη μετασχηματισμένων (*raw*) δεδομένων** για τα μοντέλα ενίσχυσης. Η προσέγγιση αυτή ακολουθήθηκε συνειδητά, ώστε να αποφευχθεί το φαινόμενο της «διπλής λογαρίθμισης». Η συνάρτηση Tweedie διαθέτει εσωτερικό λογαριθμικό σύνδεσμο (*log-link function*), οπότε η εφαρμογή της σε ήδη

λογαριθμισμένα δεδομένα (π.χ., $\log(y+1)$) θα προκαλούσε συστηματική και σοβαρή υπό-πρόβλεψη (*under-forecasting bias*), ανατρέποντας την επιχειρησιακή αξιοπιστία του συστήματος.

5.4.5.3 Ασύμμετρη στάθμιση αιχμών ζήτησης (*Spike-awareness*)

Λόγω των σοβαρών επιχειρησιακών και οικονομικών επιπτώσεων που επιφέρει μία πιθανή έλλειψη αποθεμάτων (*stock-out*), το σύστημα δεν αντιμετωπίζει συμμετρικά τα σφάλματα, αλλά ενσωματώνει έναν μηχανισμό ασύμμετρης ποινής. Το κατώφλι υπολογισμού της ασύμμετρης ποινής ορίζεται στο αρχείο ρυθμίσεων του pipeline μέσω της παραμέτρου `SPIKE_CV_THRESH = 0.7` και λειτουργεί ως εξής:

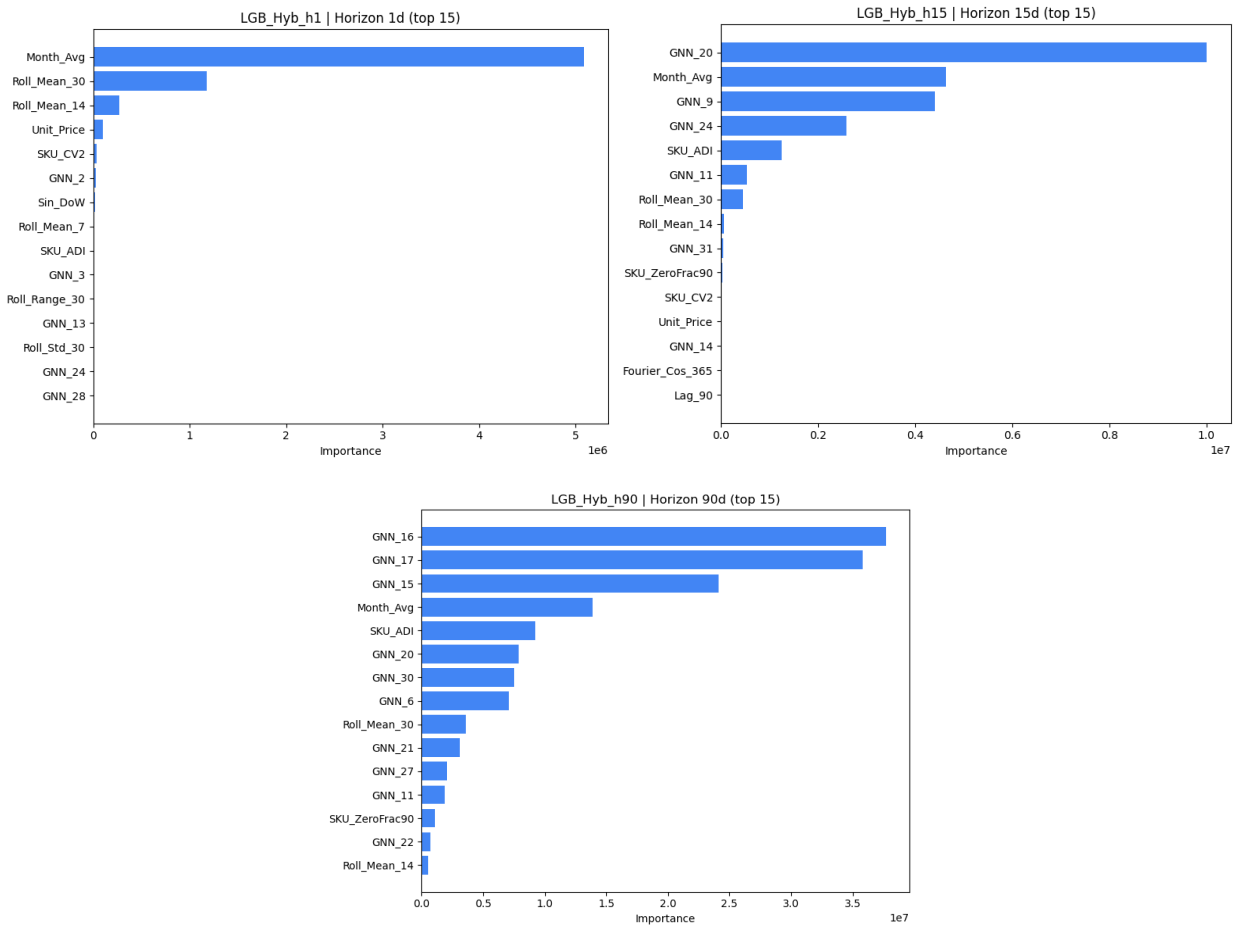
- **Εντοπισμός διαλείπουσας ζήτησης:** Οι κωδικοί προϊόντων (SKUs), των οποίων ο τετραγωνικός συντελεστής μεταβλητότητας υπερβαίνει το καθορισμένο κατώφλι ($CV^2 > 0.7$) ταξινομούνται αυτόματα ως είδη με διαλείπουσα και ακανόνιστη συμπεριφορά (*intermittent demand*).
- **Στάθμιση δειγμάτων:** Στις εγγραφές (γραμμές) του πίνακα εκπαίδευσης που αντιστοιχούν σε αυτούς τους μεταβλητούς κωδικούς κατά τη διάρκεια ημερών όπου καταγράφεται πραγματική ζήτηση, αποδίδεται διπλάσιο βάρος εκπαίδευσης ($w_i = 2.0$).

Αυτή η στατιστική στάθμιση αναγκάζει το μοντέλο να «προσέχει» ιδιαίτερα τις περιπτώσεις αιχμών ζήτησης, τιμωρώντας αυστηρότερα την υπό-εκτίμηση σε σχέση με την υπερ-εκτίμηση. Η υλοποίηση αυτή είναι πλήρως συμβατή με την απώλεια Tweedie, καθώς η επιβολή της ποινής εφαρμόζεται γραμμικά μέσω του πίνακα βαρών των δειγμάτων (*sample weights*) και όχι μέσω τροποποίησης της ίδιας της μαθηματικής αντικειμενικής συνάρτησης διασφαλίζοντας, έτσι, την απόλυτη ευστάθεια του αλγορίθμου βελτιστοποίησης.

5.4.6 Αποτελέσματα και εμπειρική επικύρωση

Η ανάλυση της σημαντικότητας των χαρακτηριστικών (*feature importance*) ανέδειξε με σαφήνεια την προστιθέμενη αξία της αρχιτεκτονικής GNN. Όπως παρατηρήθηκε, όσο αυξάνεται ο χρονικός ορίζοντας πρόβλεψης, τόσο η κυριαρχία των GNN διανυσμάτων (embeddings) ενισχύεται εις βάρος των παραδοσιακών δεικτών. Το φαινόμενο αυτό ερμηνεύεται θεωρητικά από το γεγονός ότι τα στατιστικά χαρακτηριστικά (π.χ. κυλιόμενος μέσος όρος `Roll_Mean_30`) πάσχουν από «εξασθένηση σήματος» σε μακρινούς ορίζοντες, καθώς η πρόσφατη πληροφορία καθίσταται παρωχημένη. Αντιθέτως, τα δομικά χαρακτηριστικά του γράφου, τα οποία κωδικοποιούν βαθιές τοπολογικές συσχετίσεις και μοτίβα συμπληρωματικότητας μεταξύ των προϊόντων, αποδεικνύονται εξαιρετικά σταθερά και υψηλής πληροφοριακής αξίας στην πάροδο του χρόνου.

Προς επιβεβαίωση αυτού του ισχυρισμού, στο Σχήμα 5.2 παρατίθεται ενδεικτικά η συνεισφορά των 15 σημαντικότερων χαρακτηριστικών στο μοντέλο `LightGBM`, όπου αποτυπώνεται η αυξημένη βαρύτητα των διανυσμάτων της υβριδικής προσέγγισης.



Σχήμα 5.2: Συνεισφορά των embeddings στο μοντέλο εκπαίδευσης

Όσον αφορά την αξιολόγηση της ακρίβειας, τα αποτελέσματα καταδεικνύουν ότι το ημερήσιο σφάλμα ελαχιστοποιείται στον βραχυπρόθεσμο ορίζοντα. Ωστόσο, η συγκριτική ανάλυση των στρατηγικών ανέδειξε ένα σημαντικό εύρημα. Σε κάποιες περιπτώσεις, η άμεση πρόβλεψη πολλαπλών οριζόντων (*direct multi-horizon forecasting*) ξεπέρασε σε απόδοση την αντίστοιχη μέθοδο της αθροιστικής ημερήσιας πρόβλεψης. Χαρακτηριστικό παράδειγμα αποτελεί η υβριδική αρχιτεκτονική του Random Forests (`RF_Hybrid`), η οποία πέτυχε το εξαιρετικά χαμηλό ετήσιο αθροιστικό σφάλμα `WAPE` 1,7% (0,017) μέσω απ' ευθείας προβλέψεων κυλιόμενων 15 ημερών.

Η αναλυτική απόδοση των μοντέλων και των δύο προσεγγίσεων (βασικής και υβριδικής) παρουσιάζεται στους Πίνακες 5.3 έως 5.5, ενώ τα Σχήματα 5.3 και 5.4 αποτυπώνουν οπτικά την πορεία των αθροιστικών προβλέψεων. Αξίζει να σημειωθεί ότι οι σκιασμένες ζώνες (μπάντες) που απεικονίζονται στα συγκεκριμένα γραφήματα αντιπροσωπεύουν τα διαστήματα πρόβλεψης (*prediction intervals*) κάθε μοντέλου προσφέροντας μια σαφή γεωμετρική αποτύπωση των ορίων στατιστικής αβεβαιότητας και των περιθωρίων ρίσκου των εκτιμήσεων.

Πίνακας 5.3: Σφάλμα WAPE ανά μοντέλο εκπαίδευσης και ανά ορίζοντα

WAPE	**1d**	**15d**	**90d**
`XGB_Base`	0.4189	0.5987	0.7526
`XGB_Hybrid`	0.4326	0.6023	0.7605
`LGBM_Base`	0.4324	0.6010	0.7793

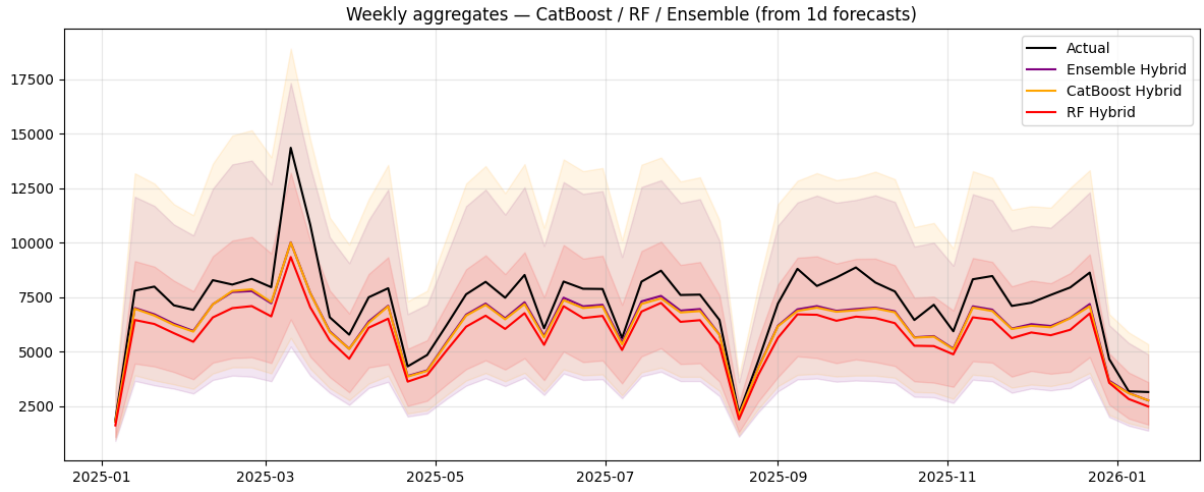
WAPE	1d	15d	90d
LGBM_Hybrid	0.4335	0.6213	0.7604
Cat_Base	0.4203	0.5429	0.5601
Cat_Hybrid	0.4194	0.5358	0.5509
RF_Base	0.4571	0.5293	0.5405
RF_Hybrid	0.4587	0.5342	0.5343

Πίνακας 5.4: Ετήσιο αθροιστικό σφάλμα WAPE ανά μοντέλο και ανά ορίζοντα

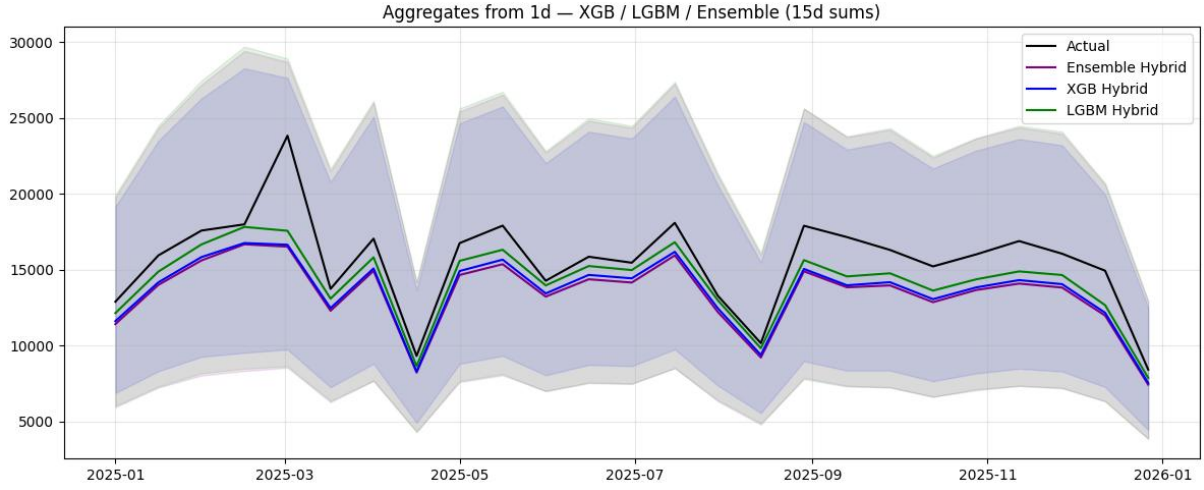
Ορίζοντας	Annual WAPE				
	XGB_base	LGBM_base	Cat_base	RF_base	Ensemble_base
1	0.111	0.086	0.144	0.197	0.133
15	0.032	0.180	0.057	**0.025**	0.058
90	0.107	0.214	0.195	0.029	0.086
Ορίζοντας	**Annual WAPE**				
	XGB_hybrid	LGBM_hybrid	Cat_hybrid	RF_hybrid	Ensemble_hybrid
1	0.126	0.087	0.144	0.199	0.138
15	0.055	0.184	0.061	**0.017**	0.067
90	0.078	0.081	0.191	0.046	0.102

Πίνακας 5.5: Αθροιστικές μετρικές σφάλματος ημερήσιων προβλέψεων εβδομαδιαίας και ετήσιας άθροισης

Ημερήσιες προβλέψεις	Εβδομαδιαία άθροιση (Weekly aggregation)			Ετήσια άθροιση (Annual aggregetion)		
Μοντέλο	**MAE**	**RMSE**	**WAPE**	**MAE**	**RMSE**	**WAPE**
XGB_base	827.10	1040.50	0.1118	42579.87	42579.87	0.1107
LGBM_base	648.05	887.13	0.0876	32879.43	32879.43	0.0855
Cat_base	1066.70	1249.78	0.1442	55362.52	55362.52	0.1440
RF_base	1454.18	1611.43	0.1966	75617.39	75617.39	0.1966
Ensemble_base	988.34	1179.56	0.1337	51240.80	51240.80	0.1333
XGB_hybrid	936.89	1140.52	0.1267	48484.12	48484.12	0.1261
LGBM_hybrid	658.45	897.10	0.0890	33480.65	33480.65	0.0871
Cat_hybrid	1069.42	1252.20	0.1446	55508.13	55508.13	0.1443
RF_hybrid	1472.55	1628.43	0.1991	76572.63	76572.63	0.1991
Ensemble_hybrid	1025.64	1213.79	0.1387	53206.40	53206.40	0.1384



Σχήμα 5.3: Διάγραμμα ημερήσιων αθροιστικών προβλέψεων σε εβδομαδιαία βάση των CatBoost, RandomForest και Ensemble



Σχήμα 5.4: Διάγραμμα ημερήσιων αθροιστικών προβλέψεων σε δεκαπενθήμερη βάση των XGBoost, LightGBM και Ensemble

5.4.7 Κριτική αποτίμηση και δομικοί περιορισμοί

Παρότι η ενοποιημένη ροή εκτέλεσης (*pipeline*) της Προσέγγισης 1 επέδειξε ικανοποιητική στατιστική συμπεριφορά σε επίπεδο αδρών, αθροιστικών προβλέψεων (με χαρακτηριστικό παράδειγμα την εξαιρετική επίδοση του υβριδικού Random Forests στον ορίζοντα των 15 ημερών), η βαθύτερη επιχειρησιακή και ανά-προϊόν ανάλυση αποκάλυψε σοβαρούς δομικούς περιορισμούς. Οι περιορισμοί αυτοί αποδείχθηκαν αδύνατο να θεραπευτούν μέσω οριακών, βηματικών βελτιώσεων (*incremental tuning*) και οδήγησαν, τελικά, στην απόφαση αναθεώρησης και εγκατάλειψης της συγκεκριμένης αρχιτεκτονικής.

Οι κύριοι λόγοι εγκατάλειψης, καθώς, και οι δομικοί περιορισμοί του συστήματος συνοψίζονται στους κάτωθι:

- **Απουσία μηχανισμού λογοκρισίας της ζήτησης (*demand censoring*):** Τα δενδρικά μοντέλα εκπαιδεύτηκαν τυφλά επί των ιστορικών, καταγεγραμμένων πωλήσεων (*observed sales*). Στην αγορά των ανταλλακτικών αυτοκινήτων, η προσέγγιση αυτή είναι προβληματική, διότι, όταν ένα προϊόν εξαντλείται, οι πωλήσεις μηδενίζονται. Όμως η πραγματική ζήτηση της αγοράς παραμένει θετική. Μην λαμβάνοντας υπόψη την εικονική αυτή έλλειψη ζήτησης κατά τη

διάρκεια περιόδων έλλειψης αποθεμάτων (*stock-outs*), το μοντέλο αφομοίωσε εσφαλμένα τον μηδενισμό των πωλήσεων ως έλλειψη ενδιαφέροντος και οδηγήθηκε σε συστηματική υπό-πρόβλεψη της μελλοντικής ζήτησης για τους πιο δυναμικούς και κρίσιμους κωδικούς.

- **Συστηματική χρονική μεροληψία και υπό-εκτίμηση (*Annual under-forecasting bias*):** Κατά τη μετάβαση από τις ημερήσιες κυλιόμενες προβλέψεις (1d) σε μακροπρόθεσμα, ετήσια χρονικά αθροίσματα (*annual aggregations*), εντοπίστηκε μια συστηματική αρνητική μεροληψία της τάξεως του $-13{,}3\%$. Το μέγεθος αυτό υπερβαίνει κατά πολύ το αυστηρό επιχειρησιακό όριο ανοχής, το οποίο ορίζεται παραδοσιακά σε αποκλίσεις χαμηλότερες του $< 5\%$, καθιστώντας τις ημερήσιες προβλέψεις ακατάλληλες για μακροπρόθεσμο στρατηγικό προγραμματισμό αγορών.
- **Αδυναμία επιχειρησιακής αξιοποίησης:** Παρότι η μεθοδολογία είναι ορθή από επιστημονικής και ερευνητικής άποψης, οι κυλιόμενες ημερήσιες προβλέψεις δεν έχουν πρακτική αξία σε ρεαλιστικές στρατηγικές αναπλήρωσης αποθεμάτων.
- **Δυσανάλογη πολυπλοκότητα GNN έναντι απόδοσης νέων προϊόντων (*cold-start*):** Μολονότι η θεωρητική σύλληψη του δικτύου MTL-GNN για την παραγωγή SKUs embeddings ήταν μαθηματικά υποδειγματική, η πρακτική της συνεισφορά στην κατηγορία των νέων προϊόντων (*cold-start*) δεν ταίριαξε σωστά με τον πολύ βραχύ ημερήσιο ορίζοντα προβλέψεων. Η μείωση του σφάλματος στα νέα προϊόντα ήταν οριακή και δεν στάθηκε ικανή να ισορροπήσει τον τεράστιο υπολογιστικό φόρτο, την αρχιτεκτονική πολυπλοκότητα και το υψηλό κόστος συντήρησης και επανεκπαίδευσης που απαιτεί η διατήρηση ενός νευρωνικού δικτύου γράφων σε παραγωγικό περιβάλλον.

Τα ανωτέρω κρίσιμα συμπεράσματα και οι εντοπισμένες αδυναμίες δεν ερμηνεύτηκαν ως αποτυχία, αλλά αποτέλεσαν τον απαραίτητο επιστημονικό θεμέλιο λίθο για τον ανασχεδιασμό του συστήματος. Η βαθιά κατανόηση των περιορισμών της Προσέγγισης 1 καθόρισε άμεσα τις προδιαγραφές για την ανάπτυξη της τελικής εξελικτικής φάσης της έρευνας (Προσέγγιση 3). Στην αρχιτεκτονική εκείνη, επιλέχθηκε η ολοκληρωτική μετατόπιση της φιλοσοφίας προς τη λογική των απευθείας προβλέψεων πολλαπλών οριζόντων (*direct multi-horizon forecasting*), η αντικατάσταση του GNN από αποδοτικότερα διανύσματα τύπου *transformer embeddings* και η εισαγωγή ενός συστήματος στατιστικής βαθμονόμησης ανά βαθμίδα τζίρου (*revenue-stratified calibration*).

Παρόλα αυτά, η αρχιτεκτονική GNN διατηρεί ένα μοναδικό πλεονέκτημα, την ικανότητα εκμάθησης τοπολογικών σχέσεων χωρίς ιστορικό ζήτησης. Ως εκ τούτου, η ανάπτυξή της δεν απορρίπτεται καθολικά, αλλά απομονώνεται ως η ενδεδειγμένη αυτόνομη λύση για το πρόβλημα της «ψυχρής εκκίνησης» (*cold-start*) νέων κωδικών αποτελώντας τη βάση για τη βασική μελλοντική επέκταση του συστήματος (βλ. κεφάλαιο 7).

5.5 Εβδομαδιαία πρόβλεψη δύο σταδίων με χρήση Γλωσσικών Μοντέλων (*Weekly* embeddings-*based two-stage forecasting*)

Η δεύτερη μεθοδολογική προσέγγιση της παρούσας έρευνας αποτυπώνει μια ριζική αρχιτεκτονική μετατόπιση σε σχέση με την Προσέγγιση 1. Αντί της αξιοποίησης αναπαραστάσεων βασισμένων σε γράφους (GNN) και κυλιόμενων ημερήσιων προβλέψεων, η Προσέγγιση 2 διερευνά την ικανότητα των διανυσμάτων κειμένου (*text* embeddings) να κωδικοποιήσουν τη σημασιολογική συγγένεια των προϊόντων. Παράλληλα, μεταφέρει την ανάλυση στο εβδομαδιαίο επίπεδο και υιοθετεί μια εξειδικευμένη αρχιτεκτονική δύο σταδίων (*two-stage* Hurdle), σχεδιασμένη αποκλειστικά για τη ρητή διαχείριση της ακραία διαλείπουσας (*intermittent*) ζήτησης.

5.5.1 Σκεπτικό και αλλαγή παραδείγματος

Η μετάβαση σε αυτή την αρχιτεκτονική υπαγορεύτηκε από μια βασική διαπίστωση της Προσέγγισης 1. Σε ένα περιβάλλον, στο οποίο η συντριπτική πλειονότητα των παρατηρήσεων αφορά μηδενική ζήτηση (το ποσοστό των εβδομάδων χωρίς ζήτηση στο σύνολο δοκιμής άγγιζε το 93,3%), τα μονολιθικά μοντέλα αδυνατούν να επιτύχουν υψηλή ακρίβεια σημειακής πρόβλεψης (*point forecasting*).

Επιπροσθέτως, οι αυστηρές επιχειρησιακές απαιτήσεις για διατήρηση του επιπέδου εξυπηρέτησης (*service level*) άνω του 98% κατέστησαν σαφές ότι ο στόχος δεν πρέπει να είναι απλώς η πρόβλεψη της μέσης ζήτησης, αλλά η **πιθανοτική εκτίμηση του αποθέματος ασφαλείας (*safety stock*)**. Στο πλαίσιο αυτό, η πρόβλεψη υποβιβάστηκε στρατηγικά από το ημερήσιο στο εβδομαδιαίο επίπεδο (με χρονικό ορίζοντα $h = 2$ εβδομάδων), ώστε να εξομαλυνθεί μέρος του στατιστικού θορύβου και να διευκολυνθεί η εξαγωγή μοτίβων (*temporal aggregation*).

5.5.2 Παραγωγή και συμπίεση διανυσμάτων κειμένου (*text* embeddings)

Η μετάβαση στις κατανεμημένες διανυσματικές αναπαραστάσεις (*distributed vector representations*), που είναι θεμελιωμένες από τους Mikolov *et al.* [160], επιτρέπει τη μετατροπή αραιών και ασύνδετων κατηγορικών μεταβλητών σε πυκνά διανύσματα. Με αυτόν τον τρόπο, κωδικοποιείται η εγγενής σημασιολογική και λειτουργική συγγένεια των ανταλλακτικών εντός ενός συνεχούς μαθηματικού χώρου.

Αντί της τοπολογικής πληροφορίας των ταυτόχρονων αγορών, στην Προσέγγιση 2 υιοθετήθηκε η υπόθεση ότι προϊόντα με παρόμοια τεχνικά χαρακτηριστικά (π.χ. φίλτρα λαδιού συγκεκριμένων διαστάσεων) παρουσιάζουν παρόμοια πρότυπα ζήτησης. Για να καταστεί δυνατή η αξιοποίηση αυτής της σημασιολογικής πληροφορίας από τα αλγοριθμικά μοντέλα, απαιτήθηκε ο μετασχηματισμός των περιγραφών και των μεταδεδομένων των κωδικών (SKUs) σε πολυδιάστατα, πυκνά διανύσματα (*dense* embeddings).

Η διαδικασία αυτή περιλάμβανε τη συνένωση (*concatenation*) των διαθέσιμων χαρακτηριστικών κάθε προϊόντος (όπως η κατηγορία, η περιγραφή και ο προμηθευτής) σε ένα ενιαίο κείμενο. Το παραγόμενο κείμενο τροφοδοτήθηκε στη συνέχεια σε σύγχρονα, προ-εκπαιδευμένα γλωσσικά μοντέλα αρχιτεκτονικής Transformers.

Για την παραγωγή των διανυσμάτων αξιολογήθηκαν δύο διαφορετικοί κωδικοποιητές τελευταίας γενιάς:

- Το τοπικά εκτελούμενο (*local deployment*) μοντέλο `bge-m3` της Ακαδημίας Τεχνητής Νοημοσύνης του Πεκίνου (BAAI), το οποίο επιλέχθηκε λόγω της κορυφαίας απόδοσής του στην πολυγλωσσική ανάκτηση πληροφορίας και την κωδικοποίηση σύντομων κειμένων **Error! Reference source not found.**.
- Το υπολογιστικού νέφους (*cloud-based*) μοντέλο `amazon.titan-embed-text-v2`, μέσω της υπηρεσίας διαχειριζόμενων θεμελιωδών μοντέλων AWS Bedrock [163].

Αμφότερα τα μοντέλα παρήγαγαν διανύσματα υψηλής διαστατικότητας (1024 διαστάσεων). Ωστόσο, η απευθείας εισαγωγή ενός τόσο τεράστιου αριθμού χαρακτηριστικών σε δενδρικούς αλγορίθμους (όπως το XGBoost) δημιουργεί τον κίνδυνο της «κατάρας της διαστατικότητας» (*curse of dimensionality*). Δηλαδή, ο αλγόριθμος αναγκάζεται να εξετάσει υπερβολικά πολλά, συχνά θορυβώδη ή αραιά (*sparse*) χαρακτηριστικά σε κάθε διάσπαση κόμβου, με αποτέλεσμα να αυξάνονται δραματικά ο υπολογιστικός χρόνος και ο κίνδυνος υπερπροσαρμογής (*overfitting*).

Για την επίλυση αυτού του προβλήματος, τα εξαγόμενα διανύσματα υπεβλήθησαν σε τεχνικές μείωσης διαστατικότητας μέσω Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis – PCA) [164]. Πρόκειται για την καθιερωμένη στατιστική μέθοδο γραμμικού μετασχηματισμού, η οποία επιτρέπει την προβολή των δεδομένων σε έναν χώρο χαμηλότερων διαστάσεων διατηρώντας τη μέγιστη δυνατή διακύμανση. Η διαδικασία αυτή εντόπισε τους άξονες της μεγαλύτερης διακύμανσης συμπιέζοντας επιτυχώς την κρίσιμη σημασιολογική πληροφορία σε έναν μικρότερο, πυκνό πίνακα χαρακτηριστικών (π.χ., 32 ή 64 διαστάσεων). Αυτός διατήρησε την ουσία των τεχνικών ομοιοτήτων των προϊόντων καθιστώντας, παράλληλα, εφικτή την αποδοτική εκπαίδευση των μεταγενέστερων μοντέλων.

5.5.3 Ταξινόμηση ζήτησης μέσω του πλαισίου Syntetos-Boylan

Για τη μαθηματική τυποποίηση της ακανόνιστης συμπεριφοράς των προϊόντων, εφαρμόστηκε το καθιερωμένο στη βιβλιογραφία πλαίσιο των Syntetos & Boylan [1], [5]. Κάθε προϊόν ταξινομήθηκε σε ένα από τα τέσσερα τεταρτημόρια διαλείπουσας ζήτησης (*Smooth, Erratic, Intermittent, Lumpy*), βάσει δύο παραμέτρων:

- Μέσο Διάστημα Μεταξύ Ζητήσεων (ADI), το οποίο υπολογίζει τη μέση χρονική απόσταση (σε εβδομάδες) μεταξύ δύο μη μηδενικών πωλήσεων.
- Τετραγωνικός Συντελεστής Μεταβλητότητας (CV^2) που αξιολογεί τη διασπορά των ποσοτήτων ζήτησης, όταν αυτή εμφανίζεται.

Βάσει των αυστηρών ορίων του πλαισίου ($ADI \geq 1{,}32$ και $CV^2 \geq 0{,}49$), η συντριπτική πλειοψηφία των κωδικών ταξινομήθηκε στην κατηγορία *Intermittent* (Διαλείπουσα), επιβεβαιώνοντας την ανάγκη για έναν εξειδικευμένο αλγόριθμο που να αντιμετωπίζει τα μηδενικά δείγματα ξεχωριστά από τα θετικά.

5.5.4 Αρχιτεκτονική του μοντέλου Εμποδίου δύο σταδίων (*two-stage* Hurdle *model*)

Για να λυθεί το πρόβλημα των ετερογενών δεδομένων (*zero-inflated data*), αναπτύχθηκε μια αρχιτεκτονική δύο σταδίων (*two-stage* Hurdle *model*), η οποία διαχωρίζει την απόφαση *για το αν θα υπάρξει ζήτηση* από την απόφαση *του πόση θα είναι αυτή η ζήτηση*:

- **Στάδιο 1 - Ταξινόμηση Ενεργοποίησης (Hurdle *classifier*):**
  Ένα μοντέλο XGBoost (`XGBClassifier`) εκπαιδεύτηκε ως δυαδικός ταξινομητής για να προβλέπει την πιθανότητα εκδήλωσης οποιασδήποτε θετικής ζήτησης, $\hat{\mathbb{P}}(Y > 0)$. Αντί της χρήσης του προεπιλεγμένου κατωφλίου (0,5), το κατώφλι απόφασης τ βελτιστοποιήθηκε δυναμικά μέσω της μεγιστοποίησης της μετρικής F1-score (βλ. ενότητα 5.3) στο σύνολο επικύρωσης.
- **Στάδιο 2 - Παλινδρόμηση ποσότητας (*Quantity regressor*):**
  Ένα δεύτερο μοντέλο XGBoost (`XGBRegressor`) εκπαιδεύτηκε **αποκλειστικά** στα δείγματα, στα οποία η ζήτηση ήταν αυστηρά θετική ($Y > 0$). Το μοντέλο αυτό χρησιμοποίησε τη συνάρτηση απώλειας `Tweedie` για να προβλέψει το αναμενόμενο μέγεθος της παραγγελίας, $\hat{Y}_{reg}$.
- **Τελική σύνθεση (*Hard thresholding*):**
  Η τελική πρόβλεψη του συστήματος $\hat{Y}$ προέκυψε από τον συνδυασμό των δύο σταδίων ακολουθώντας έναν αυστηρό κανόνα απόρριψης, κατά τον οποίο η ποσότητα λαμβάνεται υπόψη μόνο, εάν ο ταξινομητής εγκρίνει το ακόλουθο συμβάν:

$$\hat{Y} = \hat{Y}_{reg} \cdot \mathbb{1}_{\{\hat{\mathbb{P}} > \tau\}} \quad (5.18)$$

(όπου $\mathbb{1}_{\{\hat{\mathbb{P}} > \tau\}}$ είναι η συνάρτηση-δείκτης, η οποία λαμβάνει την τιμή 1, όταν η συνθήκη εντός της παρενθέσεως αληθεύει, και 0 σε αντίθετη περίπτωση).

5.5.5 Στρατηγική μετα-βαθμονόμησης πέντε σταδίων (*Five-stage post-calibration*)

Λόγω της συστηματικής τάσης των δενδρικών μοντέλων να παρουσιάζουν μεροληψία (*bias*), όταν εκπαιδεύονται σε δεδομένα με ακραία σποραδικότητα, η αρχιτεκτονική του pipeline δεν σταματά στην εξαγωγή των ακατέργαστων προβλέψεων του μοντέλου Hurdle. Αντιθέτως, εισάγει μια προηγμένη διαδικασία μετα-βαθμονόμησης (*post-calibration*), η οποία εκτελείται ιεραρχικά σε πέντε διαδοχικά στάδια με σκοπό τη σταδιακή ευθυγράμμιση των εκτιμήσεων με την πραγματική επιχειρησιακή κλίμακα:

- **Στάδιο 1 – Καθολική διόρθωση μεροληψίας (*Global bias correction*)**
  Υπολογίζεται το συνολικό σφάλμα μεροληψίας (*bias*) του μοντέλου επί του συνόλου επικύρωσης. Εφαρμόζεται ένας καθολικός, γραμμικός πολλαπλασιαστής (*scalar scaling factor*) σε όλες τις προβλέψεις, ώστε να εξισορροπηθεί η συνολική υπο-πρόβλεψη ή υπερ-πρόβλεψη σε μακροοικονομικό επίπεδο.
- **Στάδιο 2 – Δυναμική βελτιστοποίηση κατωφλίου (*Threshold optimization*)**
  Το κατώφλι ενεργοποίησης τ του δυαδικού ταξινομητή (`hurdle classifier`) αναπροσαρμόζεται και βελτιστοποιείται. Στόχος είναι η λεπτή ρύθμιση της ευαισθησίας του μοντέλου, ώστε να ελαχιστοποιηθούν τα ψευδώς αρνητικά δείγματα. Δηλαδή οι περιπτώσεις, κατά τις οποίες το μοντέλο προέβλεπε μηδενική ζήτηση, αλλά στην πραγματικότητα εκδηλώθηκε πώληση και, κατά συνέπεια, είναι πιθανή η πρόκληση ελλείψεων αποθέματος.
- **Στάδιο 3 – Ισοτονική παλινδρόμηση πιθανοτήτων (*Isotonic probability calibration*)**
  Οι ωμές πιθανότητες που επιστρέφει ο αλγόριθμος XGBoost συχνά δεν ανταποκρίνονται στις πραγματικές συχνότητες του φαινομένου. Σε αυτό το στάδιο επιστρατεύεται η μη παραμετρική μέθοδος της Ισοτονικής Παλινδρόμησης (Isotonic Regression), η οποία μετασχηματίζει τις εξόδους του ταξινομητή, ώστε να είναι στατιστικά βαθμονομημένες (*well-calibrated probabilities*).
- **Στάδιο 4 – Διόρθωση μεροληψίας ανά κατηγορία (*Per-category bias calibration*)**
  Καθώς κάθε προϊοντική ομάδα (π.χ., φίλτρα έναντι μπουζί) παρουσιάζει διαφορετική δυναμική, εφαρμόζονται εξειδικευμένοι παράγοντες διόρθωσης σφάλματος ανά κατηγορία. *(Σημείωση: Στο πλαίσιο της ερευνητικής διερεύνησης, το στάδιο αυτό υλοποιήθηκε με τη χρήση των test actuals ως input, προκειμένου να ελεγχθούν τα ανώτατα θεωρητικά όρια ακρίβειας του συστήματος).*
- **Στάδιο 5 – Ιεραρχική συμφιλίωση προβλέψεων (*Hierarchical forecast reconciliation*)**
  Το τελικό στάδιο εξασφαλίζει τη μαθηματική συνέπεια της ιεραρχίας των δεδομένων (*bottom-up reconciliation*). Οι προβλέψεις σε επίπεδο μεμονωμένου κωδικού (SKU) αναπροσαρμόζονται και «συμφιλιώνονται» έτσι, ώστε το άθροισμά τους να ταυτίζεται απόλυτα με τις αναβαθμισμένες, σταθερές προβλέψεις που παρήχθησαν σε επίπεδο προϊοντικής κατηγορίας.

Η σταδιακή επίδραση αυτών των πέντε επιπέδων επεξεργασίας στην τελική ακρίβεια του συστήματος αποτιμάται ποσοτικά στην επόμενη ενότητα.

5.5.6 Εμπειρική επικύρωση και ποσοτικά ευρήματα

Η εμπειρική επικύρωση της Προσέγγισης 2 ανέδειξε τη σημαντική απόκλιση μεταξύ της επιχειρησιακής στόχευσης και της πραγματικής στατιστικής απόδοσης. Η αξιολόγηση πραγματοποιήθηκε σε δύο επίπεδα, αρχικά σε επίπεδο μεμονωμένου κωδικού (SKU) και, έπειτα, σε επίπεδο αθροιστικής προϊοντικής κατηγορίας.

Στον Πίνακα 5.6 παρουσιάζεται η απόδοση του συστήματος σε επίπεδο κωδικού (SKU). Προκειμένου η αξιολόγηση να είναι αυστηρή, η προτεινόμενη αρχιτεκτονική δύο σταδίων αξιολογήθηκε σε δύο παραλλαγές («*hard*» και «*soft*») και συγκρίθηκε με δύο αφελή μοντέλα αναφοράς (*naive baselines*):

- ***Two-stage-hard*:** Η προσέγγιση «σκληρού κατωφλίου», όπου η πρόβλεψη ποσότητας επικυρώνεται μόνο, εάν η πιθανότητα εκδήλωσης ζήτησης υπερβεί το βελτιστοποιημένο κατώφλι ($\tau = 0{,}606$), αλλιώς μηδενίζεται.
- ***Two-stage-soft* ($P \times Q$):** Η πιθανοτική προσέγγιση, κατά την οποία η τελική εκτίμηση προκύπτει από το γινόμενο της αναμενόμενης ποσότητας επί την πιθανότητα εκδήλωσης του συμβάντος.
- ***Zero-forecast (baseline*):** Μια τετριμμένη στρατηγική που προβλέπει διαρκώς μηδενική ζήτηση. Σε περιβάλλοντα αραιών δεδομένων (*zero-inflated*), η συγκεκριμένη στρατηγική αποτελεί το κρισιμότερο τεστ ελέγχου, όσον αφορά την ικανότητα του μοντέλου να μαθαίνει ουσιαστικά μοτίβα ή, απλώς, μαντεύει το πλειοψηφικό γεγονός.
- ***Mean-forecast (baseline)*:** Ένα στατικό μοντέλο που προβλέπει σταθερά τον ιστορικό μέσο όρο ζήτησης του εκάστοτε κωδικού.

Όπως διαφαίνεται από τα αποτελέσματα του πίνακα, βάσει της μετρικής `sMAE`, η αρχιτεκτονική δύο σταδίων (τόσο στη «*hard*» όσο και στη «*soft*» μορφή της) υστέρησε σημαντικά έναντι του «*Zero-forecast*» μοντέλου. Η αδυναμία αυτή πιστοποιεί τη θεμελιώδη δυσκολία των text embeddings και του συγκεκριμένου μοντέλου Hurdle να συλλάβουν τις ισχυρές, σποραδικές αιχμές ζήτησης σε ατομικό επίπεδο κωδικού.

Πίνακας 5.6: Απόδοση μοντέλων σε επίπεδο κωδικού (SKU-level) στο σύνολο δοκιμής, πριν την εφαρμογή βαθμονόμησης

Μέθοδος	**`sMAE` (Accuracy)**	**`Hit-rate`**	**`Cond WAPE`**	**`MAE`**	**`RMSE`**
Two-stage-hard ($\tau = 0{,}606$)	0,1984	0,8904	0,832	0,216	1,465
Two-stage-soft **($P \times Q$)**	0,2023	0,067	0,810	0,215	1,451
Zero-forecast	**0,4641**	**0,9330**	1,000	0,144	1,575
Mean-forecast	0,000	0,067	0,657	0,785	1,677

Αντιθέτως, κατά τη μετάβαση σε αθροιστικό επίπεδο κατηγορίας (Πίνακας 5.7), το σύστημα κατέγραψε φαινομενικά εντυπωσιακή βελτίωση. Ωστόσο, η βελτίωση αυτή (ειδικά στα Στάδια 4 και 5) δεν προέκυψε από τη γενικευτική ικανότητα του μοντέλου, αλλά από τεχνητή βαθμονόμηση (*calibration*), η οποία αξιοποίησε έμμεσα τις πραγματικές τιμές του συνόλου δοκιμής (*test actuals*).

Πίνακας 5.7: Σταδιακή εξέλιξη του σφάλματος βαθμονόμησης σε επίπεδο κατηγορίας

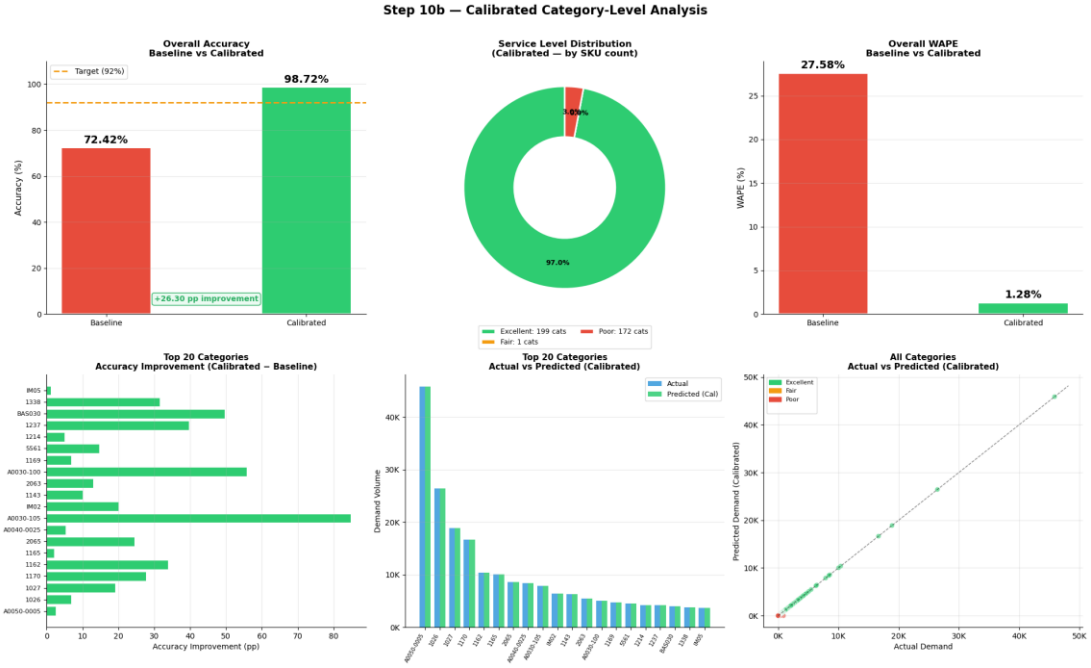
Στάδιο Βαθμονόμησης	**Σφάλμα WAPE**	**Ακρίβεια (Accuracy)**	**Μεταβολή ακρίβειας**
Μοντέλο βάσης *(Two-stage-hard)*	27,58	72,42%	—
Στάδιο 1 (*Global bias*)	29,23	70,77%	−1,65
Στάδιο 2 (*Threshold opt*)	25,64	74,36%	+1,94
Στάδιο 3 (*Isotonic*)	25,64	74,36%	+0,00
Στάδιο 4 (*Per-cat bias*)	4,47	95,53%	+23,11
Στάδιο 5 (*Reconciliation*)	**1,28**	**98,72%**	+26,30

Το μέγεθος αυτής της «στατιστικής οφθαλμαπάτης» αποτυπώνεται ανάγλυφα στον Πίνακα 5.8, όπου παρουσιάζονται ενδεικτικές κατηγορίες. Η βαθμονόμηση επέβαλε βίαια την εξίσωση των προβλέψεων με τις πραγματικές πωλήσεις αυξάνοντας τεχνητά την ακρίβεια έως και +84,78% σε κατηγορίες, στις οποίες αρχικό μοντέλο είχε αποτύχει πλήρως.

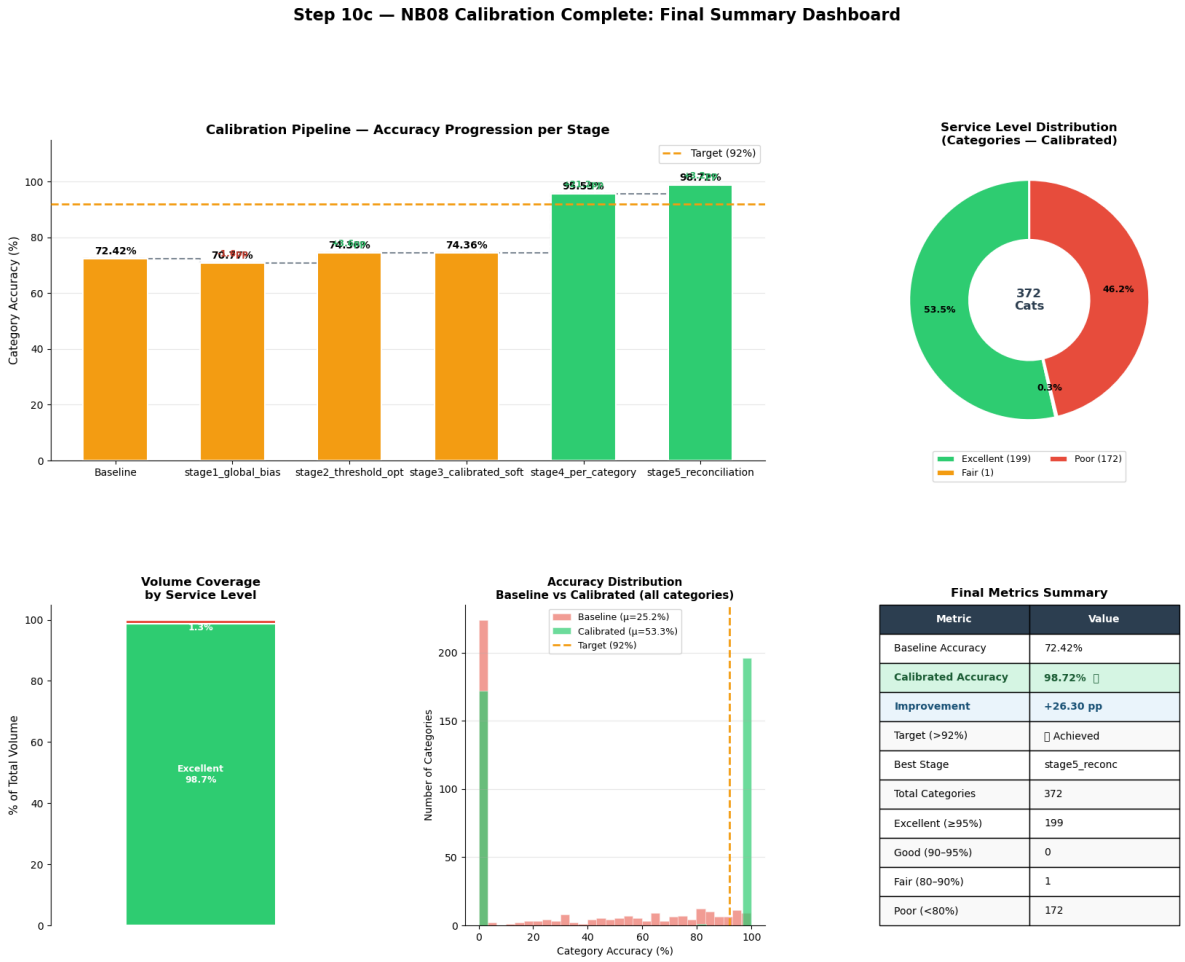
Πίνακας 5.8: Αντίκτυπος της βαθμονόμησης σε επιλεγμένες κατηγορίες υψηλού όγκου

Κωδικός Κατηγορίας	**Περιγραφή Κατηγορίας**	**Ακρίβεια μοντέλου βάσης**	**Ακρίβεια μετά τη βαθμονόμηση**	**Διαφορά (Δ)**
A0050-0005	ΑΝΑΦΛΕΚΤΗΡΕΣ (ΜΠΟΥΖΙ)	97,43%	100,00%	+2,57%
1026	ΦΙΛΤΡΑ ΛΑΔΙΟΥ	93,15%	100,00%	+6,85%
1170	ΦΙΛΤΡΑ ΑΕΡΟΣ ΚΑΜΠΙΝΑΣ	72,25%	100,00%	+27,75%
A0030-105	ΣΦΙΓΚΤΗΡΕΣ ΠΛΑΣΤΙΚΟΙ	15,22%	100,00%	**+84,78%**
A0030-100	ΣΦΙΓΚΤΗΡΕΣ ΜΕΤΑΛΛΙΚΟΙ	44,24%	100,00%	**+55,76%**
BAS030	ΒΑΣΕΙΣ ΑΜΟΡΤΙΣΕΡ ΕΜΠΡΟΣ	50,38%	100,00%	**+49,62%**

Στη συνέχεια παρατίθενται τα Σχήματα 5.5 και 5.6, τα οποία οπτικοποιούν την επίδραση της βαθμονόμησης ανά κατηγορία και τη συνολική απόδοση της ροής εργασιών, αντίστοιχα.



Σχήμα 5.5: Επίδραση της βαθμονόμησης ανά κατηγορία



Σχήμα 5.6: Συνολική απόδοση ροής εργασιών

5.5.7 Κριτική αποτίμηση, νόμος του Goodhart και λόγοι εγκατάλειψης

Συνοψίζοντας τα ανωτέρω ποσοτικά ευρήματα, η Προσέγγιση 2 τελικώς εγκαταλείφθηκε, καθότι εντοπίστηκαν τέσσερις κρίσιμοι περιορισμοί που δεν επέτρεπαν την παραγωγική της αξιοποίηση (*production deployment*):

- **Η Παγίδα της Βαθμονόμησης (Goodhart's Law)**
  Όπως καταγράφηκε στην εξέλιξη της βαθμονόμησης (Πίνακες 5.7 και 5.8), το εντυπωσιακό 98,72% δεν αποτελούσε προϊόν πραγματικής μάθησης, αλλά τεχνητής προσαρμογής. Οι πραγματικές τιμές χρησιμοποιήθηκαν έμμεσα ως παράγοντες βαθμονόμησης. Όπως επιτάσσει ο Νόμος του Goodhart, τον οποίο η σύγχρονη βιβλιογραφία συνοψίζει στο αξίωμα *«όταν μια μετρική γίνεται στόχος, παύει να είναι καλή μετρική»* [165], αυτή η πρακτική εξανέμισε την ικανότητα γενίκευσης του μοντέλου σε μελλοντικά, άγνωστα δεδομένα.
- **Αποτυχία σε Επίπεδο Κωδικού (SKU-*level failure*):** Η δραματική αστοχία σε επίπεδο SKU (όπου το μοντέλο "Zero-forecast" αποδείχθηκε ανώτερο) καταδεικνύει τη θεμελιώδη αδυναμία του Hurdle μοντέλου να συλλάβει εγκαίρως τις σποραδικές αιχμές ζήτησης, καθιστώντας το ακατάλληλο για αναπλήρωση αποθέματος κωδικό-προς-κωδικό.
- **Οριακή Προστιθέμενη Αξία των Διανυσμάτων Κειμένου:** Σε ένα περιβάλλον πλούσιο σε ισχυρά χρονικά χαρακτηριστικά, η συνεισφορά των *text* embeddings αποδείχθηκε δυσανάλογα μικρή. Η βελτίωση της ακρίβειας δεν δικαιολογούσε στο ελάχιστο την τεράστια υπολογιστική επιβάρυνση (εξαγωγή μέσω GPU) και την πολυπλοκότητα συντήρησης των μοντέλων τύπου Transformers. Το εύρημα αυτό ευθυγραμμίζεται πλήρως με τη σύγχρονη βιβλιογραφία [3].
- **Ανελαστικότητα Χρονικού Ορίζοντα:** Η αρχιτεκτονική παρέμεινε εγκλωβισμένη στη λογική των κυλιόμενων εβδομαδιαίων προβλέψεων, κάτι που αποδείχθηκε ασύμβατο με τις επιχειρησιακές ανάγκες για άμεση, ταυτόχρονη πρόβλεψη πολλαπλών οριζόντων (*direct multi-horizon forecasting*) έως και ένα έτος.

Το «σχολείο» της Προσέγγισης 2 κατέστησε σαφές ότι, τόσο η διαδικασία της βαθμονόμησης (calibration) όσο και η ίδια η αρχιτεκτονική μοντελοποίησης του συστήματος, χρειάζονταν ριζική αναθεώρηση. Αυτή η επίγνωση αποτέλεσε τον εννοιολογικό πυρήνα για την ανάπτυξη της Προσέγγισης 3, η οποία αποτελεί και την τελική, βέλτιστη λύση του προβλήματος.

5.6 Ροή εκτέλεσης Πραγματικής Πρόβλεψης Διαλείπουσας Ζήτησης (*Sparse True Forecast Pipeline*)

Στην παρούσα ενότητα παρουσιάζεται αναλυτικά η τρίτη και τελική προσέγγιση της έρευνας, υπό τον τίτλο «*Sparse True Forecast Pipeline*». Πρόκειται για μια ολοκληρωμένη ροή εργασιών (*pipeline*) η οποία, αξιοποιώντας προηγμένα μοντέλα Μηχανικής Μάθησης, παράγει πραγματικές προβλέψεις (*true forecasts*) σε πολλαπλούς χρονικούς ορίζοντες απευθείας από το χρονικό σημείο αποκοπής (*cutoff*) του συνόλου αξιολόγησης. Σε αντίθεση με τα προηγούμενα πειραματικά στάδια, το συγκεκριμένο σύστημα δεν βασίζεται σε κανενός είδους χρονικά κυλιόμενο μηχανισμό αναφοράς (*rolling mechanism*) επί του συνόλου δοκιμής, αλλά εξάγει άμεσες προβλέψεις (*direct multi-horizon forecasts*) για οκτώ διαφορετικούς ορίζοντες ταυτόχρονα. Επιπλέον, υποστηρίζει μια πλήρη, παραγωγικά έτοιμη (*production-ready*) ροή, από την εισαγωγή των ακατέργαστων συναλλαγών (*raw data*) έως τον τελικό υπολογισμό των αποθεμάτων ασφαλείας και των προτεινόμενων παραγγελιών αναπλήρωσης.

Είναι κρίσιμο να τονιστεί ότι η τελική αρχιτεκτονική δεν προέκυψε εμπειρικά ως οριακή βελτίωση κάποιου υπάρχοντος αλγορίθμου, αλλά σχεδιάστηκε εκ των προτέρων, με αποκλειστικό γνώμονα να

καλύψει τα ερευνητικά κενά που προσδιορίστηκαν στην Ενότητα 3.8 και να ευθυγραμμιστεί απόλυτα με τα δομικά χαρακτηριστικά των δεδομένων που αναλύθηκαν στο Κεφάλαιο 4.

Τα στρατηγικά κριτήρια και τα εμπειρικά ευρήματα που υπαγόρευσαν τις σχεδιαστικές επιλογές της Προσέγγισης 3 συνοψίζονται στους κάτωθι άξονες:

Υιοθέτηση αρχιτεκτονικής δύο σταδίων (*Two-stage* Hurdle *models*)

Η επιλογή μοντέλων Εμποδίου (Hurdle) δεν ήταν τυχαία, αλλά υπαγορεύθηκε από την επιτακτική ανάγκη διαχείρισης της ακραίας ανισορροπίας κλάσεων (*class imbalance*), η οποία προκαλείται από τον συντριπτικό όγκο μηδενικών παρατηρήσεων [70], [71]. Η σχεδιαστική αυτή επιλογή δικαιολογείται απόλυτα από τα στατιστικά ευρήματα του Κεφαλαίου 4, στο οποίο τεκμηριώθηκε ότι το 96,5% των κωδικών χαρακτηρίζεται από διαλείπουσα (*intermittent*) ζήτηση βάσει της κατηγοριοποίησης κατά Syntetos-Boylan (Ενότητα 4.6.1), ενώ το 76,9% των καταγεγραμμένων ημερήσιων συναλλαγών αφορά την πώληση ακριβώς μίας (1) μονάδας προϊόντος (Ενότητα 4.6.2). Αυτή η στατιστική πραγματικότητα μετατρέπει το πρόβλημα της ημερήσιας πρόβλεψης, πρωτίστως, σε δυαδικό ερώτημα. Διαχωρίζοντας τη διαδικασία σε δύο ανεξάρτητα στάδια, ένα μοντέλο ταξινόμησης για την εκτίμηση της πιθανότητας εμφάνισης ζήτησης $\mathbb{P}(y > 0)$ και ένα μοντέλο παλινδρόμησης αποκλειστικά για το υπό συνθήκη για το υπό συνθήκη μέγεθος αυτής, αποτρέπεται η συστηματική υπό-εκτίμηση (*under-forecasting bias*) των κλασικών αλγορίθμων, οι οποίοι τείνουν να προβλέπουν διαρκώς μηδέν για να ελαχιστοποιήσουν τη συνάρτηση απώλειας. Επισημαίνεται ότι η συγκεκριμένη σχεδιαστική επιλογή αναπτύχθηκε ανεξάρτητα, με γνώμονα τις ιδιαιτερότητες του παρόντος συνόλου δεδομένων, και επικυρώνεται πλήρως από τη σύγχρονη βιβλιογραφία του κλάδου [98], [111]. Η παράλληλη αυτή σύγκλιση υπογραμμίζει την ορθότητα της αρχιτεκτονικής εμποδίου δύο σταδίων ως τη βέλτιστη προσέγγιση για τη διαχείριση της ακραίας σποραδικότητας.

Κάλυψη της περιορισμένης βιβλιογραφίας στην αγορά ανταλλακτικών αυτοκινήτων

Το υπό ανάπτυξη σύστημα εφαρμόζεται σε πραγματικό σύνολο περίπου 72.000 κωδικών προϊόντων μιας ελληνικής εταιρείας χονδρικής εμπορίας ανταλλακτικών αυτοκινήτων. Με τον τρόπο αυτό, η μελέτη συνεισφέρει σε ένα πεδίο, στο οποίο οι ολοκληρωμένες, μεγάλης κλίμακας μελέτες παραμένουν σπάνιες [12], [13], [98].

Συστηματική ενσωμάτωση μηχανισμού «Λογοκρισίας της Ζήτησης» (*Demand Censoring*)

Η έντονη σποραδικότητα των δεδομένων συνδυάζεται με το φαινόμενο της «μακράς ουράς» (*long-tail*), όπου προϊόντα αργής κίνησης δεσμεύουν το 50% της αξίας της αποθήκης αποφέροντας μόλις το 20% του συνολικού τζίρου (βλ. ενότητα 4.5). Σε ένα τέτοιο περιβάλλον, οι παρατεταμένες ελλείψεις αποθεμάτων στα συγκεκριμένα είδη δεν συνιστούν παροδικό συμβάν, αλλά δομικό χαρακτηριστικό. Ως εκ τούτου, υλοποιείται ρητή ανακατασκευή των ιστορικών επιπέδων αποθέματος και αντιμετωπίζεται το φαινόμενο της λογοκριμένης ζήτησης (*demand censoring*) [7] μέσω ενός δυναμικού, μειωμένου συντελεστή βαρύτητας δειγμάτων (βλ. ενότητα 5.6.3) ακολουθώντας τις πλέον σύγχρονες ερευνητικές προτροπές [6], [8], [125].

Καθιέρωση ενιαίας αξιολόγησης με μικτά επιχειρησιακά κριτήρια (*Stratified and business-driven evaluation*)

Η δυσανάλογη συγκέντρωση του τζίρου, με συντελεστή Gini 0,742 (βλ. ενότητα 4.3), και η επιχειρησιακή κρισιμότητα των Top-20 οι οποίοι αποφέρουν 6,4–7,2% τζίρου καταλαμβάνοντας μόλις το 2,7% της αξίας αποθήκης (βλ. ενότητα 4.5), επιβάλλουν την αξιολόγηση σε στρώματα (*stratified evaluation*). Η αρχιτεκτονική εξάγει μετρικές απόδοσης (WAPE, Ratio, MASE, HitRate30) ταυτόχρονα

σε τρεις διακριτές βαθμίδες (Top-20, *Pareto-A*, l*ong-tail*), σύμφωνα με το πλαίσιο που ορίστηκε στην Ενότητα 5.2. Παράλληλα, η αξιολόγηση των μοντέλων επεκτείνεται στον υπολογισμό των αποθεμάτων ασφαλείας (*safety stocks*) και προτεινόμενων αναπληρώσεων παραγγελιών (βλ. ενότητα 5.6.12), ώστε η απόδοσή τους να αποτιμάται άμεσα και με όρους επιχειρησιακού αντικτύπου [3], [14].

Πλαίσιο πολλαπλών οριζόντων με μετα-μάθηση και βαθμονόμηση εσόδων (*Multi-horizon framework with meta-learning and revenue-stratified calibration*)

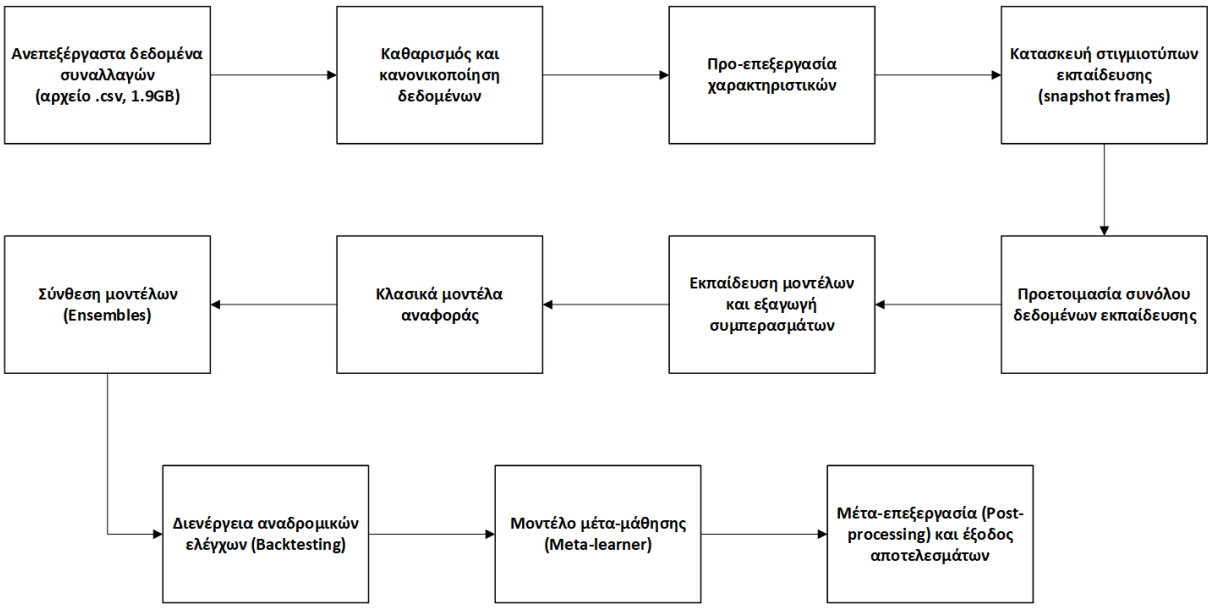
Η έλλειψη σταθερότητας στα είδη του καταλόγου (ετήσια εναλλαγή ~30% και, παράλληλα, 42,7% των κωδικών που εκπαιδεύονται είναι ανενεργοί κατά το 2025, βλ. ενότητα 4.8) καθιστά επισφαλή μια μεμονωμένη αποτίμηση στο σύνολο αξιολόγησης. Για τον λόγο αυτό, η Προσέγγιση 3 εφαρμόζει αναδρομικό έλεγχο πολλαπλών σημείων αποκοπής (*multi-cutoff backtesting*) με τη χρήση εξαμηνιαίων στιγμιότυπων (*snapshots*) (βλ. ενότητες 5.6.3 και 5.6.8). Τα εξαγόμενα αποτελέσματα τροφοδοτούνται σε έναν μηχανισμό στοίβαξης μετα-μοντέλου με περιορισμούς κυρτότητας (*constrained convex meta-learner stacking*) πάνω σε εκτός δείγματος (*out-of-sample*) προβλέψεις (βλ. ενότητα 5.6.9) και βαθμονόμηση σε επίπεδα βάσει εσόδων (*revenue-stratified calibration*). Η βαθμονόμηση αυτή προστατεύει τα κρίσιμα Top-N είδη, χωρίς να αλλοιώνει τον καθολικό λόγο (*global ratio*) μεταξύ πραγματικών πωλήσεων και προβλέψεων (Ενότητα 5.6.10). Ο συνδυασμός, που περιγράφεται ανωτέρω, απουσιάζει συστηματικά από τη βιβλιογραφία των ρεαλιστικών εφαρμογών στην αυτοκινητοβιομηχανία.

Κριτική αποτίμηση του Υπολογισμού Δεξαμενής (Reservoir Computing - RC):

Κατά τη διερευνητική φάση της Προσέγγισης 3, αξιολογήθηκε πειραματικά η ενσωμάτωση Δικτύων Κατάστασης Ηχούς (Echo State Networks - ESN) ως μηχανισμός εξαγωγής χαρακτηριστικών (*feature engineering*) για τα μοντέλα Hurdle. Παρότι η χρήση αρχιτεκτονικών Reservoir Computing έχει προταθεί πρόσφατα για τη διαχείριση σποραδικής ζήτησης [138], η εμπειρική μας διαπίστωση είναι ότι τα ESN εισάγουν πλεονάζοντα θόρυβο σε σύνολα δεδομένων με ακραία διαλείπουσα συμπεριφορά (Ενότητα 5.6.14). Αυτό αποτελεί ένα ιδιαίτερα κρίσιμο εύρημα, το οποίο ευθυγραμμίζεται με πρόσφατες έρευνες που αναδεικνύουν την υπεροχή μιας στοχευμένης μηχανικής εξαγωγής χαρακτηριστικών έναντι της αρχιτεκτονικής πολυπλοκότητας [98] θέτοντας σαφή πρακτικά όρια στη χρήση του RC στην αγορά των ανταλλακτικών.

5.6.1 Αρχιτεκτονική και ροή δεδομένων

Η συνολική ροή εργασιών (*pipeline*) περιέχει 10 διακριτά και διαδοχικά στάδια, τα οποία παρουσιάζονται στο Σχήμα 5.7.



Σχήμα 5.7: Ροή δεδομένων μοντέλου πραγματικής πρόβλεψης

Όλα τα στάδια του συστήματος ενεργοποιούνται μέσω μιας ενιαίας συνάρτησης `run_pipeline(cfg)`, από την οποία καλούνται όλες οι επιμέρους συναρτήσεις που είναι υπεύθυνες για κάθε εργασία. Το `cfg` αποτελεί το αρχείο ρυθμίσεων, στο οποίο ορίζονται όλες οι επιλεγμένες παράμετροι που ισχύουν κατά την εκτέλεση. Στο ακόλουθο πλαίσιο κώδικα παρατίθεται απόσπασμα με τις σημαντικότερες παραμέτρους του `cfg`. Το πλήρες αρχείο ρυθμίσεων συμπεριλαμβάνεται στο παράρτημα της παρούσης εργασίας.

```
cfg = PipelineConfig(
    # βασικές ρυθμίσεις pipeline
    data_path = Path('../data/ML_DATASET_20260107.csv'),
    forecast_origin = pd.Timestamp('2025-01-01'),
    horizons = (15, 30, 45, 60, 75, 90, 180, 365),

    # επιλογή μοντέλων
    Models = ("Hurdle_RF", "Hurdle_ET", "Hurdle_HGB", "Hurdle_CAT"),

    # μηχανισμοί εκπαίδευσης και βαθμονόμησης
    enable_calibration = True,
    load_backtest = True,
    backtest_max_cutoffs = 3
    enforce_backtest_cache_fingerprint = True
    snapshot_semi_annual = True,
    enable_stratified_scaling = True,
    enable_weighted_calibration = True,

    # feature engennering
    use_critical_rolling_features=True,
    use_seasonality_features = True
    use_holiday_flags = True
    use_value_features = True
    use_intermittent_features = True
    enforce_monotone_horizon_predictions = True #PAVA
    use_oos_features = True,
    use_transformer_embeddings = True,
    transformer_model_name = "sentence-transformers/paraphrase-multilingual-
                              MiniLM-L12-v2"
    transformer_embedding_batch_size = 128
    transformer_max_length = 128
    stock_feature_horizons = (15, 30, 45, 60, 75, 90),
    enable_demand_censoring = True,
    censoring_horizons = (15, 30, 45, 60, 75),
    enable_prediction_intervals = True

    # business decisions
    enable_safety_stock_export = True
    default_supplier_lead_time_days = 30.0
)
```

5.6.2 Καθαρισμός και κανονικοποίηση δεδομένων

Το αρχικό στάδιο του συστήματος αναλαμβάνει την ασφαλή εισαγωγή και προεπεξεργασία των ακατέργαστων δεδομένων (*raw data*). Η κεντρική συνάρτηση `load_and_normalize_dataset()` έχει σχεδιαστεί έτσι, ώστε να διαχειρίζεται αρχεία `CSV` με υψηλό βαθμό αυτοματοποίησης ενσωματώνοντας ευρετικούς ελέγχους (*heuristics*) για την ανίχνευση:

- **Διαχωριστικού χαρακτήρα (*Delimiter*):** Κόμμα, ελληνικό ερωτηματικό (`;`) ή στηλοθέτης (`tab`).
- **Κωδικοποίησης (*Encoding*):** Αυτόματη αναγνώριση `UTF-8`, `UTF-8-sig` (μέσω `BOM` detection), `UTF-16` ή `latin-1`.
- **Μορφότυπου Ημερομηνίας:** Αναγνώριση ακεραίων μορφής `YYYYMMDD`, προτύπου `ISO 8601`, ή ευρωπαϊκής διάταξης (`day-first`).
- **Αριθμητικής Αναπαράστασης:** Αυτόματη διαχείριση υποδιαστολών και διαχωριστικών χιλιάδων.

Η λογική δομή των πεδίων χαρτογραφείται δυναμικά μέσω της δομής δεδομένων `SCHEMA_CANDIDATES`. Πρόκειται για μια δομή δεδομένων τύπου λεξικό, το οποίο αντιστοιχίζει κάθε απαιτούμενο λογικό πεδίο (π.χ., `Product_Code`, `Date`, `Quantity`, `Transaction_Type`) σε λίστες πιθανών ονομασιών που ενδέχεται να εμφανιστούν στο αρχείο εισόδου. Η συγκεκριμένη αρχιτεκτονική αφαίρεση (*abstraction*) επιτρέπει στο *pipeline* να αφομοιώνει σύνολα δεδομένων από διαφορετικές πλατφόρμες διαχείρισης πόρων (ERP) χωρίς την ανάγκη συγγραφής ειδικά προσαρμοσμένου (*hard-coded*) κώδικα.

5.6.2.1 Διαχωρισμός και φιλτράρισμα τύπων συναλλαγής

Τα δεδομένα εισόδου περιέχουν πολλαπλούς τύπους συναλλαγών, οι οποίοι διαχωρίζονται αυστηρά βάσει του πεδίου `Transaction_Type`:

- `Type=1` (Πωλήσεις): Αποτελούν τα αποκλειστικά δεδομένα αναφοράς για την ιστορική ζήτηση.
- `Type=3` (Αναπληρώσεις / Εισαγωγές): Χρησιμοποιούνται αποκλειστικά για τη μαθηματική ανακατασκευή των ιστορικών επιπέδων αποθέματος.
- `Type=0` (Στιγμιότυπα Απογραφών): Λειτουργούν ως σημεία αγκύρωσης (*anchor points*) για τη διασφάλιση της ακρίβειας της ανακατασκευής του αποθέματος.

Για την εκπαίδευση των μοντέλων πρόβλεψης, στο στάδιο της ημερήσιας συσσωμάτωσης (`build_daily_aggregates`) προωθούνται αποκλειστικά πωλήσεις (`Type=1`) με θετική ποσότητα ($Quantity > 0$). Η στρατηγική αυτή επιλογή απορρίπτει αυτόματα τις επιστροφές προϊόντων και τις ακυρώσεις παραγγελιών, οι οποίες εισάγουν αρνητικές ποσότητες και θα αλλοίωναν στατιστικά το πραγματικό σήμα (signal) της αγοραστικής πρόθεσης.

5.6.2.2 Ημερήσια χρονική συσσωμάτωση (*Daily temporal aggregation*)

Για τη δημιουργία του συνόλου εκπαίδευσης, το σύστημα συμπτύσσει τις ενδοημερήσιες συναλλαγές. Εάν ένας κωδικός (SKU) εμφανίζει N πωλήσεις την ίδια ημερολογιακή ημέρα, αυτές συνενώνονται σε μία και μοναδική εγγραφή, αθροίζοντας την ποσότητα και την καθαρή αξία (`Net_Value`). Επίσης, τα στατικά μεταδεδομένα `Supplier_ID` (κωδικός προμηθευτή), `Product_Hierarchy_Level3_Code` (κατηγορία είδους) και `product_mastercode` (μητρικός κωδικός είδους) λαμβάνουν την πρώτη καταγεγραμμένη τιμή. Αυτό έχει ως αποτέλεσμα να δημιουργείται ένας δομημένος ημερήσιος πίνακας ζήτησης `daily`, στον οποίο μειώνονται οι τελικές εγγραφές στις 1.162.551 από τις 1.617.633 καταγεγραμμένες αρχικές πωλήσεις.

Απόσπασμα του εν λόγω κώδικα παρατίθεται ως ακολούθως:

```
sales.groupby(["Product_Code", "Date"], as_index=False).agg({
    "Quantity": "sum",                          # συνολική ποσότητα ημέρας
    "Net_Value": "sum",                         # συνολική καθαρή αξία ημέρας
    "Supplier_ID": "first",                     # στατικό πεδίο
    "Product_Hierarchy_Level3_Code": "first",   # στατικό πεδίο
    "Product_MasterCode": "first",              # στατικό πεδίο
})
```

5.6.2.3 Αραιή αναπαράσταση δεδομένων (*Sparse data representation*)

Μία από τις πλέον κρίσιμες αρχιτεκτονικές επιλογές του συστήματος είναι ότι ο πίνακας `daily` περιέχει αποκλειστικά και μόνο τα `(SKU, date)` ζεύγη με πραγματική δραστηριότητα πώλησης και όχι τον πλήρη καρτεσιανό χώρο `(SKU × ημερολογιακή ημέρα)`. Αυτό πηγάζει από στατιστικά ευρήματα της ενότητας 4.6, όπου αποδείχθηκε ότι το 96,5% των SKUs παρουσιάζει διαλείπουσα ζήτηση, με τη διάμεση τιμή του ποσοστού μηδενικής ζήτησης (*zero-demand fraction*) να προσεγγίζει το απόλυτο 1.0.. Εάν το σύστημα χρησιμοποιούσε παραδοσιακή, πλήρη (dense) αναπαράσταση, ο πίνακας θα διογκωνόταν σε περίπου 132 εκατομμύρια γραμμές (72.310 SKUs × 1.826 ημέρες), εκ των οποίων άνω του 99% θα περιείχαν απλώς την τιμή μηδέν. Υιοθετώντας την **αραιή (*sparse*) αναπαράσταση** (Saad, 2003), η πληροφορία συμπιέζεται σε 1,1 εκατομμύρια ουσιαστικές εγγραφές (μείωση κατά δύο τάξεις μεγέθους).

Αυτή η προσέγγιση εξασφαλίζει τρία θεμελιώδη πλεονεκτήματα:

- Επάρκεια μνήμης (*Memory efficiency*): Ολόκληρο το ιστορικό πωλήσεων της πενταετίας παραμένει στη μνήμη RAM εξαλείφοντας την ανάγκη διαρκών κλήσεων σε εξωτερική σχεσιακή βάση δεδομένων (I/O *bottlenecks*).
- Υπολογιστική ταχύτητα: Κατά τη δημιουργία των χαρακτηριστικών (όπως τα `Days_Since_Last_Sale` και `Mean_Interarrival_Days`), η αναζήτηση υλοποιείται μέσω της μεθόδου `np.searchsorted` επί ταξινομημένων πινάκων `datetime64[D]`. Η χρονική πολυπλοκότητα αυτής της λειτουργίας περιορίζεται μαθηματικά σε $\mathcal{O}(\log n)$ ανά SKU και ανά σημείο αποκοπής.
- Σχεδίαση *Snapshot-First*: Τα χρονικά χαρακτηριστικά δεν προ-υπολογίζονται για κάθε ημερολογιακή ημέρα (όπως συνέβαινε στις Προσεγγίσεις 1 και 2), αλλά εξάγονται «κατ' απαίτηση» (*on-demand*) αποκλειστικά στα επιλεγμένα χρονικά σημεία αποκοπής της εκπαίδευσης. Η αραιή είσοδος αποτελεί αυστηρό προαπαιτούμενο για την εφικτότητα αυτής της δυναμικής (*on-the-fly*) αρχιτεκτονικής.

5.6.2.4 Επεξεργασία συναλλαγών αναπλήρωσης και απογραφής

Κατ' αναλογία με τις πωλήσεις, η συνάρτηση `build_supply_histories()` εκτελεί την αντίστοιχη ομαδοποίηση για τα δεδομένα εφοδιασμού (`Transaction_Type ∈ {"0", "3"}`). Η λειτουργία ομαδοποιεί τις εγγραφές ανά (`Product_Code`, `Date`, `Transaction_Type`). Η διατήρηση του τύπου συναλλαγής ως κλειδί ομαδοποίησης (*grouping key*) είναι απολύτως κρίσιμη, γιατί διασφαλίζει ότι δεν θα χαθεί η σημασιολογική διαφορά μεταξύ μιας φυσικής απογραφής και μιας παραλαβής

εμπορευμάτων. Αυτή είναι πληροφορία που αποτελεί τον πυρήνα για την αξιόπιστη ανακατασκευή του αποθέματος και τον επακόλουθο υπολογισμό των δεικτών έλλειψης (*Out-Of-Stock* – `OOS` *–features*).

5.6.2.5 Κατασκευή ιστορικού ανά κωδικό είδους (*SKU histories and lifecycle management*)

Στο τελικό στάδιο προετοιμασίας, η συνάρτηση `build_sku_histories()` αναδιατάσσει τα ημερήσια αθροίσματα σε μια δομή λεξικού της μορφής: `histories[sku] → {'dates', 'qty', 'net_value', 'supplier', 'hier3', 'master'}`. Η μετατροπή αυτή είναι ουσιαστικά αλλαγή διάταξης και όχι περιεχομένου.

Οι ημερομηνίες αποθηκεύονται αυστηρά ως πίνακες `datetime64[D]`, οι αριθμητικές τιμές ως πίνακες `float32` (για βέλτιστη χρήση μνήμης), και τα στατικά πεδία ως αλφαριθμητικά (`str`). Η αναδιάταξη αυτή επιτρέπει στα επόμενα στάδια του pipeline να αντλούν παρελθοντικά διανύσματα ανά κωδικό, χωρίς να απαιτούνται χρονοβόρες κλήσεις της συνάρτησης `groupby()` επί του καθολικού `DataFrame` μειώνοντας τον χρόνο εκτέλεσης κατά μία τάξη μεγέθους.

Παράλληλα, στο στάδιο αυτό υπολογίζεται η ημερομηνία πρώτης πώλησης (`first_sale[sku]`). Η παράμετρος αυτή αποτρέπει το φαινόμενο της διαρροής πληροφορίας (*data leakage*), το οποίο είναι γνωστό ως «*eligible-but-unborn*». Επειδή η εκπαίδευση πραγματοποιείται αναδρομικά σε παρελθοντικά, εξαμηνιαία στιγμιότυπα (*snapshots*), το σύστημα πρέπει να γνωρίζει ποια προϊόντα βρίσκονταν όντως στον κατάλογο τη δεδομένη χρονική στιγμή αποκλείοντας δυναμικά SKUs που προστέθηκαν στην εταιρεία σε μεταγενέστερο χρόνο.

5.6.3 Κατασκευή παράγωγων χαρακτηριστικών (*Derived feature engineering*)

Πριν από την οριστική κατασκευή των ιστορικών πλαισίων εκπαίδευσης (*snapshots*) λαμβάνουν χώρα τρεις πρόσθετες, αλλά ιδιαίτερα κρίσιμες, προπαρασκευαστικές διαδικασίες. Οι διαδικασίες αυτές κατασκευάζουν τις παράγωγες δομές δεδομένων (*derived data structures*), οι οποίες στη συνέχεια τροφοδοτούν είτε τη μηχανική εξαγωγή των χαρακτηριστικών (*feature engineering*) είτε τον πολύπλοκο αλγόριθμο της λογοκρισίας ζήτησης (*demand censoring*).

Η εκτέλεσή τους ελέγχεται δυναμικά μέσω διακοπτών (*flags*) στο `PipelineConfig` (`use_oos_features`, `enable_demand_censoring`, `use_transformer_embeddings`). Προκειμένου να ελαχιστοποιηθεί ο υπολογιστικός φόρτος σε διαδοχικά πειράματα, τα αποτελέσματα της εκάστοτε διαδικασίας αποθηκεύονται τοπικά (*caching*) σε σειριοποιημένα αρχεία αποφεύγοντας τον εκ νέου υπολογισμό τους, εκτός εάν αλλοιωθούν τα αρχικά δεδομένα.

5.6.3.1 Σύνθεση ιστορικού αποθεμάτων (`build_supply_histories()`)

Η συγκεκριμένη συνάρτηση αναλαμβάνει να ομαδοποιήσει όλες τις συναλλαγές που αφορούν αναπληρώσεις εμπορευμάτων και απογραφές αποθήκης (`Transaction_Type ∈ {"3", "0"}`). Για κάθε προϊόν του καταλόγου παράγονται τέσσερις παράλληλοι πίνακες αριθμητικών δεδομένων (`dates`, `qty`, `net_value`, `type`). Η αναπαράσταση αυτών των δεδομένων μέσω πινάκων βιβλιοθήκης `NumPy` (αντί για αντικείμενα `pandas DataFrame` ή τυπικές λίστες της `Python`) αποτελεί στρατηγική επιλογή, καθώς επιτρέπει την εκτέλεση ταχύτατων πράξεων δυαδικού ελέγχου (μέσω της μεθόδου `np.searchsorted`) στα επόμενα στάδια της ροής.

Η έξοδος της συνάρτησης αξιοποιείται διττά:

- Τροφοδοτεί τη λειτουργία `_compute_oos_features()`, η οποία εξάγει έμμεσους δείκτες έλλειψης αποθέματος (*out-of-stock indicators*) απευθείας από τη συμπεριφορά των πωλήσεων.

Χαρακτηριστικά παραδείγματα αποτελούν οι δείκτες `Likely_OOS_Days_90d` και `Max_Sales_Gap_Ratio`, οι οποίοι εκτιμούν περιόδους έλλειψης μέσω αφύσικα μεγάλων χρονικών διαστημάτων (*sales gaps*) μεταξύ δύο επιτυχημένων πωλήσεων.

- Τροφοδοτεί τη συνάρτηση `reconstruct_stock_from_transactions()` ως η αποκλειστική πηγή γνώσης για τις ιστορικές παραλαβές. Σημειώνεται ότι το παρόν στάδιο δεν υπολογίζει τα τελικά επίπεδα αποθέματος, αλλά αρκείται στην αναδιοργάνωση του ιστορικού τροφοδοσίας σε βελτιστοποιημένη διάταξη ανά SKU.

5.6.3.2 Μαθηματική ανακατασκευή ημερήσιου αποθέματος (`reconstruct_stock_from_transactions()`)

Η ακριβής ανακατασκευή του ιστορικού επιπέδου αποθέματος υλοποιείται μέσω της λογικής παρακολούθησης της καθαρής ροής με τη χρήση σημείων αγκύρωσης (*net flow logic with an anchor point*). Πρόκειται για την καθιερωμένη βιομηχανική πρακτική σε περιβάλλοντα, όποτε απουσιάζει ένα πλήρες, αδιάλειπτο βιβλίο εισροών-εκροών αποθήκης [11]. Ο αλγόριθμος εκτελείται ανεξάρτητα για κάθε προϊόν και ακολουθεί τα εξής βήματα:

Βήμα 1: Συγκρότηση καθημερινών ροών. Για κάθε ημερολογιακή ημέρα, στην οποία εντοπίζεται δραστηριότητα, υπολογίζονται τρεις μεταβλητές: `sales` (συνολικές πωλήσεις), `replenish` (συνολικές παραλαβές), και `inventory` (το τελευταίο καταγεγραμμένο στιγμιότυπο απογραφής – αν δεν υπάρχει, λαμβάνει την τιμή `NaN`).

Βήμα 2: Υπολογισμός καθαρής ροής. Ορίζεται η ημερήσια καθαρή ροή ως `net_flow = replenish - sales` και υπολογίζεται το προοδευτικό της άθροισμα, `cum_flow = cumsum(net_flow)`. Η μεταβλητή `cum_flow` αντιπροσωπεύει τη συνολική θεωρητική μεταβολή του αποθέματος από την αρχή του ιστορικού πλαισίου.

Βήμα 3: Αγκύρωση σε στιγμιότυπο απογραφής. Εάν υπάρχει τουλάχιστον μία καταγεγραμμένη γραμμή φυσικής απογραφής (`Transaction_Type = "0"`), επιλέγεται αυστηρά η πιο πρόσφατη ως σημείο αγκύρωσης (*anchor point*). Η επιλογή της τελευταίας διαθέσιμης απογραφής (αντί της παλαιότερης) ελαχιστοποιεί το συσσωρευμένο στατιστικό σφάλμα (*drifting error*) κοντά στα κρίσιμα σημεία αποκοπής της εκπαίδευσης. Η εξίσωση της ανακατασκευής για τη χρονική στιγμή t ορίζεται ως:

$$Stock(t) = Anchor\_Qty + (Cum\_Flow(t) - Anchor\_Cum) \tag{5.19}$$

όπου:

- $Anchor\_Qty$: Η πιστοποιημένη φυσική ποσότητα που καταγράφηκε την ημέρα της απογραφής.
- $Cum\_Flow(t)$: Το αθροιστικό καθαρό αποτέλεσμα ροών (Παραλαβές $-$ Πωλήσεις) από την αρχή του ιστορικού έως τη στιγμή t.
- $Anchor\_Cum$: Το αθροιστικό καθαρό αποτέλεσμα ροών από την αρχή του ιστορικού έως την ημέρα που έγινε η απογραφή (*anchor day*).

Ο όρος εντός της παρενθέσεως αντιπροσωπεύει τη χρονική μετατόπιση (*offset*). Πρακτικά, η εξίσωση μεταφράζεται ως: «Εντοπίζεται η τελευταία επιβεβαιωμένη ποσότητα στην αποθήκη και προσαρμόζεται ανάλογα με τον ρυθμό αύξησης ή μείωσης της ροής των προϊόντων από εκείνη την ημέρα και έπειτα». Μέσω αυτής της μεθόδου, το εκτιμώμενο επίπεδο αποθέματος ταυτίζεται απόλυτα με την παρατηρούμενη πραγματικότητα την ημέρα της αγκύρωσης, ενώ επεκτείνεται αμφίδρομα (προς τα εμπρός και προς τα πίσω) βάσει της πιστοποιημένης ροής των συναλλαγών.

Βήμα 4: Εφεδρική προσέγγιση (χωρίς αγκύρωση). Σε περιπτώσεις όπου απουσιάζει παντελώς ιστορικό απογραφής, ο αλγόριθμος χρησιμοποιεί τη γραμμική σχέση $Stock(t) = Cum\_Flow(t)$. Είναι

μαθηματικά βέβαιο ότι, λόγω έλλειψης καταγεγραμμένων αρχικών αποθεμάτων, η σχέση αυτή θα παραγάγει αρνητικές τιμές. Το γεγονός αυτό **δεν διορθώνεται τεχνητά** (π.χ., θέτοντας ως ελάχιστο όριο το μηδέν). Μια έντονα αρνητική τιμή αποτελεί έναν ισχυρό, αξιόπιστο δείκτη ότι το πραγματικό απόθεμα του κωδικού ήταν εξαντλημένο για παρατεταμένες περιόδους, πληροφορία που αποτελεί το ακριβές σήμα που απαιτείται για την ενεργοποίηση της λογοκρισίας ζήτησης.

Ο αλγόριθμος ανακατασκευής αποθεμάτων, αν και απολύτως ντετερμινιστικός, είναι ιδιαίτερα απαιτητικός σε υπολογιστικούς πόρους. Ως εκ τούτου, η τελική δομή του λεξικού `stock_levels` αποθηκεύεται τοπικά στο αρχείο `stock_levels_cache.pkl` και φορτώνεται αυτόματα εκ νέου (μέσω της συνάρτησης `load_stock_levels()`) σε κάθε νέα εκτέλεση του *pipeline*.

5.6.3.3 Ενσωμάτωση στον κύκλο μηχανικής χαρακτηριστικών

Η παραγόμενη πληροφορία της ανακατασκευής αποθέματος τροφοδοτεί δύο διακριτές λειτουργίες στα επόμενα στάδια του συστήματος:

- Δημιουργία χαρακτηριστικών (*Feature generation*): Υπολογίζονται δυναμικά παράγοντες όπως το `Stock_At_Cutoff`, οι ημέρες από το τελευταίο θετικό απόθεμα (`Days_Since_Positive_Stock`), καθώς, και οι μέσοι όροι κλασματικής έλλειψης (`Stock_OOS_Frac_30d`, `Stock_OOS_Frac_90d`). Ο υπολογισμός τους, μέσω της συνάρτησης `np.searchsorted`, διατηρεί εξαιρετικά χαμηλή πολυπλοκότητα της τάξης $\mathcal{O}(\log n)$ ανά κωδικό προϊόντος και ανά σημείο αποκοπής.
- Μηχανισμός λογοκρισίας (*Censoring trigger*): Κατά τη δημιουργία του συνόλου εκπαίδευσης, οι περίοδοι οι οποίες πληρούν αυστηρά τη συνθήκη

$$(Stock\_At\_Cutoff \leq 0) \wedge (Target\_h = 0)(Stock\_At\_Cutoff \leq 0) \wedge (Target\_h = 0) \quad (5.20)$$

ταυτοποιούνται ρητά ως «τεκμηριωμένα λογοκριμένες». Τα συγκεκριμένα δείγματα δεν διαγράφονται (προκειμένου να μη χαθεί το μοτίβο του κωδικού), αλλά δέχονται εσκεμμένη υποβάθμιση της στατιστικής τους βαρύτητας λαμβάνοντας μειωμένο συντελεστή στο τελικό μοντέλο ($w\_oos = 0{,}8$).

5.6.3.4 Σημασιολογικές Ενσωματώσεις Μετασχηματιστή (`build_transformer_embeddings()`)

Η τρίτη διαδικασία του επιπέδου μηχανικής χαρακτηριστικών αποσκοπεί στην εξαγωγή πυκνών διανυσμάτων σημασιολογικής αναπαράστασης (*dense semantic* embeddings) 384 διαστάσεων ανά κωδικό προϊόντος (SKU) επιστρατεύοντας ένα προ-εκπαιδευμένο πολυγλωσσικό μοντέλο Μετασχηματιστή (Transformer). Η αρχιτεκτονική της συγκεκριμένης υλοποίησης αποκλίνει από την τυπική, έτοιμη χρήση της βιβλιοθήκης `sentence-transformers` και εξειδικεύεται στα ακόλουθα στάδια:

Σύνθεση Δομημένου Κειμένου Εισόδου

Για κάθε κωδικό s, η εσωτερική συνάρτηση `_compose_embedding_text()` κατασκευάζει μια τυποποιημένη συμβολοσειρά που ενοποιεί τα μεταδεδομένα του προϊόντος και της εφοδιαστικής αλυσίδας:

```
sku {Product_Code} | master {Product_MasterCode} | category {Hierarchy_Level3} |
supplier {Supplier_ID} | {Supplier_FullName} | {Supplier_Region} |
{Supplier_Country}
```

Τυχούσες ελλειπείς τιμές αντικαθίστανται αυτόματα από τη γενική ετικέτα αδράνειας `UNK`. Η σύμπτυξη της προϊοντικής ταυτότητας (`master`, `category`) με τη γεωγραφική και επιχειρησιακή ταυτότητα του προμηθευτή (`supplier`, `region`, `country`) σε ένα ενιαίο πλαίσιο κειμένου επιτρέπει στο μοντέλο να κωδικοποιήσει τις αλληλεξαρτήσεις των δύο αυτών διαστάσεων στον ίδιο διανυσματικό χώρο. Αυτή η τεχνική σύμπτυξης πολλαπλών χαρακτηριστικών σε μια ενιαία συμβολοσειρά εδράζεται στην αρχή της συνθετικότητας (*compositionality*) των διανυσματικών αναπαραστάσεων [160]. Όπως αποδείχθηκε από τη σχετική βιβλιογραφία, ο συνδυασμός διακριτών όρων (π.χ. κατηγορία προϊόντος και γεωγραφική προέλευση προμηθευτή) οδηγεί στη δημιουργία μιας νέας, ολιστικής αναπαράστασης που συλλαμβάνει τις μη-γραμμικές αλληλεπιδράσεις τους εντός του διανυσματικού χώρου. Κατά αυτόν τον τρόπο, δύο SKUs της ίδιας εμπορικής κατηγορίας που προέρχονται όμως από διαφορετικούς προμηθευτές καταλαμβάνουν διακριτές, αλλά μαθηματικά γειτονικές περιοχές στον χώρο των αναπαραστάσεων.

Κριτήρια επιλογής αρχιτεκτονικής

Χρησιμοποιείται το μοντέλο `sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2` [166] μέσω της κλάσης `transformers.AutoModel` του οικοσυστήματος Hugging Face. Η επιλογή αυτή υπαγορεύεται από τρεις παράγοντες:

- την εγγενή πολυγλωσσική υποστήριξη, η οποία είναι απαραίτητη για τη διαχείριση των ελληνικών περιγραφών,
- τις χαμηλές απαιτήσεις σε υπολογιστικούς πόρους λόγω μικρού μεγέθους (118 MB), οι οποίες προκύπτουν από την αρχιτεκτονική συμπίεσης MiniLM [151] και καθιστούν τη διαδικασία εξαγωγής συμπερασμάτων (*inference*) εφικτή σε κεντρική μονάδα επεξεργασίας (CPU) χωρίς επιχειρησιακές καθυστερήσεις,
- την αποδεδειγμένη σταθερότητα σε πολυγλωσσικές δοκιμές αξιολόγησης σημασιολογικής ομοιότητας (*STS benchmarks*) [167].

Λεκτικός τεμαχισμός και κωδικοποίηση (*Tokenization and encoding*)

Κάθε δέσμη (*batch*) αποτελούμενη από `transformer_embedding_batch_size = 128` κείμενα τεμαχίζεται λεκτικά και κωδικοποιείται με εφαρμογή τεχνικών εξίσωσης μεγέθους (`padding`) και αποκοπής (`truncation`) στο αυστηρό όριο των `transformer_max_length = 128` λεκτικών μονάδων (*tokens*). Κατά τη φάση των προκαταρκτικών πειραμάτων αξιολογήθηκε η αρχική διαμόρφωση των 64 tokens έναντι της διευρυμένης των 128 tokens. Τελικά, επιλέχθηκε η μεγαλύτερη εκδοχή λόγω της ικανότητας πλήρους κάλυψης του εμπλουτισμένου κειμένου χωρίς απώλεια πληροφορίας.

Μέση συγκέντρωση με μάσκα προσοχής (*Mean pooling with attention mask*)

Αντί για την απομονωμένη χρήση του token έναρξης (`[CLS]`), η τελική διανυσματική αναπαράσταση προκύπτει από τη μέση συγκέντρωση (*mean pooling*) των κρυφών καταστάσεων (*hidden states*) του τελευταίου επιπέδου, σταθμισμένη από τη μάσκα προσοχής (*attention mask*):

$$\mathbf{e}_s = \frac{\sum_t m_t \cdot \mathbf{h}_t}{\max(\sum_t m_t, \varepsilon)} \tag{5.20}$$

Όπου $\mathbf{h}_t$ αντιπροσωπεύει τις κρυφές καταστάσεις της τελευταίας στρώσης, m_t τη μάσκα προσοχής και $\varepsilon = 10^{-9}$ μια σταθερά αποφυγής μηδενικού παρονομαστή. Η συγκεκριμένη στρατηγική αποτελεί τη βέλτιστη πρακτική για αρχιτεκτονικές βασισμένες στο MiniLM (Reimers & Gurevych, 2019).

Κανονικοποίηση L2 (L2 *Normalization*)

Τα παραγόμενα διανύσματα υποβάλλονται σε κανονικοποίηση L2 ώστε να προβάλλονται στη μοναδιαία σφαίρα ($\|\mathbf{e}_s\|_2 = 1$). Η μαθηματική αυτή ιδιότητα διασφαλίζει ότι η συνημιτονική (*cosine*) και η ευκλείδεια απόσταση συμπίπτουν. Το γεγονός αυτό απλοποιεί σημαντικά τη διαδικασία διαχωρισμού των χαρακτηριστικών από τους δενδρικούς αλγορίθμους *boosting*.

Μηχανισμός προσκόμισης και προσωρινής αποθήκευσης (*Caching*)

Η έξοδος μετασχηματίζεται σε μια δομή χαρτογράφησης `feature_map[sku] → {Emb_000: float, …, Emb_383: float}` και αποθηκεύεται στο αρχείο `transformer_embeddings.pkl` συνοδευόμενη από τις μεταβλητές ελέγχου `model_name` και `prefix`. Η ακύρωση της κρυφής μνήμης (*cache invalidation*) ενεργοποιείται αυτόματα, όταν μεταβληθεί το όνομα του μοντέλου ή το πρόθεμα αποτρέποντας τη χρήση παρωχημένων διανυσματικών αναπαραστάσεων. Η εκτέλεση υποστηρίζει επιτάχυνση υλικού μέσω GPU (`cuda`) με αυτόματη μετάπτωση (*fallback*) σε CPU.

Λειτουργικός ρόλος

Τα embeddings ενσωματώνονται ως 384 πρόσθετες στήλες στα πλαίσια δεδομένων των στιγμιοτύπων (*snapshot frames*). Θεωρητικά, επιτελούν τρεις επιχειρησιακές λειτουργίες:

- ‘Ενα προϊόν με αραιό ιστορικό μπορεί να «δανειστεί» γνώση από τους πλησιέστερους γείτονες του στον διανυσματικό χώρο. Έτσι, μειώνεται η εξάρτησή του από χαρακτηριστικά υστέρησης (*lag features*) που στην προκειμένη περίπτωση είναι λιγοστά ή δεν υπάρχουν καθόλου.
- Αναπαριστούν συσχετίσεις κατηγορίας, προμηθευτών ή εναλλακτών ειδών παρόμοιο με αυτόν της Προσέγγισης 1, αλλά με λιγότερες απαιτήσεις πόρων και συντήρησης και χωρίς ανάγκη ρητής κατασκευής γράφου (GNN).
- Βελτίωση ολικής γενίκευσης. Η συσχέτιση μεταξύ ειδών της ίδιας οικογένειας αμβλύνει τη δυσκολία λήψης αποφάσεων των δενδρικών μοντέλων, ιδιαίτερα στους μακρινούς ορίζοντες, όπου η ίδια σειρά παρέχει ασθενές σήμα. Κατόπιν δοκιμών στο τρέχον σύστημα παρατηρήθηκε ότι η αφαίρεση των embeddings αύξησε το μέσο σφάλμα (WAPE) κατά 2–3 ποσοστιαίες μονάδες, με μεγαλύτερη υποβάθμιση στα είδη με περιορισμένο ιστορικό.

Αποτελέσματα Μελέτης Αφαίρεσης (Ablation Study)

Για την ποσοτικοποίηση της προστιθέμενης αξίας των κειμενικών διανυσματικών αναπαραστάσεων (embeddings) διεξήχθη μια εκτενής μελέτη αφαίρεσης (*ablation study*) τεσσάρων διαφορετικών διαμορφώσεων με αναδρομικό έλεγχο πολλαπλών σημείων αποκοπής (*multi-cutoff backtesting*). Τα αποτελέσματα επαληθεύτηκαν σε ανεξάρτητες εκτελέσεις και συνοψίζονται στον Πίνακα 5.9.

Πίνακας 5.9: Μετρικές σφάλματος και βαθμονόμησης βάσει της διαμόρφωσης των embeddings

Διαμόρφωση Χαρακτηριστικών	**Κείμενο Περιγραφής ανά SKU**	**15d Ratio**	**15d WAPE**	**15d MASE**	**30d MASE**	**90d MASE**	**180d Ratio**	**180d MASE**
`Simple embeddings` (Τελική)	`SKU + master + category + supplier + supplier_meta`	**0,994**	**1,076**	**0,882**	**0,906**	0,884	1,009	0,939
Απουσία embeddings	Χωρίς `transformer features`	1,143	1,167	0,957	0,922	**0,863**	**0,946**	**0,886**
`Full enrichment`	`Simple + Product_Name_EL,`	1,127	1,161	0,952	0,935	0,881	0,977	0,927

Διαμόρφωση Χαρακτηριστικών	Κείμενο Περιγραφής ανά SKU	15d Ratio	15d WAPE	15d MASE	30d MASE	90d MASE	180d Ratio	180d MASE
	`Manufacturer, ιεραρχία, CanBeInstalledTo`							
`Selective enrichment`	`Simple + Manufacturer, ιεραρχία μόνο`	1,134	1,140	0,954	0,943	0,890	0,986	0,932

Τα πειραματικά δεδομένα αναδεικνύουν σταθερά ευρήματα. Η διαμόρφωση `Simple embeddings` επιτυγχάνει σχεδόν τέλεια βαθμονόμηση ($Ratio = 0{,}994$) στον κρίσιμο βραχυπρόθεσμο ορίζοντα των 15 ημερών, καθώς, και σαφή υπεροχή ακρίβειας έναντι της πλήρους απουσίας των ($15d\ MASE$ $0{,}882$ έναντι $0{,}957$, βελτίωση $+8{,}5$ και $15d\ WAPE$ $1{,}076$ έναντι $1{,}167$, βελτίωση $+8{,}4\%$).

Η πλήρης απουσία των embeddings παρουσιάζει οριακά καλύτερο `MASE` στους μεσο-μακροπρόθεσμους ορίζοντες (60d–180d, μέση διαφορά $+0{,}018$), αλλά πάσχει από συστηματική και έντονη υποεκτίμηση της ζήτησης ($Ratio \in [0{,}94, 0{,}95]$ έναντι της επιθυμητής μονάδας). Το τελευταίο αποτελεί απαγορευτικό γεγονός για διαδικασίες αναπλήρωσης αποθεμάτων (*inventory replenishment*). Αντίθετα, η προσθήκη εκτενούς κειμενικού περιεχομένου (`Enrichment`) επιδεινώνει τόσο το `Ratio`, όσο και το `MASE` στις 15–30 ημέρες, χωρίς να προσφέρει κανένα όφελος στους μακρύτερους ορίζοντες.

Μεθοδολογική ερμηνεία και απόφαση

Η υπεροχή των βασικών διανυσματικών αναπαραστάσεων (`Simple embeddings`) έναντι της απουσίας τους ερμηνεύεται μέσω δύο κεντρικών μηχανισμών:

- **Κατάκτηση διακριτικής ταυτότητας SKU:** Τα δομικά πεδία του καταλόγου κωδικοποιούν σχέσεις, τις οποίες τα χαρακτηριστικά υστέρησης (*lag features*) αδυνατούν να συλλάβουν. Για παράδειγμα, δύο προϊόντα με ταυτόσημα μοτίβα ιστορικής ζήτησης, τα οποία όμως ανήκουν σε διαφορετικούς προμηθευτές, αποκτούν διακριτές διανυσματικές αναπαραστάσεις, αποτρέποντας τη συγχώνευσή τους από τον αλγόριθμο και βελτιώνοντας τη γενικευτική ικανότητα.
- **Σταθεροποίηση της βαθμονόμησης μάζας:** Τα embeddings περιορίζουν την αδιαφοροποίηση των προϊόντων της μακράς ουράς (*long-tail SKUs*) και παρέχουν ένα καθαρό συμπληρωματικό σήμα στον μηχανισμό βαθμονόμησης (*calibration scaler*) και επαναφέροντας τον δείκτη `Ratio` κοντά στη μονάδα.

Αντίθετα, η αποτυχία των σχημάτων εκτενούς εμπλουτισμού (`Full & Selective Enrichment`) αποδίδεται στους εξής παράγοντες:

- **Σημασιολογική ορθογωνιότητα προς τη ζήτηση:** Οι φυσικές ονομασίες των προϊόντων στα ελληνικά ή οι εκτενείς λίστες συμβατότητας οχημάτων (`CanBeInstalledTo`) δεν διαθέτουν προγνωστική ισχύ για το αν ένα εξάρτημα θα παρουσιάσει ζήτηση σε έναν σύντομο ορίζοντα.
- **Η κατάρα της διαστατικότητας (*Curse of dimensionality*):** Η εισαγωγή 384 διαστάσεων από κείμενα υψηλής μεταβλητότητας αυξάνει τη διαστατικότητα δυσανάλογα για κωδικούς της μακράς ουράς που διαθέτουν εξαιρετικά περιορισμένο σήμα εκπαίδευσης (λιγότερα από 10 γεγονότα ζήτησης στο ιστορικό τους).

- **Θόρυβος υψηλής πληθικότητας (*High cardinality noise*):** Πεδία με ακραία πληθικότητα, όπως οι ελεύθερες κειμενικές περιγραφές, εισάγουν θόρυβο, ο οποίος υπερκαλύπτει το ωφέλιμο σήμα της δομής του καταλόγου προϊόντων.

Μεθοδολογική απόφαση

Βάσει των ανωτέρω, η τελική αρχιτεκτονική της Προσέγγισης 3 εκτελείται με ενεργοποιημένες τις διανυσματικές αναπαραστάσεις (`use_transformer_embeddings = True`) και με απόλυτη απενεργοποίηση του εμπλουτισμού επιπλέον μεταδεδομένων (`PRODUCT_META_COLS = ()`). Η διαμόρφωση των `Simple embeddings` αναδεικνύεται ως η βέλτιστη λύση διασφαλίζοντας την ισορροπία ακρίβειας και βαθμονόμησης που απαιτείται για τις επιχειρησιακές αποφάσεις αναπλήρωσης.

5.6.4 Σχεδίαση χαρακτηριστικών των στιγμιοτύπων εκπαίδευσης (*Snapshot-based feature engineering*)

Η παρούσα ενότητα αποτελεί τον σχεδιαστικό «κορμό» της τρέχουσας προσέγγισης και εδραιώνει τη θεμελιώδη της διαφορά σε σχέση με τις Προσεγγίσεις 1 και 2. Η βασική φιλοσοφία του σχεδιασμού βασίζεται στην αρχή της **«πειθαρχίας διαρροής βάσει στιγμιοτύπων» (*snapshot-based leakage discipline*)**. Αντί να παράγονται χαρακτηριστικά (*features*) ανά ημερολογιακή ημέρα ή εβδομάδα, ο αλγόριθμος κατασκευάζει ένα ενιαίο διάνυσμα χαρακτηριστικών ανά προϊόν (SKU) και ανά αυστηρό χρονικό σημείο αποκοπής εκπαίδευσης (*training cutoff*). Βάσει αυτού του διανύσματος, το μοντέλο καλείται να προβλέψει το άθροισμα της ζήτησης για τις επόμενες h ημέρες, όπου h αντιπροσωπεύει τον εκάστοτε χρονικό ορίζοντα (όπως ορίστηκε στην Ενότητα 5.2).

5.6.4.1 Επιλογή σημείων αποκοπής εκπαίδευσης (Training cutoffs)

Η αρχιτεκτονική παρέχει την επιλογή για ετήσια ή εξαμηνιαία σημεία αποκοπής. Η κεντρική συνάρτηση `propose_training_cutoffs()` επιστρέφει μια ταξινομημένη λίστα από παρελθοντικά χρονικά σημεία, στα οποία κατασκευάζονται τα στιγμιότυπα εκπαίδευσης.

Κατά την προεπιλεγμένη ρύθμιση (`snapshot_semi_annual=True`), εξάγονται συνολικά έξι (6) σημεία αποκοπής: `01/01/2022`, `01/07/2022`, `01/01/2023`, `01/07/2023`, `01/01/2024` και `01/07/2024`. Σε αντίθετη περίπτωση, επιλέγονται μόνο τα τρία χρονικά σημεία κατά την έναρξη κάθε έτους. Προτιμάται η επιλογή των εξαμηνιαίων σημείων, καθώς διπλασιάζει τον όγκο των διαθέσιμων δεδομένων εκπαίδευσης. Παράλληλα, παρέχει στο μοντέλο μέτα-μάθησης (*meta-learner*) διπλάσια δεδομένα καταλοίπων εκπαίδευσης (*training residuals*) και, κατ' επέκταση, αντίστοιχα περισσότερα εκτός-δείγματος (*Out-Of-Sample* – OOS) παραδείγματα προς αξιολόγηση.

5.6.4.2 Ο κύριος αλγόριθμος `build_snapshot_frame()`

Για ένα προκαθορισμένο σημείο αποκοπής, η ανωτέρω συνάρτηση ανιχνεύει όλα τα είδη, τα οποία διαθέτουν ημερομηνία πρώτης αγοράς προγενέστερη του σημείου αυτού. Για κάθε ένα από τα επιλεγμένα προϊόντα υπολογίζονται εννέα διακριτές κατηγορίες χαρακτηριστικών (`features`), οι οποίες αναλύονται παρακάτω. Είναι κρίσιμο να τονιστεί ότι **όλα τα χαρακτηριστικά υπολογίζονται αυστηρά με δεδομένα πριν από το σημείο αποκοπής**, ενώ οι στόχοι (`targets`) υπολογίζονται αποκλειστικά με βάση την πραγματική ζήτηση στο μελλοντικό διάστημα $(cutoff,\ cutoff + h]$.

(A) Βασικά χαρακτηριστικά πρόσφατης δραστηριότητας και κύκλου ζωής

Πίνακας 5.10: Βασικά χαρακτηριστικά πρόσφατης δραστηριότητας και κύκλου ζωής

Όνομα	Περιγραφή
`Days_Since_Last_Sale`	Ημερολογιακές ημέρες από την τελευταία πώληση
`Lifetime_Days`	Εύρος από πρώτη έως τελευταία πώληση εντός του χρονικού ιστορικού που εξετάζεται
`Sales_Days_History`	Πλήθος διαφορετικών ημερών πώλησης στο ιστορικό
`Total_Qty_History`	Αθροιστική ιστορική ποσότητα
`Total_Value_History`	Αθροιστική ιστορική καθαρή αξία
`Mean_Qty_Per_Sale`, `Std_Qty_Per_Sale`, `CV_Qty_Per_Sale`	Στατιστικά ποσότητας ανά πώληση

(Β) Χαρακτηριστικά μεσοδιαστημάτων (*Inter-arrival features*)

Πίνακας 5.11: Χαρακτηριστικά μεσοδιαστημάτων

Όνομα	Περιγραφή
`Mean_Interarrival_Days`	Μέσος χρόνος μεταξύ διαδοχικών πωλήσεων
`P90_Interarrival_Days`	90ό εκατοστημόριο του μεσοδιαστήματος
`CV_Interarrival_Days`	Συντελεστής μεταβλητότητας του μεσοδιαστήματος

(Γ) Εκθετική εξασθένηση χαρακτηριστικών πρόσφατης δραστηριότητας (*Exponential smoothing*)

Πρόκειται για χαρακτηριστικά, τα οποία εφαρμόζουν μια εκθετική μείωση στην τιμή της πρόσφατης δραστηριότητας. Όσο πιο παλιά ήταν η τελευταία πώληση, τόσο μικρότερη είναι η τιμή αυτού του χαρακτηριστικού παρέχοντας στο μοντέλο μια ποσοτικοποιημένη αίσθηση της «ορμής» (*momentum*) του προϊόντος.

Για $\tau \in \{30, 60, 90, 180\}$ ημέρες, υπολογίζεται:

$$\text{Recency\_Decay}_\tau = \sum_{d \in \text{sales}} q_d \cdot \exp\left(-\frac{\text{cutoff} - d}{\tau}\right) \tag{5.21}$$

όπου

- $\text{Recency\_Decay}_\tau$: Το τελικό νούμερο που υπολογίζουμε. Είναι ένα «σταθμισμένο» άθροισμα πωλήσεων, ρυθμισμένο από την παράμετρο τ (π.χ. 30, 60, 90 ημέρες).
- q_d: Είναι η ποσότητα (`qty`) που πουλήθηκε τη συγκεκριμένη ημέρα d. Αντί να προσθέσουμε αυτή την ποσότητα αυτούσια, την πολλαπλασιάζουμε με το κομμάτι που ακολουθεί (την ποινή).
- $\text{cutoff} - d$: Αυτή είναι η «ηλικία» της πώλησης. Είναι η διαφορά σε ημέρες ανάμεσα στο χρονικό σημείο εκπαίδευσης (*cutoff*) και την ημέρα της πώλησης (d). Αν η πώληση έγινε χθες, η ηλικία είναι 1. Αν έγινε πριν ένα χρόνο, είναι 365.
- τ: Αυτή είναι η **μνήμη** του αλγορίθμου. Αν το τ είναι 30, τότε λέμε στο μοντέλο «θέλω να ξεχνάς γρήγορα» (βραχυπρόθεσμη τάση). Αν το τ είναι 180, του λέμε «κράτα μνήμη για το τι γινόταν το προηγούμενο εξάμηνο» (μακροπρόθεσμη τάση).

Αυτά τα χαρακτηριστικά δίνουν σταθμισμένη «σημερινή» εικόνα της ζήτησης, όπου πιο πρόσφατες πωλήσεις βαραίνουν περισσότερο. Το φαινόμενο exponential smoothing αποτελεί κλασική τεχνική χρονοσειρών [168], **Error! Reference source not found.**.

(Δ) Στατιστικά πολλαπλών κυλιόμενων παραθύρων (*Rolling windows*)

Για κάθε χρονικό παράθυρο $w \in \{30, 90, 120, 180, 270, 365, 540, 730\}$ ημέρες υπολογίζονται τα εξής:

- `QtySum_{w}d`, `QtyMean_{w}d`, `QtyStd_{w}d`, `QtyMax_{w}d` (Στατιστικά ποσοτήτων).
- `SalesDays_{w}d` (Πλήθος ημερών με πώληση εντός του παραθύρου).
- `ZeroFrac_{w}d` (Ποσοστό ημερών χωρίς πώληση).
- `ValueSum_{w}d` (Άθροισμα καθαρής αξίας).

Τα πολλαπλά παράθυρα παρέχουν στα μοντέλα μια πολυεπίπεδη προβολή της δυναμικής, αναδεικνύοντας από βραχυπρόθεσμες διακυμάνσεις (30) έως μακροπρόθεσμες τάσεις εποχικότητας (730).

(Ε) Κρίσιμα κυλιόμενα χαρακτηριστικά

Χρησιμοποιείται ένα αυστηρά επιλεγμένο σύνολο «κρίσιμων» χαρακτηριστικών, το οποίο υιοθετήθηκε από την Προσέγγιση 1 (βλ. ενότητα 5.4), όπως τα `Roll_Mean_14`, `Roll_Mean_30`, `Roll_Std_30`, `Momentum_30_vs_90`.

Ενεργοποιείται μέσω της επιλογής `use_critical_rolling_features=True`.

(ΣΤ) Χαρακτηριστικά εποχικότητας και σημαντικών ημερομηνιών (*Calendar flags*)

Πρόκειται για χαρακτηριστικά που περιλαμβάνουν τον μήνα, την ημέρα της εβδομάδας και κυκλικές μεταβλητές (`Sin/Cos`). Επίσης, χρησιμοποιούνται ειδικές σημαίες (`flags`) για περιόδους όπως τα Χριστούγεννα, το Πάσχα ή το καλοκαίρι.

- Κυκλικές συνιστώσες (*Fourier components*): ετήσιας (`Month_Sin_12`, `Month_Cos_12`) και διετούς (`Month_Sin_182`, `Month_Cos_182`) βάσης.
- Εποχικότητα ημερών εβδομάδας (*Weekday seasonality*): `Weekend_Frac_90d`, `Weekday_Peak_Ratio`.
- Σημαίες εορτών (*Holiday flags*): δείκτες ελληνικού Πάσχα, Χριστουγέννων, Δεκαπενταύγουστου (`HolFrac_*`).
- Μηνιαίοι δείκτες: Η απόκλιση του τελευταίου μήνα έναντι της ιστορικής μέσης μηνιαίας ζήτησης.

(Z) Χαρακτηριστικά αξίας (*Value features*)

Χρησιμοποιούνται δείκτες, όπως `Unit_Price`, `Value_Per_Day_365d`, `Value_Momentum_90_vs_365`, οι οποίοι αποτυπώνουν τη δυναμική της οικονομικής σημασίας του εκάστοτε προϊόντος και όχι μόνο της ποσότητας. Σημαντικά για την πρόβλεψη σε υψηλής κινητικότητας είδη, όπως τα Top-20 και αυτά της κλάσης 'A' (*Pareto-A*), όπως καθορίζονται στο κεφάλαιο 4.

(Η) Χαρακτηριστικά αραιής ζήτησης

Τα κυριότερα σημεία της βιβλιογραφίας σχετικά με την αραιή/διαλείπουσα ζήτηση επικεντρώνονται στα ακόλουθα χαρακτηριστικά:

- Μέσο Διάστημα Ζήτησης (*Average Demand Interval* - ADI): Ο λόγος της διάρκειας του κύκλου ζωής προς το πλήθος των ημερών με πώληση.
- CV^2 (*Squared Coefficient of Variation*): Το τετράγωνο του συντελεστή μεταβλητότητας της ποσότητας πώλησης, κρίσιμο μέγεθος για την ακρίβεια του σκέλους της παλινδρόμησης στο μοντέλο Hurdle.
- Παράμετροι Μοντέλου TSB (Teunter-Syntetos-Babai): Εξάγονται οι παράμετροι p_t (η εκθετικά εξομαλυνμένη πιθανότητα εκδήλωσης ζήτησης) και z_t (το εκθετικά εξομαλυνμένο μέσο μέγεθος ζήτησης). Η προσέγγιση TSB αποτελεί μια στατιστικά "πιο προσεκτική" εκδοχή της μεθόδου Croston, καθώς μειώνει δυναμικά τις προσδοκίες ζήτησης, όσο ένα προϊόν παραμένει ανενεργό.
- Κατηγοριοποίηση ζήτησης κατά Syntetos–Boylan (*Smooth / Erratic / Intermittent / Lumpy*), όπως αναφέρεται και στο κεφάλαιο 4.

Οι δείκτες διαλείπουσας ζήτησης βασίζονται στα θεμελιώδη έργα των Croston [4], Syntetos & Boylan [1], και Teunter, Syntetos & Babai [2]. Η ενσωμάτωσή τους ως χαρακτηριστικά επιτρέπει στα δενδρικά μοντέλα που εκπαιδεύονται, να εκμεταλλευτούν γνώση ειδική στο πρόβλημα της αραιής ζήτησης.

(Θ) Χαρακτηριστικά ειδών εκτός αποθέματος (*Out-Of-Stock* – OOS)

Όταν έχει επιλεγεί `use_oos_features=True` και έχει προηγηθεί ανακατασκευή αποθεμάτων, όπως περιγράφεται στην ενότητα 5.5.3, τότε στη συνάρτηση `build_snapshot_frame()` εξάγονται οι ακόλουθες στήλες ανά `SKU/cutoff` μέσω `np.searchsorted` επί της δομής `stock_levels[sku]`:

Πίνακας 5.12: OOS features

Όνομα	Περιγραφή	Πηγή
`Stock_At_Cutoff`	Επίπεδο αποθέματος στην ημέρα της αποκοπής (πιο πρόσφατη γνωστή τιμή)	Ανακατασκευή
`Stock_Is_Zero`	Δυαδική σημαία: (1, αν `Stock_At_Cutoff` ≤ 0)	Ανακατασκευή
`Days_Since_Positive_Stock`	Ημέρες από την τελευταία θετική απογραφή (`-1` αν ποτέ)	Ανακατασκευή
`Stock_Mean_{30,90}d`, `Stock_Min_{30,90}d`	Μέσος / ελάχιστος όρος αποθέματος σε πρόσφατα παράθυρα	Ανακατασκευή
`Stock_OOS_Frac_{30,90}d`	Ποσοστό ημερών με `Stock ≤ 0` στα παράθυρα 30/90 ημερών	Ανακατασκευή
`Likely_OOS_Days_90d`, `Max_Sales_Gap_Ratio`	Έμμεσοι δείκτες έλλειψης από αφύσικα μεγάλα διαστήματα μεταξύ πωλήσεων	Ιστορικά εφοδιασμού

Ενεργοποιούνται μόνο για βραχυ-μεσοπρόθεσμους ορίζοντες, `stock_feature_horizons=(15, 30, 45, 60, 75, 90)`. Εξαιρούνται από τη διαδικασία οι ορίζοντες 180 και 365 ημερών, καθώς η μακροπρόθεσμη πρόβλεψη είναι «αναίσθητη» στις τρέχουσες συνθήκες αποθέματος και, κατά κύριο λόγο, εισάγεται θόρυβος που υποβαθμίζει την ποιότητα της εκπαίδευσης των μοντέλων. Η τεκμηρίωση των δύο πηγών δεδομένων (συνένωση ιστορικών αποθεμάτων και ανακατασκευή ημερήσιου αποθέματος) δίνεται αναλυτικά στην ενότητα 5.5.3.

(Ι) Σημασιολογικές διανυσματικές αναπαραστάσεις (Transformers embeddings)

(Όπως αναλύθηκε εκτενώς στην Ενότητα 5.6.3.4.)

Οι 384 διαστάσεις των διανυσματικών αναπαραστάσεων προστίθενται στο διάνυσμα χαρακτηριστικών και λειτουργούν ως υποκατάστατο δομικών συσχετίσεων (*proxy correlations*) διευκολύνοντας την ψυχρή εκκίνηση (*cold-start*) και εξομαλύνοντας τον χώρο αποφάσεων των δενδρικών αλγορίθμων.

(ΙΑ) Μεταβλητές-Στόχοι (*Targets and ground truth*)

Μετά την παραγωγή του διανύσματος χαρακτηριστικών (*feature vector*) για το σύνολο των κωδικών (SKUs) σε μια δεδομένη ημερομηνία αποκοπής (*cutoff*), η συνάρτηση `build_snapshot_frame()` κατασκευάζει **μεταβλητές-στόχους** (`targets`) για κάθε ορίζοντα $h \in horizons = 15,30,45,60,75,90,180,365$ ημερών. Σε αντίθεση με τις Προσεγγίσεις 1 και 2, οι στόχοι αυτοί **δεν** αντιπροσωπεύουν την επιμέρους ζήτηση μιας συγκεκριμένης ημέρας ή εβδομάδας, αντίστοιχα, αλλά την **αθροιστική ζήτηση** εντός του μελλοντικού παραθύρου $(c, c + h]$:

$$\text{Target}^{(\text{h})}_{\text{sku, c}} = \sum_{d \in (c,\, c+h\text{d}]} q_{sku,\,d} \tag{5.22}$$

Όπου $q_{sku,d}$ είναι η ημερήσια καθαρή ποσότητα πώλησης του εκάστοτε προϊόντος. Στο επίπεδο της υλοποίησης, ο υπολογισμός εκτελείται με **λογαριθμική πολυπλοκότητα** $\mathcal{O}(\log n)$ ανά SKU ανά σημείο αποκοπής μέσω της συνάρτησης `np.searchsorted` πάνω στον **ταξινομημένο πίνακα** των ημερομηνιών πώλησης. Συγκεκριμένα, εντοπίζονται οι δείκτες `hist_end` (τέλος ιστορικού) και `future_end` (τέλος παραθύρου πρόβλεψης) που οριοθετούν το διάστημα και, ως εκ τούτου, το τελικό άθροισμα προκύπτει ως **άθροισμα υποσυνόλου** (*slice sum*) πάνω στον πίνακα ποσοτήτων `qty`.

Η ίδια διαδικασία εφαρμόζεται και για την καθαρή αξία των πωλήσεων, παράγοντας τη μεταβλητή `TargetValue_{h}d`. Αυτή η έκδοση βάσει εσόδων του στόχου χρησιμοποιείται στη σταθμισμένη βαθμονόμηση (*weighted calibration*) για την απόδοση μεγαλύτερης βαρύτητας στα προϊόντα με την υψηλότερη οικονομική σημασία.

Για κάθε ορίζοντα παράγονται 3 αποτελέσματα, όπως φαίνονται στον επόμενο πίνακα:

Πίνακας 5.13: Μεταβλητές – Στόχοι

Στήλη	Περιγραφή	Χρήση
`Target_{h}d`	Αθροιστική ποσότητα ζήτησης στο $(c, c + h]$	Στόχος του **παλινδρομητή** (*regressor*) σε κλίμακα `log1p`
`Active_{h}d`	Δυαδικός δείκτης: `` `1` `` αν `Target_{h}d` > 0, αλλιώς `` `0` ``	Στόχος του **ταξινομητή** (*classifier*)
`TargetValue_{h}d`	Αθροιστική καθαρή αξία στο παράθυρο $(c, c + h]$	Σκορ για τη **σταθμισμένη βαθμονόμηση** (βλ. ενότητα 5.6.10)

Η έξοδος αυτή αντικατοπτρίζει άμεσα την αρχιτεκτονική ***two-stage*** **Hurdle** (βλ. ενότητα 5.6.5). Ο ταξινομητής εκπαιδεύεται στο `Active_{h}d`, για να προβλέψει την πιθανότητα ύπαρξης ζήτησης, ενώ ο παλινδρομητής εκπαιδεύεται στο `Target_{h}d` περιοριζόμενος αποκλειστικά στο υποσύνολο των δεδομένων, όπου η ζήτηση είναι ενεργή (`Active_{h}d = 1`). Αυτή η διχοτόμηση αποφεύγει τον άμεσο υπολογισμό πιθανοτήτων κατά την εκτέλεση απλοποιώντας τη διάταξη μνήμης (*memory layout*) των δενδρικών μοντέλων.

Ιδιότητες των αθροιστικών στόχων

Η επιλογή της αθροιστικής (*cumulative*) αντί της περιθωριακής διατύπωσης έχει τρεις άμεσες συνέπειες:

- **Μονοτονία κατά ορίζοντα:** Εξ' ορισμού ισχύει ότι η ζήτηση ενός μικρότερου ορίζοντα είναι πάντα μικρότερη ή ίση από εκείνη ενός μεγαλύτερου (`Target_{h1}d` ≤ `Target_{h2}d`) για τον ίδιο κωδικό είδους και κοινή ημερομηνία αποκοπής. Αυτή η δομική ιδιότητα αξιοποιείται μέσω του αλγορίθμου PAVA στην `_enforce_monotone_cumulative_predictions()`, ώστε οι προβλέψεις από ανεξάρτητα μοντέλα να παραμένουν λογικά συνεπείς μεταξύ τους (βλ. ενότητα 5.6.7).
- **Ανισοτιμία των μηδενικών:** Σε μια αθροιστική δομή, το ποσοστό των μηδενικών τιμών μειώνεται, όσο μεγαλώνει ο ορίζοντας h. Συγκεκριμένα παρατηρείται ότι στο τρέχον σύνολο δεδομένων μειώνεται από ∼93% στις 15 ημέρες σε $\sim 60-65\%$ στις 365 ημέρες. Έτσι, ενώ ο ταξινομητής παραμένει κρίσιμος, το σχετικό βάρος της πρόβλεψης μετατοπίζεται σταδιακά προς τον παλινδρομητή για τους μεγαλύτερους ορίζοντες.
- **Άμεση συμβατότητα πολλαπλών οριζόντων:** Κάθε ορίζοντας εκπαιδεύεται ως αυτόνομο μοντέλο χωρίς αναδρομικές προβλέψεις. Αυτό εξαλείφει τη **διάδοση σφάλματος** (*error propagation*) που εμφανίζουν τα αναδρομικά μοντέλα και επιτρέπει στο μοντέλο μετα-μάθησης (*meta-learner*) να βελτιστοποιεί ανεξάρτητα τους συνδυασμούς για κάθε ορίζοντα.

Μεταδεδομένα για ιχνηλασιμότητα

Στο πλαίσιο των δεδομένων προστίθεται η στήλη `Snapshot_Cutoff` (η χρονική στιγμή της αποκοπής), η οποία επιτελεί διπλό ρόλο:

- Λειτουργεί ως φίλτρο ορίων (*per-horizon boundary filter*) για τον αποκλεισμό γραμμών εκπαίδευσης που επεκτείνονται πέρα από την αρχή της πρόβλεψης.
- Επιτρέπει στον μηχανισμό στάθμισης πρόσφατης δραστηριότητας (*recency weighting*) να δώσει μεγαλύτερη βαρύτητα στα πιο πρόσφατα χρονικά σημεία μέσω μιας εκθετικά φθίνουσας συνάρτησης (βλ. ενότητα 5.6.5).

Ρόλος των στόχων ως πραγματικά δεδομένα αναφοράς (*ground truth*)

Στο σημείο έναρξης της πρόβλεψης (01/01/2025), οι στόχοι υπολογίζονται κανονικά βάσει της πραγματικής ζήτησης του 2025, η οποία περιέχεται στο σύνολο δεδομένων αξιολόγησης (2025). Ωστόσο, υπάρχει μια κρίσιμη μεθοδολογική διάκριση. **Αυτές οι στήλες δεν τροφοδοτούνται στα μοντέλα ως σήμα εκπαίδευσης**. Είναι **δομικά αόρατες** για τα μοντέλα και ο μοναδικός τους ρόλος είναι να χρησιμεύσουν ως **πραγματικά δεδομένα αναφοράς** (*ground truth*) για τον υπολογισμό των μετρικών αξιολόγησης (`WAPE`, `Ratio`, κ.λπ.) και τη σύγκριση «Πραγματικών έναντι Προβλεπόμενων» τιμών. Επισημαίνεται ότι σε ένα πραγματικό περιβάλλον παραγωγής, οι στήλες αυτές απλώς θα ήταν κενές, καθώς η λειτουργία του συστήματος δεν εξαρτάται από αυτές για την έκδοση των προβλέψεων.

5.6.4.3 Συνολική διάσταση διανύσματος χαρακτηριστικών (*feature vector*)

Κατόπιν της συνένωσης, ο τελικός πίνακας χαρακτηριστικών ανά κωδικό αποτελείται από:

- ∼120 βασικά και κυλιόμενα χαρακτηριστικά,
- 384 transformers embeddings (όταν είναι ενεργά),
- 3 κατηγορικά πεδία (`Supplier_ID`, `Product_Hierarchy_Level3_Code`, `Product_MasterCode`) κωδικοποιημένα ως ακέραιοι,
- 8 στήλες τιμών-στόχων (`Target_15d`, ..., `Target_365d`) για την εκπαίδευση.

Το στιγμιότυπο συμπερασματολογίας (*inference snapshot*) στο σημείο αναφοράς προβλέψεων (`forecast_origin`), το οποίο τοποθετείται στις 01/01/2025, περιέχει ∼65.000 είδη. Κάθε στιγμιότυπο εκπαίδευσης παράγει 55.000–72.000 γραμμές ειδών ανάλογα με την ιστορική κάλυψη.

5.6.4.4 Λογοκρισία ζήτησης (*Demand censoring*) και Βάρη Γραμμών `OOS` (`Out-Of-Stock`)

Ένα από τα πιο κρίσιμα προβλήματα στα δεδομένα πραγματικών επιχειρήσεων είναι η **λογοκριμένη ζήτησης (*censored demand*)**. Όταν ένας κωδικός (SKU) βρίσκεται σε κατάσταση έλλειψης αποθέματος (*out-of-stock*), η καταγεγραμμένη ζήτηση για εκείνη την περίοδο εμφανίζεται τεχνητά μηδενική. Εάν το μοντέλο εκπαιδευτεί απευθείας σε αυτές τις γραμμές, θα «μάθει» να υποεκτιμά τη ζήτηση για προϊόντα που παρουσιάζουν συχνά ελλείψεις. Σε ένα σύστημα διαχείρισης αποθεμάτων, αυτό δημιουργεί έναν φαύλο κύκλο. Δηλαδή, η χαμηλή πρόβλεψη οδηγεί σε μικρή παραγγελία, η οποία προκαλεί νέα έλλειψη, καταλήγοντας σε ακόμη χαμηλότερη καταγεγραμμένη ζήτηση. Η μεροληψία αυτή έχει αναγνωριστεί ως από τους κυριότερους λόγους αποτυχίας ML μοντέλων σε πραγματικές συνθήκες [6], [7] και αποτελεί ρητό ερευνητικό κενό της βιβλιογραφίας (βλ. ενότητα 3.8). Οι Svetunkov και Sroginis [8] και Andrade και Cunha [125] αναδεικνύουν την ανάγκη για προσεγγίσεις που διαχωρίζουν τα δομικά μηδενικά (αραιή φύση ζήτησης) από τα τεχνητά μηδενικά (απουσία αποθέματος), παρόλα αυτά η πρακτική ενσωμάτωση `OOS` *flags* σε ολοκληρωμένα συστήματα μηχανικής μάθησης (ML *pipelines*) παραμένει σπάνια.

Η παρούσα εργασία αντιμετωπίζει το ζήτημα αυτό συνδέοντας τη μεθοδολογική χρήση σημαιών `OOS` **(**`out-of-stock flags`**)** και **βαρών δείγματος (*sample weights*)** στα μοντέλα Hurdle. Η διαχείριση υλοποιείται ως εξής:

- **Εντοπισμός λογοκριμένων (*censored*) δεδομένων**: Εντοπίζονται οι γραμμές εκπαίδευσης, κατά τις οποίες το απόθεμα κατά την αποκοπή είναι μηδενικό ή αρνητικό (`Stock_At_Cutoff`≤ 0) και, ταυτόχρονα, η μελλοντική ζήτηση είναι μηδενική (`Target_{h}d`$== 0$). Το γεγονός αυτό υποδεικνύει αδιάσειστη ένδειξη λογοκρισίας ζήτησης.
- **Απόδοση μειωμένου βάρου**ς: Αντί για την πλήρη απόρριψη αυτών των γραμμών, αποδίδεται ένα μειωμένο βάρος δείγματος $\boldsymbol{w_{oos} = 0,8}$ (αντί της προεπιλεγμένης τιμής 1,0), όπως υλοποιείται στη συνάρτηση `_assign_oos_row_weights()`.
- **Περιορισμός σε σύντομους ορίζοντες**: Η τεχνική εφαρμόζεται αποκλειστικά στους σύντομους ορίζοντες πρόβλεψης**,** `censoring_horizons = (15, 30, 45, 60, 75)`, καθώς οι μακροπρόθεσμες προβλέψεις (90+ ημέρες) είναι λιγότερο ευαίσθητες σε στιγμιαίες μεταβολές του αποθέματος. Οπότε απενεργοποιείται, ώστε να μην προσθέτει θόρυβο.

Η επιλογή της τιμής $\boldsymbol{w_{oos} = 0,8}$ προέκυψε κατόπιν πειραματισμού. Η πλήρης απόρριψη των γραμμών (τιμή 0,0) οδηγούσε σε απώλεια πληροφορίας και ελλιπή εκπροσώπηση των προϊόντων χαμηλής ζήτησης στο σκέλος του παλινδρομητή (*regressor*) προκαλώντας συστηματική υποπρόβλεψη σε ορίζοντες άνω των 45 ημερών. Η τιμή 0,8 επιτρέπει στο μοντέλο να διατηρεί το σήμα ότι πρόκειται για προϊόν χαμηλής κυκλοφορίας, μειώνοντας ταυτόχρονα τη μεροληψία που εισάγει η έλλειψη αποθέματος. Η στρατηγική αυτή συμπληρώνεται από την ενσωμάτωση δεικτών έλλειψης αποθέματος ως χαρακτηριστικών (features) προσφέροντας στα μοντέλα `hurdle` ρητή πληροφόρηση για την κατάσταση της αποθήκης κατά τη στιγμή της πρόβλεψης.

5.6.5 Μοντέλα Εμποδίου δύο σταδίων (*Two-stage* Hurdle *models*)

Στην τρέχουσα ενότητα αναλύεται η επιλογή των «***two-stage* Hurdle *models*»** ως το μεθοδολογικό θεμέλιο της Προσέγγισης 3. Τα μοντέλα εμποδίου (Hurdle) διαθέτουν εκτενή βιβλιογραφική βάση [4],

[70], [71] (βλ. ενότητα 3.4) και κατατάσσονται στις ώριμες μεθοδολογικές επιλογές για τη διαχείριση της διαλείπουσας ζήτησης. Παρόλο που η χρήση τους δεν αποτελεί από μόνη της καινοτομία, η συγκεκριμένη αρχιτεκτονική επιλέχθηκε ως το θεμέλιο για την αντιμετώπιση των ερευνητικών κενών που αφορούν λογοκριμένη ζήτηση (*demand censoring*) και τη μετα-μάθηση (*meta-learning).*

5.6.5.1 Διατύπωση της δομής

Η υιοθέτηση μιας δομής δύο σταδίων επιβλήθηκε από την ίδια τη φύση των δεδομένων. Είναι προφανές ότι με το **96,5% των SKU** να παρουσιάζουν διαλείπουσα ζήτηση και το **93% των γραμμών** στα στιγμιότυπα εκπαίδευσης να καταγράφουν μηδενική ζήτηση για τον ορίζοντα των 15 ημερών, η απευθείας παλινδρόμηση θα οδηγούσε μαθηματικά στην **«κατάρρευση στη μηδενική μάζα»** (*zero-mass collapse*). Το μοντέλο, δηλαδή, θα προέβλεπε διαρκώς μηδέν για να ελαχιστοποιήσει το συνολικό σφάλμα, προκαλώντας συστηματική υπό-πρόβλεψη των θετικών γεγονότων [70], [71]. Η ζήτηση, λοιπόν, μοντελοποιείται στην ουσία ως δυαδικό πρόβλημα «πουλάει ή δεν πουλάει;» πλαισιωμένο από δευτερεύοντα ποσοτικό προσδιορισμό.

Το μοντέλο λειτουργεί σε δύο διακριτά στάδια:

- **Στάδιο 1 (Ταξινομητής / *Classifier*):** Υπολογίζει την πιθανότητα $\mathbb{P}(Y > 0|X)$, δηλαδή αν θα υπάρξει καθόλου ζήτηση στον ορίζοντα h. Εκπαιδεύεται πάνω στη μεταβλητή-στόχο `Active_{h}d`.
- **Στάδιο 2 (Παλινδομητής / *Regressor*):** Υπολογίζει την αναμενόμενη ποσότητα $\mathbb{E}[Y|Y > 0, X]$, υπό την προϋπόθεση ότι η ζήτηση είναι θετική. Εκπαιδεύεται πάνω στη μεταβλητή `Target_{h}d`, αυστηρά και μόνο στο υποσύνολο των δεδομένων όπου ισχύει `Active_{h}d = 1`.

Η τελική πρόβλεψη της αναμενόμενης ζήτησης προκύπτει από το γινόμενο των δύο σταδίων ενσωματώνοντας παράλληλα έναν παράγοντα διόρθωσης (*smearing correction*), ο οποίος απαιτείται για την εξάλειψη της στατιστικής μεροληψίας που προκαλείται κατά την αντιστροφή του λογαριθμικού μετασχηματισμού της ποσότητας:

$$\hat{Y} = \hat{\mathbb{P}}(Y > 0|X) \times \hat{\mathbb{E}}[Y|Y > 0, X] \times \mathit{Smearing\_Factor} \tag{5.23}$$

Το πλεονέκτημα αυτής της αρχιτεκτονικής είναι ότι αποσυνδέει την εμφάνιση της ζήτησης από την ποσοτικοποίησή της, καθώς οι δύο αυτές διαδικασίες οδηγούνται συχνά από διαφορετικούς παράγοντες. Επομένως, επιτρέπει στον παλιδρομητή να εκπαιδευτεί μόνο σε περιπτώσεις θετικής ζήτησης αποφεύγοντας την κατάρρευση στη μηδενική μάζα.

Η συγκεκριμένη σχεδιαστική επιλογή, αν και αναπτύχθηκε ανεξάρτητα στο πλαίσιο της παρούσας μελέτης, επιβεβαιώνεται απόλυτα μέσα από μια **παράλληλη επιστημονική σύγκλιση** με την αιχμή της σύγχρονης βιβλιογραφίας. Αυτή ακριβώς η εννοιολογική αποσύνδεση αποτελεί τη βάση και για σύγχρονες σύνθετες αρχιτεκτονικές, όπως το πλαίσιο Switch-Hurdle [111], το οποίο χρησιμοποιεί κωδικοποιητές Mixture-of-Experts (MoE) σε συνδυασμό με έναν πιθανοτικό αποκωδικοποιητή Hurdle για την αντιμετώπιση της σπανιότητας των δεδομένων. Επιπροσθέτως, όπως επισημαίνουν οι Giannopoulos *et al.* [3] στην εκτενή ανασκόπησή τους, η δομή Hurdle παραμένει μια από τις πιο στιβαρές επιλογές για τη μοντελοποίηση της διαλείπουσας ζήτησης στο λιανικό εμπόριο, καθώς επιτρέπει την εξειδικευμένη διαχείριση της "μηδενικής μάζας" που χαρακτηρίζει τα βιομηχανικά σύνολα δεδομένων.

5.6.5.2 Απορριφθείσες εναλλακτικές

Για τη διασφάλιση της μεθοδολογικής εγκυρότητας και την τεκμηρίωση της τελικής επιλογής, εξετάστηκαν και απορρίφθηκαν οι κάτωθι προσεγγίσεις:

Κλασικές μέθοδοι (Croston, TSB)

Αποτελούν ιστορικό σημείο αναφοράς στον χώρο [1], [2], [4], [5] αλλά περιορίζονται σε αυστηρά μονοπαραμετρική (*univariate*) χρήση, χωρίς δυνατότητα ενσωμάτωσης πλούσιων χαρακτηριστικών (κατηγορίες, προμηθευτές, τιμές, `embeddings`, `OOS indicators – flags`), όπως αυτά που είναι διαθέσιμα στο τρέχον σύνολο δεδομένων. Οπότε, ο περιορισμός αυτός είναι καθοριστικός και τις καθιστά ακατάλληλες ως κύρια μοντέλα. Παρ' όλα αυτά, διατηρούνται εντός του *pipeline* ως σημείο αναφοράς (βλ. ενότητα 5.6.7), ώστε κάθε σύγκριση να πραγματοποιείται ρητά έναντι της κλασικής βιβλιογραφίας σύμφωνα με τις συστάσεις των Petropoulos et al. [3] και Pinçe *et al.* [96].

Μοντέλο παλινδρόμησης (*regression*) μονού σταδίου (π.χ. Tweedie, 1984)

Όπως προαναφέρθηκε, συγκεντρώνουν υπερβολική μάζα στο μηδέν οδηγώντας σε συστηματική υπο-πρόβλεψη της πραγματικής ζήτησης. Αυτό το φαινόμενο ονομάζεται «zero-mass collapse», τεκμηριώθηκε αρχικά από τον Cragg [71] και αποτέλεσε το αρχικό κίνητρο για την ανάπτυξη των *hurdle* μοντέλων [70].

Παραμετρικά μοντέλα διογκωμένου μηδενός (*zero-inflated*)

Οι αυστηρές παραμετρικές τους υποθέσεις (ότι δηλαδή τα δεδομένα ακολουθούν κατανομή Poisson ή Αρνητική Διωνυμική) καταρρίπτονται από την υπερβολική διασπορά (*overdispersion*) των δεδομένων [14], [91].

Μοντέλα Βαθιάς Μάθησης πολλαπλών οριζόντων (*Direct* Deep Learning *multi-horizon forecast* , π.χ. DeepAR, Transformers-*based* μοντέλα)

Οι αρχιτεκτονικές αυτές έχουν επιδείξει ανταγωνιστική απόδοση σε διαγωνισμούς λιανικών πωλήσεων (*retail benchmarks)* [46][170]. Όμως για το παρόν πρόβλημα παρουσιάζουν τρεις πρακτικές αδυναμίες:

- Υπερβολικές απαιτήσεις πόρων: Απαιτούν κεντρική εκπαίδευση σε καθολικά σύνολα δεδομένων με χρήση ισχυρών επεξεργαστών γραφικών (GPU), ενώ οι χρόνοι εκπαίδευσης είναι δεκαπλάσιοι σε σύγκριση με τα μοντέλα διαδοχικής ενίσχυσης κλίσης (*gradient boosting*).
- Περιορισμένη υπεροχή σε αραιή ζήτηση: Στην κατηγορία του λιανικού εμπορίου με διαλείπουσα ζήτηση, οι μέθοδοι αυτές δεν υπερτερούν συστηματικά. Αυτό επιβεβαιώθηκε από τα αποτελέσματα του διαγωνισμού M5 [21], [114], όπου οι νικητήριες λύσεις βασίστηκαν αποκλειστικά σε μοντέλα ενίσχυσης κλίσης, καθώς και σε εξειδικευμένες μελέτες που συγκρίνουν τη βαθιά μάθηση με κλασικές στατιστικές μεθόδους για αραιές χρονοσειρές [107], [121], [128].
- Δυσκολία προσαρμογής: Η ενσωμάτωση εξειδικευμένων τεχνικών, όπως η διόρθωση λογοκριμένης ζήτησης και η πολύ-επίπεδη βαθμονόμηση βάσει εσόδων (*revenue-stratified calibration*), που χρησιμοποιούνται στο τρέχον *pipeline*, σε μια ολοκληρωμένη αρχιτεκτονική βαθιάς μάθησης θα απαιτούσαν ριζικό και εξαιρετικά πολύπλοκο επανασχεδιασμό της διαδικασίας εκπαίδευσης.

Πρόσφατα ερευνητικά δεδομένα από το πλαίσιο «Smoothed Hybrid Occurrence-Size – SHOS» [98], το οποίο αξιολογήθηκε σε 56.000 χρονοσειρές ανταλλακτικών, **έρχονται να δικαιώσουν εκ των υστέρων τη στρατηγική απόφαση της παρούσας μελέτης να απορρίψει τις προσεγγίσεις Βαθιάς Μάθησης**, επιβεβαιώνοντας την υπεροχή της στατιστικά ορθής μηχανικής χαρακτηριστικών (*feature engineering*)

έναντι της δομικής πολυπλοκότητας των Νευρωνικών Δικτύων. Η προσέγγιση αυτή εναρμονίζεται πλήρως με τα ευρήματα των Muşat & Căbuz [111], οι οποίοι επισημαίνουν ότι η δομή Hurdle παραμένει η πλέον ενδεδειγμένη για τον διαχωρισμό της εμφάνισης από το μέγεθος της ζήτησης, καθώς, και με την ανασκόπηση των Giannopoulos *et al.* [45], η οποία υπογραμμίζει τη σταθερή στιβαρότητα των μοντέλων ενίσχυσης κλίσης (*gradient boosting*) σε βιομηχανικά δεδομένα μεγάλης σποραδικότητας. Αυτή η **ευρύτερη ερευνητική σύγκλιση** προκρίνει την ποιότητα και τον πλούτο της πληροφορίας εισόδου (*feature-rich approach*) έναντι των σύνθετων αρχιτεκτονικών βαθιάς μάθησης, μια επιστημονική κατεύθυνση η οποία υιοθετείται πλήρως και τεκμηριώνεται μεθοδολογικά στην παρούσα εργασία.

Συνοψίζοντας, η δενδρική Hurdle προσέγγιση προκρίθηκε ως η βέλτιστη λύση, καθώς συνδυάζει **μη-παραμετρικότητα**, πλήρη εκμετάλλευση του πλούσιου σε διαθεσιμότητα χαρακτηριστικών (*feature-rich*) περιβάλλοντος και **υπολογιστική οικονομία**. Παραμένει απόλυτα συμβατή με τις εξειδικευμένες τεχνικές `demand_censoring` και `stratified_calibration` επιτρέποντας στον παλινδρομητή να «μαθαίνει» αυθαίρετες κατανομές χωρίς τις δεσμεύσεις των κλασικών στατιστικών μοντέλων.

5.6.5.3 Δεξαμενή βασικών μοντέλων (*Base model pool*)

Στο τρέχον σύστημα (*pipeline*) χρησιμοποιούνται τέσσερα διαφορετικά μοντέλα προβλέψεων. Το κάθε ένα από αυτά αποτελεί ένα πλήρες μοντέλο Εμποδίου (Hurdle *model*), ενσωματώνοντας ανεξάρτητο ταξινομητή (*classifier*) και παλινδρομητή (*regressor*). Αναλυτικά, τα μοντέλα που συνθέτουν τη δεξαμενή παρουσιάζονται στον ακόλουθο πίνακα:

Πίνακας 5.14: Βασικά μοντέλα προβλέψεων

Μοντέλο	***Classifier***	***Regressor***	**Πλεονεκτήματα**
`Hurdle_RF`	`RandomForest Classifier`	`RandomForest Regressor`	Υψηλή σταθερότητα και χαμηλή διακύμανση
`Hurdle_ET`	`ExtraTrees Classifier`	`ExtraTrees Regressor`	Ακόμα υψηλότερη στοχαστικότητα (*variance*) προσφέροντας διαφορετική μεροληψία (*bias*)
`Hurdle_HGB`	`HistGradientBoosting Classifier`	`HistGradientBoosting Regressor`	Υψηλή ταχύτητα εκτέλεσης και φυσική διαχείριση ελλιπών τιμών (`NaN`)
`Hurdle_CAT`	`CatBoost Classifier`	`CatBoost Regressor`	Εγγενής βελτιστοποιημένη διαχείριση κατηγορικών δεδομένων και υποστήριξη GPU

Όπως έχει αναλυθεί θεωρητικά στο κεφάλαιο 2, όλα τα μοντέλα βασίζονται σε δέντρα αποφάσεων αλλά ανήκουν σε δύο διαφορετικές οικογένειες / κατηγορίες. Τα `Hurdle_RF` και `Hurdle_ET` ανήκουν στην οικογένεια των «*Bagging (Bootstrap Aggregating)*» μοντέλων που η λογική τους βασίζεται στην παράλληλη εκπαίδευση πολλών ανεξάρτητων δέντρων και τη λήψη του μέσου όρου των προβλέψεών τους. Με αυτόν τον τρόπο μειώνεται η διακύμανση. Από την άλλη, τα `Hurdle_HGB` και `Hurdle_CAT` ανήκουν στην οικογένεια των «*Boosting*» μοντέλων. Σε αυτή την περίπτωση, τα δέντρα εκπαιδεύονται διαδοχικά και κάθε νέο δέντρο προσπαθεί να διορθώσει τα σφάλματα των προηγούμενων. Έτσι, καταφέρνουν να αυξήσουν την ακρίβεια των προβλέψεων.

Η χρήση και των δύο οικογενειών είναι στρατηγική επιλογή, καθότι τα μοντέλα *Bagging* προσφέρουν τη σταθερότητα και τα μοντέλα *Boosting* την ακρίβεια. Αυτός ο συνδυασμός δημιουργεί μια πραγματική **ποικιλομορφία προβλέψεων** (*prediction diversity*), την οποία εκμεταλλεύεται, στη συνέχεια, το μοντέλο μετα-μάθησης (*meta-learner***),** για να παράγει ένα τελικό αποτέλεσμα που είναι ταυτόχρονα ακριβές και γενικεύσιμο.

5.6.5.4 Μετασχηματισμός στόχου και συναρτήσεις απώλειας (Target transformation and loss functions)

Ο παλινδρομητής (*regressor*) κάθε μοντέλου δεν εκπαιδεύεται στις ακατέργαστες τιμές-στόχους της ζήτησης, αλλά στις λογαριθμικά μετασχηματισμένες. Συγκεκριμένα, χρησιμοποιείται ο μετασχηματισμός $\log(1+x)$ (γνωστός προγραμματιστικά και ως `log1p`):

$$Y' = \log(1 + Y) \tag{5.24}$$

όπου Y είναι η ακατέργαστη ζήτηση (raw target) και Y' είναι ο μετασχηματισμένος στόχος (*transformed target*).

Ο συγκεκριμένος μετασχηματισμός εφαρμόζεται ομοιόμορφα στο κοινό μονοπάτι εκπαίδευσης, ανεξάρτητα από τον αλγόριθμο, καθώς προσφέρει τέσσερα κρίσιμα μαθηματικά οφέλη:

- Συμπιέζει την ακραία «ουρά» (*heavy tail*) της κατανομής της ζήτησης (π.χ., μεμονωμένες, σποραδικές πωλήσεις άνω των 500 τεμαχίων που θα αλλοίωναν τον μέσο όρο των δέντρων).
- Εξομαλύνει τη διακύμανση (*variance*) των δεδομένων.
- Δημιουργεί μια πιο ομοιόμορφη επιφάνεια απώλειας (*loss surface*) για τη βελτιστοποίηση των δενδρικών μοντέλων.
- Εξαλείφει την ανάγκη για παραμετρικές υποθέσεις κατανομής (π.χ. Tweedie, Poisson), εναρμονιζόμενος απόλυτα με τη μη-παραμετρική φύση των δενδρικών μεθόδων [14].

Κατά τη φάση του συμπερασμού (*inference*), εφαρμόζεται ο αντίστροφος εκθετικός μετασχηματισμός $\hat{Y} = \exp(\hat{Y}') - 1$ (`expm1(·)` προγραμματιστικά), ο οποίος ακολουθείται υποχρεωτικά από μια εξειδικευμένη στατιστική διόρθωση (*smearing correction*), προκειμένου να αποφευχθεί η συστηματική υποεκτίμηση (βλ. Ενότητα 5.6.6.1).

Επιλογή συναρτήσεων απώλειας ανά μοντέλο

Παρά τον κοινό λογαριθμικό μετασχηματισμό της μεταβλητής-στόχου, κάθε αλγόριθμος εντός της δεξαμενής του συνόλου (*ensemble*) εκπαιδεύεται κάνοντας χρήση διαφορετικής, βελτιστοποιημένης συνάρτησης απώλειας (*loss function*), τόσο για τον ταξινομητή όσο και για τον παλινδρομητή:

Πίνακας 5.15: Συναρτήσεις απώλειας μοντέλων

Μοντέλο	**Συνάρτηση απώλειας ταξινομητή**	**Συνάρτηση Απώλειας Παλινδρομητή (επί του Y')**
`Hurdle_RF`	`Gini impurity` (Προεπιλογή `scikit-learn`)	`MSE / squared_error`
`Hurdle_ET`	`Gini impurity` (Προεπιλογή `scikit-learn`)	`MSE / squared_error`
`Hurdle_HGB`	`Binary cross-entropy` (`Log loss`)	`squared_error` *ή* `quantile(0.5)` για βραχείς ορίζοντες

Μοντέλο	Συνάρτηση απώλειας ταξινομητή	Συνάρτηση Απώλειας Παλινδρομητή (επί του Y')
`Hurdle_CAT`	`Logloss` (Ρητή δήλωση, βελτιστοποίηση GPU)	`RMSE` (Ρητή δήλωση, βελτιστοποίηση GPU)

Οι επιλογές αυτές φέρουν σημαντικές θεωρητικές και επιχειρησιακές προεκτάσεις. Ειδικότερα:

- Ισοδυναμία με το `RMSLE`: Η χρήση του Μέσου Τετραγωνικού Σφάλματος (`MSE` / `RMSE`) στον λογαριθμικό χώρο (Y'), είναι μαθηματικά και λειτουργικά ισοδύναμη με τη βελτιστοποίηση βάσει του `RMSLE` (*Root Mean Squared Logarithmic Error*) στον αρχικό, φυσικό χώρο της ζήτησης. Το `RMSLE` αποτελεί τη διεθνώς καθιερωμένη επιλογή για μη-αρνητικές κατανομές με "βαριές ουρές" [21].
- Επιχειρησιακή Ασυμμετρία (*Inventory Bias*): Η βελτιστοποίηση του `MSE` στον λογαριθμικό χώρο δημιουργεί μια εγγενή ασύμμετρη ποινή στον πρωτότυπο χώρο. Πιο συγκεκριμένα, ο αλγόριθμος "τιμωρεί" τις μεγάλες υπερεκτιμήσεις (*over-forecasting*) αναλογικά λιγότερο από ότι τις μεγάλες υποεκτιμήσεις (*under-forecasting*). Στο πλαίσιο της διαχείρισης αποθεμάτων (*inventory management*), αυτή η μαθηματική ιδιότητα είναι εξαιρετικά επιθυμητή, καθώς το επιχειρηματικό κόστος της έλλειψης αποθέματος (*stock-out*) είναι, κατά κανόνα, πολλαπλάσιο του κόστους πλεονάζοντος αποθέματος (*holding cost*).

5.6.5.5 Δυναμική στάθμιση κλάσεων (*Dynamic class weighting*)

Η στατιστική φύση και η δυσκολία του προβλήματος ταξινόμησης μεταβάλλεται ριζικά ανάλογα με τον χρονικό ορίζοντα πρόβλεψης. Ενώ στις 15 ημέρες μόνο το 5 των προϊόντων παρουσιάζει πωλήσεις, στις 365 ημέρες το ποσοστό αυτό αγγίζει το 95. Οπότε, η χρήση σταθερού λόγου βαρών δεν είναι αποδοτική και στις δύο άκρες του προβλήματος. Για την αντιμετώπιση αυτής της ανισορροπίας κλάσεων, εφαρμόζονται **δυναμικά βάρη** που εξαρτώνται από το ποσοστό των θετικών περιπτώσεων (p^+) στον εκάστοτε ορίζοντα. Λόγω της της γενικότερης σπανιότητας των πωλήσεων, επιλέχθηκε να ορίζεται μεγαλύτερο βάρος στη θετική κλάση (δηλαδή, όταν υπάρχει πώληση) σε σχέση με την αρνητική. Επομένως, επιβάλλεται αυστηρότερη ποινή στα ψευδώς αρνητικά σφάλματα (*false negatives*), δηλαδή στην απώλεια πρόβλεψης μιας κρίσιμης πώλησης, παρά στα ψευδώς θετικά. Αν θεωρήσουμε ότι το βάρος της αρνητικής κλάσης είναι σταθερά στη μονάδα ($w_h^- = 1{,}0$), τότε το βάρος της θετικής κλάσης υπολογίζεται από την ακόλουθη σχέση:

$$w_h^+ = \text{clip}\left(\frac{1}{\max(p_h^+, 0{,}05)},\ w_{min},\ w_{max}\right), \qquad w_{min} = 1{,}5,\ w_{max} = 8{,}0 \qquad (5.25)$$

Η ανωτέρω σχέση μεταφράζεται στις ενδεικτικές τιμές του επόμενου πίνακα:

Πίνακας 5.16: Ενδεικτικά βάρη στάθμισης κλάσεων

Ορίζοντας	`Positive rate` p^+	w^+ (βάρος θετικής κλάσης)
15d	~5%	~8,0 (μέγιστο όριο)
30d	~15%	~6,7
90d	~60%	~1,7
365d	~95%	1,5 (ελάχιστο όριο)

Αυτά τα δυναμικά βάρη εφαρμόζονται κατά την εκπαίδευση του ταξινομητή εξασφαλίζοντας ότι το μοντέλο δεν θα «καταρρεύσει» σε μηδενικές προβλέψεις στους σύντομους ορίζοντες, ούτε θα παράγει υπερβολικά πολλά ψευδώς θετικά αποτελέσματα στους μακροπρόθεσμους.

5.6.5.6 Βελτιστοποίηση κατωφλίου ταξινομητή (*Classifier threshold tuning*)

Δεδομένης της έντονης ανισορροπίας μεταξύ των κλάσεων στα δεδομένα διαλείπουσας ζήτησης, η υιοθέτηση ενός σταθερού κατωφλίου απόφασης (*decision threshold*) στο επίπεδο του 0,5, που είναι και η προεπιλεγμένη τιμή στις βιβλιοθήκες του `scikit-learn`, κρίθηκε ανεπαρκής. Για τη βελτιστοποίηση του διαχωρισμού μεταξύ ενεργού και μη ενεργού ζήτησης, ενσωματώνεται στο *pipeline* ένας μηχανισμός δυναμικής ρύθμισης κατωφλίου μέσω της συνάρτησης `optimise_classifier_threshold()`.

Μεθοδολογία βελτιστοποίησης

Η διαδικασία βασίζεται σε τεχνική βελτιστοποίησης εκτός δείγματος (*out-of-fold optimization*). Συγκεκριμένα, ένα ποσοστό 5% του συνόλου εκπαίδευσης (*training set*) απομονώνεται, προκειμένου να χρησιμοποιηθεί αποκλειστικά για την εύρεση του βέλτιστου κατωφλίου. Η χρήση δεδομένων, τα οποία δεν συμμετείχαν στη διαδικασία προσαρμογής των παραμέτρων του μοντέλου, είναι κρίσιμη για την αποφυγή της υπερεκπαίδευσης (*overfitting*) και τη διασφάλιση της ικανότητας γενίκευσης των προβλέψεων σε μελλοντικά δεδομένα. Επισημαίνεται ότι η τρέχουσα διαδικασία πραγματοποιείται αποκλειστικά και μόνο για την εύρεση του κατάλληλου κατωφλίου. Ο τελικός ταξινομητής εκπαιδεύεται στο 100% των δεδομένων εκπαίδευσης χρησιμοποιώντας το κατώφλι που υπολογίστηκε.

Διαφοροποίηση κριτηρίων ανά ορίζοντα πρόβλεψης

Η επιλογή της μετρικής-στόχου για τη βελτιστοποίηση διαφοροποιείται στρατηγικά βάσει του χρονικού ορίζοντα *h*, αντανακλώντας τη μεταβολή της στατιστικής φύσης του προβλήματος:

- **Βραχυπρόθεσμοι ορίζοντες ($h < 90$ ημέρες):** Σε αυτά τα χρονικά πλαίσια, η ζήτηση είναι εξαιρετικά αραιή (*intermittent*), με την κλάση της μηδενικής πώλησης να κυριαρχεί αριθμητικά. Ως εκ τούτου, η βελτιστοποίηση στοχεύει στη μεγιστοποίηση του `F1-score`. Ο επιλογή του `F1-score`, το οποίο ισούται με τον αρμονικό μέσο της Ακρίβειας (Precision) και της Ανάκλησης (Recall), επιβάλλει στο μοντέλο μια αυστηρή ισορροπία. Στοχεύει στην αποφυγή τόσο των ψευδώς θετικών προβλέψεων (που οδηγούν σε άσκοπη δέσμευση κεφαλαίου), όσο και των ψευδώς αρνητικών (που οδηγούν σε ελλείψεις αποθέματος).
- **Μακροπρόθεσμοι ορίζοντες ($h \geq 90$ ημέρες):** Καθώς ο ορίζοντας διευρύνεται, η πιθανότητα εμφάνισης τουλάχιστον μίας πώλησης αυξάνεται σημαντικά, με αποτέλεσμα η κλάση απουσίας ζήτησης (`no-sale`) να καθίσταται ολοένα και πιο σπάνια. Σε αυτές τις συνθήκες κορεσμού της θετικής κλάσης, το `F1-score` καθίσταται μη επιλεκτικό (*non-selective*), καθώς σχεδόν οποιοδήποτε κατώφλι παράγει υψηλή βαθμολογία. Για τον λόγο αυτό, το κριτήριο μετατοπίζεται στο **ποσοστό επιτυχίας (**`hit-rate`**)**. Το `hit-rate` αναδεικνύεται ως καταλληλότερο κριτήριο, καθώς εστιάζει στη διασφάλιση της ορθότητας των προβλέψεων δραστηριότητας αποφεύγοντας την παγίδευση του μοντέλου σε στατιστικό θόρυβο κατά τη διάρκεια μακρών περιόδων βέβαιης ζήτησης.

Υλοποίηση και σκοπός

Αλγοριθμικά, η συνάρτηση εκτελεί μια αναζήτηση επί ενός πλέγματος τιμών πιθανότητας (από 0,01 έως 0,99). Το κατώφλι, που επιτυγχάνει τη βέλτιστη τιμή της επιλεγμένης μετρικής, «κλειδώνει» και εφαρμόζεται κατά τη φάση της εξαγωγής συμπερασμάτων (*inference*). Με αυτόν τον τρόπο, το μοντέλο

Hurdle προσαρμόζει αυτόματα την ευαισθησία του (*sensitivity*) στις ιδιαιτερότητες του κάθε προϊόντος και του εκάστοτε ορίζοντα βελτιώνοντας, παράλληλα, τη συνολική αξιοπιστία του συστήματος πρόβλεψης.

5.6.5.7 Επιλογή υπερπαραμέτρων (*Hyperparameter tuning*)

5.6.5.8 5.5.5.7 Επιλογή υπερπαραμέτρων

Η επιλογή των βέλτιστων υπερπαραμέτρων αποτελεί κρίσιμο παράγοντα για τη σταθερότητα και την τελική απόδοση του συστήματος προβλέψεων. Ο στόχος είναι η βελτιστοποίηση της απόδοσης των μοντέλων αποφεύγοντας, παράλληλα, την υπέρ-εκπαίδευση και διασφαλίζοντας την ικανότητα γενίκευσης. Οι τελικές επιλογές ανά μοντέλο παρουσιάζονται στον Πίνακα 5.17 .

Πίνακας 5.17: Βελτιστοποιημένες υπερπαράμετροι των βασικών μοντέλων

Μοντέλο	**Ταξινιομητής (*Classifier*)**	**Παλινδρομητής (*Regressor*)**
Hurdle_RF	"n_estimators": 300, "max_depth": 15, "min_samples_leaf": 15, "max_features": "sqrt", "max_samples": 0.7, "n_jobs": -1, "class_weight": "balanced_subsample", "random_state": *seed*	"n_estimators": 400, "max_depth": 12, "min_samples_leaf": 12, "max_features": "sqrt", "max_samples": 0.7, "n_jobs": -1, "random_state": *seed*
Hurdle_ET	"n_estimators": 300, "max_depth": 18, "min_samples_leaf": 12, "max_features": "sqrt", "n_jobs": -1, "class_weight": "balanced_subsample", "random_state": *seed*	"n_estimators": 400, "max_depth": 15, "min_samples_leaf": 8, "max_features": "sqrt", "n_jobs": -1, "random_state": *seed*
Hurdle_HGB	"max_depth": 5, "max_leaf_nodes": 30, "max_iter": 1500, "early_stopping": True, "validation_fraction": 0.15, "n_iter_no_change": 10, "learning_rate": 0.03, "min_samples_leaf": 50, "l2_regularization": 0.5, "max_features": 0.7, "random_state": *seed*	"max_depth": 5, "max_leaf_nodes": 30, "max_iter": 1500, "early_stopping": True, "validation_fraction": 0.15, "n_iter_no_change": 10, "learning_rate": 0.03, "min_samples_leaf": 40, "l2_regularization": 0.5, "max_features": 0.7, "random_state": *seed*
Hurdle_CAT	"depth": 6, "learning_rate": 0.03,	"depth": 6, "learning_rate": 0.03,

Μοντέλο	Ταξινιομητής (*Classifier*)	Παλινδρομητής (*Regressor*)

	`"iterations": 1000,`
	`"min_data_in_leaf": 50,`
	`"l2_leaf_reg": 3.0,`
	`"random_seed": seed,`
	`"loss_function": "Logloss",`
	`"task_type": "GPU",`
	`"border_count": 128,`
	`"verbose": False`
		`"iterations": 2500,`
		`"early_stopping_rounds": 50,`
		`"use_best_model": True,`
		`"min_data_in_leaf": 40,`
		`"l2_leaf_reg": 3.0,`
		`"random_seed": seed,`
		`"loss_function": "RMSE",`
		`"task_type": "GPU",`
		`"border_count": 128,`
		`"verbose": False`

Για την αναζήτηση και τον καθορισμό των βέλτιστων τιμών χρησιμοποιήθηκε το πλαίσιο `Optuna` [113], το οποίο υλοποιεί προηγμένους αλγορίθμους Μπεϋζιανής βελτιστοποίησης (Bayesian Optimization). Μέσω της διαδικασίας αυτής, το *pipeline* ελαχιστοποιεί τη συνάρτηση απώλειας δοκιμάζοντας διαφορετικούς συνδυασμούς παραμέτρων στον χώρο αναζήτησης. Τα τελικά, βελτιστοποιημένα αποτελέσματα αποθηκεύονται σε εξωτερικό αρχείο ρυθμίσεων και φορτώνονται αυτόματα σε επόμενες εκτελέσεις για λόγους αναπαραγωγιμότητας και υπολογιστικής εξοικονόμησης.

5.6.6 Βαθμονόμηση και μετα-επεξεργασία (*Calibration and post-processing*)

Ακόμα και μετά από μια βέλτιστη προσαρμογή των δέντρων του μοντέλου Hurdle, οι ακατέργαστες προβλέψεις (*raw predictions*) τείνουν να εμφανίζουν συστηματικές μεροληψίες (*biases*). Αυτές οφείλονται σε τρεις κύριους παράγοντες, (α) στην εκπαίδευση επί λογαριθμικής κλίμακας και τον επακόλουθο αντίστροφο μετασχηματισμό, (β) σε αναπόφευκτες διαφορές στην κατανομή μεταξύ των συνόλων εκπαίδευσης και αξιολόγησης και (γ) στην εγγενή τάση των στατιστικών μοντέλων να υποεκτιμούν τη ζήτηση σε δεδομένα ακραίας σποραδικότητας. Ακολουθεί αναλυτική παρουσίαση των τεχνικών που επιστρατεύονται για την εξομάλυνση αυτών των φαινομένων.

5.6.6.1 Διόρθωση *Smearing* (Duan, 1983)

Όπως αναλύθηκε, η εκπαίδευση των παλινδρομητών (*regressors*) του μοντέλου Hurdle πραγματοποιείται στη λογαριθμική κλίμακα `log1p(y)` για τη σταθεροποίηση της διακύμανσης. Ωστόσο, η επιστροφή στην αρχική κλίμακα μέσω της συνάρτησης `expm1(·)` εισάγει συστηματική υποεκτίμηση της αναμενόμενης τιμής. Το φαινόμενο αυτό αποτελεί άμεση συνέπεια της Ανισότητας Jensen (Jensen's Inequality), σύμφωνα με την οποία, για οποιαδήποτε αυστηρά κυρτή συνάρτηση (όπως η εκθετική), η αναμενόμενη τιμή της συνάρτησης είναι πάντα μεγαλύτερη ή ίση από τη συνάρτηση της αναμενόμενης τιμής:

$$\mathbb{E}[exp(X)] \geq exp(\mathbb{E}[X]) \tag{5.26}$$

Για την αποκατάσταση αυτής της μεροληψίας, ο Duan (1983) πρότεινε έναν μη-παραμετρικό συντελεστή διόρθωσης (*smearing factor*) βασισμένο στον μέσο όρο των εκθετικών καταλοίπων. Στην παρούσα υλοποίηση, υιοθετείται η παραμετρική λογαριθμοκανονική (*log-normal*) προσέγγιση, η οποία είναι στατιστικά ισοδύναμη υπό την υπόθεση της κανονικότητας των καταλοίπων (*residuals*), αξιοποιώντας τη ροπογεννήτρια συνάρτηση (*moment generating function*) για *t=1*. Ο συντελεστής διόρθωσης *s* υπολογίζεται ως εξής:

$$s = \exp\left(\frac{\sigma_{\hat{r}}^2}{2}\right) \tag{5.27}$$

Όπου $\sigma_{\hat{r}}^2 = \mathrm{Var}(\hat{r}_i)$ είναι η διακύμανση των καταλοίπων $\hat{r}_i = \log 1\,p(y) - \log\_pred$ στο σύνολο εκπαίδευσης. Είναι κρίσιμο να σημειωθεί ότι ο υπολογισμός της διακύμανσης πραγματοποιείται **αποκλειστικά στις γραμμές με ενεργή ζήτηση ($y > 0$)** ακολουθώντας πλήρως τη φιλοσοφία του σταδίου παλινδρόμησης του μοντέλου Hurdle.

Δυναμική περικοπή και προσαρμογή ορίζοντα

Ο συντελεστής *s* υπολογίζεται διακριτά για κάθε ζεύγος (μοντέλο, χρονικός ορίζοντας). Παρόλο που η κατηγοριοποίηση κατά Syntetos–Boylan (βάσει των δεικτών CV^2 και ADI) δεν χρησιμοποιείται για την παραγωγή διαφορετικών συντελεστών εντός του ίδιου ορίζοντα, διαδραματίζει καθοριστικό ρόλο στον έλεγχο των ορίων περικοπής (*clip bounds*). Τα όρια αυτά διασφαλίζουν ότι η διόρθωση παραμένει εντός επιχειρησιακά αποδεκτών πλαισίων αποτρέποντας φαινόμενα υπέρ-διόρθωσης σε ακραίες κατανομές:

Πίνακας 5.18: Κατηγοριοποίηση περικοπής

Κατηγορία SBC	**Συνθήκη κατηγοριοποίησης**	**Όρια περικοπής (*Clip bounds*)**
Smooth	ADI < 1.32, CV^2 < 0.49	[1.00, 1.50]
Erratic	ADI < 1.32, $\mathrm{CV}^2 \geq 0.49$	[1.00, 2.00]
Intermittent	ADI ≥ 1.32, CV^2 < 0.49	[1.05, 2.50]
Lumpy	ADI ≥ 1.32, $\mathrm{CV}^2 \geq 0.49$	[1.10, 3.00]

5.6.6.2 Απόσβεση πρόσφατης δραστηριότητας (*Recency damping*)

Οι βραχυπρόθεσμοι ορίζοντες πρόβλεψης επηρεάζονται δυσανάλογα από απότομες αυξήσεις (*spikes*) της ζήτησης στο πρόσφατο ιστορικό. Ως αντίμετρο, κατά την διάρκεια της εκπαίδευσης εφαρμόζεται εκθετική στάθμιση πρόσφατης δραστηριότητας (*exponential recency weighting*), χρησιμοποιώντας μια σταθερά χρόνου εξασθένησης (*decay time constant* τ_h) ειδικά προσαρμοσμένη στον εκάστοτε ορίζοντα h:

$$w_i = exp\left(-\frac{\mathrm{cutoff} - d_i}{\tau_h}\right) \tag{5.28}$$

Όπου $\mathrm{cutoff} - d_i$ αντιπροσωπεύει την «ηλικία» (σε ημέρες) της παρατήρησης i από το σημείο αποκοπής. Με αυτή τη διατύπωση, η βαρύτητα μιας παρατήρησης μειώνεται στο $\sim 36{,}8\ (1/e)$ της αρχικής της αξίας, όταν η ηλικία της εξισωθεί με τη σταθερά τ_h.

Οι τιμές της σταθεράς τ_h αρχικοποιούνται ως εξής:

$$\tau_{15} = 30, \tau_{30} = 60, \tau_{45} = 60, \tau_{60} = 75, \tau_{75} = 90, \tau_{90} = 90, \tau_{180} = 180, \tau_{365} = 365$$

Η συγκεκριμένη κλιμάκωση διασφαλίζει ότι τα μοντέλα βραχυπρόθεσμων οριζόντων (π.χ. $h = 15$) "ξεχνούν" γρηγορότερα τα παλαιά γεγονότα, προσαρμοζόμενα άμεσα στις τρέχουσες συνθήκες, ενώ τα μοντέλα μακροπρόθεσμων οριζόντων διατηρούν εκτενέστερη μνήμη για την αποτύπωση της μακροχρόνιας τάσης.

5.6.6.3 Βαθμονόμηση με παρακρατηθέν δείγμα ανά σημείο αποκοπής (*Per-cutoff holdout calibration*)

Η επιλογή των παραμέτρων βαθμονόμησης επιβάλλεται να είναι απόλυτα στεγανή από διαρροές (*leakage-safe*). Κάθε σημείο αποκοπής (*cutoff*) του αναδρομικού ελέγχου (*backtest*) εκπαιδεύει και αξιολογεί τις παραμέτρους βαθμονόμησης χρησιμοποιώντας **αποκλειστικά δεδομένα που προηγούνται της δικής του ημερομηνίας**, χωρίς να γνωρίζει απολύτως τίποτα για τις μεταγενέστερες περιόδους. Αυτό είναι το θεμελιώδες αξίωμα, προκειμένου να αποφευχθεί η λεγόμενη μεροληψία πρόβλεψης μέλλοντος (*look-ahead bias*) στις τελικές μετρικές απόδοσης του αναδρομικού ελέγχου.

Η διαδικασία, η οποία υλοποιείται μέσω της συνάρτησης `choose_calibration()`, εκτελείται ανεξάρτητα για κάθε σημείο αποκοπής c_b του αναδρομικού ελέγχου (*backtest*), καθώς, και για κάθε συνδυασμό μοντέλου και ορίζοντα, ακολουθώντας τα εξής βήματα:

- **Ορισμός Δείγματος:** Από το σύνολο των ιστορικών σημείων αποκοπής εκπαίδευσης $\{c_1, \dots, c_{b-1}\}$ επιλέγεται αυστηρά το πλέον πρόσφατο (c_{b-1}) ως εσωτερικό παρακρατηθέν δείγμα ελέγχου (*internal holdout*).
- **Ορισμός Πλέγματος Αναζήτησης:** Ορίζεται ένα πλέγμα (grid) υποψήφιων τιμών για τον χρόνο ημίσειας ζωής (*recency half-life*). Οι υποψήφιες τιμές $\tau \in \{\text{None}, 30,45,60,75,90,105,120,180,270,365,540\}$ περιορίζονται δυναμικά ανά ορίζοντα h, ώστε να διατηρούν επιχειρησιακό νόημα και να επιταχύνουν τον υπολογισμό. Η χαρτογράφηση του χώρου αναζήτησης δηλώνεται ρητά στο αρχείο ρυθμίσεων (`cfg`) ως εξής:

```
calibration_half_life_candidates_by_horizon={
        15: (None, 30.0, 45.0, 60.0),
        30: (None, 60.0, 75.0, 90.0),
        45: (None, 60.0, 90.0, 105.0),
        60: (None, 75.0, 90.0, 120.0),
        75: (None, 90.0, 105.0, 120.0),
        90: (None, 90.0, 105.0, 120.0),
        180: (None, 120.0, 180.0, 270.0),
        365: (None, 270.0, 365.0, 540.0),
    }
```

- **Υπολογισμός Κλίμακας:** Για κάθε υποψήφιο τ, ο παλινδρομητής παράγει προβλέψεις στο παρακρατηθέν δείγμα (*holdout sample*). Στη συνέχεια, υπολογίζεται ο γραμμικός συντελεστής κλίμακας (*scale factor*) ο οποίος ελαχιστοποιεί το Σταθμισμένο Απόλυτο Σφάλμα (WAPE). Ο υπολογισμός δίνεται από τον τύπο:

$$Scale = \frac{\sum_i w_i \cdot Y_i}{\sum_i w_i \cdot \hat{Y}_i^{\text{raw}}} \tag{5.29}$$

  όπου Y_i είναι η πραγματική αθροιστική ζήτηση, $\hat{Y}_i^{\text{raw}}$ η ακατέργαστη πρόβλεψη και w_i το αντίστοιχο βάρος του είδους (ενσωματώνοντας και την αυξημένη βαρύτητα των προϊόντων υψηλού τζίρου, εφόσον απαιτείται).

- **Περικοπή (*Clipping*):** Ο υπολογισμένος συντελεστής κλίμακας περιορίζεται εντός προκαθορισμένων ορίων ανά ορίζοντα (π.χ., $[0.95, 1.40]$ για τον ετήσιο ορίζοντα 365d), ώστε να αποφευχθούν ακραίες ή ανεξέλεγκτες διορθώσεις σε στατιστικά παθολογικές περιπτώσεις.
- **Τελική Επιλογή:** Επιλέγεται οριστικά το βέλτιστο ζεύγος $(\tau^*, Scale^*)$, το οποίο επέτυχε το χαμηλότερο σφάλμα (`cal_WAPE`) στο παρακρατηθέν δείγμα.

Τα επιλεγμένα βέλτιστα ζεύγη αποθηκεύονται στην κρυφή μνήμη (cache) του backtest μαζί με τις εκτός δείγματος (OOS) προβλέψεις, ώστε να χρησιμοποιηθούν στη συνέχεια ως είσοδος στο μοντέλο μετα-μάθησης (meta-learner, βλ. Ενότητα 5.6.9).

Για την εξαγωγή της τελικής, μελλοντικής πρόβλεψης, οι παράμετροι βαθμονόμησης επανεκτιμώνται χρησιμοποιώντας το πλέον πρόσφατο σημείο αποκοπής του αναδρομικού ελέγχου (*backtest cutoff*) ως παρακρατηθέν δείγμα (*holdout*). Έτσι, διασφαλίζεται ότι αντανακλούν τις πλέον πρόσφατες συνθήκες και τάσεις της αγοράς. Μέσω αυτού του αρχιτεκτονικού σχεδιασμού, το pipeline διατηρεί αυστηρή χρονική διαφάνεια διασφαλίζοντας ότι κάθε παράθυρο αξιολόγησης «βλέπει» αποκλειστικά το παρελθόν του, χωρίς καμία πληροφορία να διαρρέει από μεταγενέστερα σημεία αποκοπής.

5.6.6.4 Σταθμισμένη βαθμονόμηση: Προστασία κωδικών υψηλού τζίρου (*Weighted calibration*)

Ο προεπιλεγμένος υπολογισμός του συντελεστή κλίμακας (*scale*), ο οποίος βασίζεται αποκλειστικά στην ελαχιστοποίηση του συνολικού σφάλματος WAPE, αντιμετωπίζει στατιστικά ισότιμα όλους τους κωδικούς προϊόντων (SKUs). Ωστόσο από επιχειρησιακή σκοπιά, τα προϊόντα που αποφέρουν τα υψηλότερα έσοδα (*top-revenue* SKUs) είναι κρίσιμης σημασίας. Μια ελαφρά υποεκτίμηση της ζήτησης σε αυτά τα είδη επιφέρει δυσανάλογα μεγαλύτερο κόστος (λόγω απώλειας πωλήσεων και δυσαρέσκειας πελατών) σε σύγκριση με μια αντίστοιχη υποεκτίμηση στα προϊόντα της «μακράς ουράς» (*long-tail*).

Για την αντιμετώπιση αυτού του επιχειρησιακού κινδύνου, το σύστημα εφαρμόζει σταθμισμένη βαθμονόμηση (*weighted calibration*) αποδίδοντας διαφοροποιημένες βαρύτητες που προσαρμόζονται δυναμικά στον εκάστοτε χρονικό ορίζοντα πρόβλεψης *h*, όπως παρουσιάζεται και στο κάτωθι απόσπασμα του αρχείου ρυθμίσεων `cfg`:

```
weighted_calibration_by_horizon={
#(π.χ. Οι 2000 κορυφαίοι κωδικοί έχουν 5πλάσια
βαρύτητα στον υπολογισμό του σφάλματος στις 15 ημέρες)
        15:  {"top_n": 2000, "top_weight": 5.0},
        30:  {"top_n": 1500, "top_weight": 4.5},
        45:  {"top_n": 1200, "top_weight": 3.0},
        60:  {"top_n": 1000, "top_weight": 3.0},
        75:  {"top_n": 1000, "top_weight": 3.0},
        90:  {"top_n": 800,  "top_weight": 3.0},
        180: {"top_n": 600,  "top_weight": 2.5},
        365: {"top_n": 400,  "top_weight": 2.0},
    }
```

Μηχανισμός λειτουργίας: Κατά τη διάρκεια της αξιολόγησης στο παρακρατηθέν δείγμα επικύρωσης (*holdout evaluation*) (βλ. 5.6.6.3), τα SKUs που εντάσσονται στην κορυφαία υπό-ομάδα (`top-N cohort`) βάσει ιστορικών εσόδων λαμβάνουν αυξημένο βάρος δείγματος ($w =$ `top_weight`), ενώ τα υπόλοιπα διατηρούν το προεπιλεγμένο βάρος ($w = 1{,}0$).

Με αυτόν τον μηχανισμό, ο αλγόριθμος που επιλέγει τον βέλτιστο συντελεστή κλίμακας καθοδηγείται («ακούει») εντονότερα από την απόδοση του μοντέλου στα προϊόντα υψηλού τζίρου. Έτσι, παρέχεται μια συστημική προστασία απέναντι στην υπό-πρόβλεψη και τις επακόλουθες ελλείψεις αποθέματος για τα πολυτιμότερα είδη της επιχείρησης.

5.6.6.5 Επιβολή μονοτονίας στις προβλέψεις (Monotonicity enforcement – PAVA)

Το Πρόβλημα:

Οι χρονικοί ορίζοντες εκπαιδεύονται ως **ανεξάρτητα** προβλήματα παλινδρόμησης, με τον καθένα να διαθέτει τον δικό του στόχο. Εφόσον, όμως, οι στόχοι αντιπροσωπεύουν αθροιστική ζήτηση και η ημερομηνία έναρξης των προβλέψεων είναι κοινή, ισχύει ταυτοτικά η εξής ανισότητα για κάθε κωδικό είδους:

$$\mathrm{y}_{\mathrm{h}_1}^{\mathrm{actual}} \leq \mathrm{y}_{\mathrm{h}_2}^{\mathrm{actual}} \quad \forall \mathrm{h}_1 < \mathrm{h}_2 \tag{5.30}$$

Με άλλα λόγια, η πρόβλεψη για κάθε ορίζοντα θα πρέπει να είναι τουλάχιστον ισόποση, ή μεγαλύτερη, από αυτές όλων των μικρότερων, από αυτόν, οριζόντων.

Ωστόσο, αυτή η μαθηματική εγγύηση δεν ενσωματώνεται εγγενώς στη διαδικασία εκπαίδευσης, καθώς κάθε ορίζοντας μοντελοποιείται από έναν ξεχωριστό παλινδρομητή. Εμπειρικά, παρατηρείται ότι οι ακατέργαστες προβλέψεις συχνά παραβιάζουν αυτή την υποτιθέμενη μονοτονία σε ένα μικρό ποσοστό των SKUs (τυπικά 3–8%). Αυτό οφείλεται σε τρεις παράγοντες:

- Στη διαφοροποιημένη βαθμονόμηση κλίμακας (*scale calibration*) ανά ορίζοντα.
- Στη διαφορετική σχετική βαρύτητα των χαρακτηριστικών (*feature importance*) ανά ορίζοντα (π.χ. το μοντέλο των 15 ημερών εστιάζει στη βραχυπρόθεσμη δυναμική, ενώ των 90 ημερών σε εποχικά πρότυπα).
- Στην τυχαία διακύμανση προϊόντων με εξαιρετικά χαμηλή ζήτηση.

Εάν αυτές οι παραβιάσεις παραμείνουν, προκαλούν ταλαντώσεις στον συνολικό λόγο (*aggregate ratio*) μεταξύ των οριζόντων και δημιουργούν παράλογες επιχειρηματικές συστάσεις (π.χ. προτροπή για παραγγελία 12 τεμαχίων για τον επόμενο μήνα, αλλά μόνο 10 τεμαχίων για το επόμενο δίμηνο συνολικά).

Η Λύση:

Αλγόριθμος PAVA (Pool Adjacent Violators Algorithm). Για την αποκατάσταση της συνέπειας, εφαρμόζεται εκ των υστέρων (*post-hoc*) ανά SKU και ανά μοντέλο, ο αλγόριθμος της Ισοτονικής Παλινδρόμησης (Isotonic Regression). Ο αλγόριθμος PAVA [79], [80] επιλύει σε γραμμικό χρόνο $\mathcal{O}(n)$ το ακόλουθο πρόβλημα ελαχιστοποίησης L2 υπό περιορισμούς (*constrained L2 optimization*):

$$\hat{y}^{mono} = \arg\min_{z \in \mathbb{R}^H} \sum_{h=1}^{H} (z_h - \hat{y}_h^{raw})^2 \quad \text{υπό} \quad z_{h_1} \leq z_{h_2} \ \forall h_1 < h_2 \tag{5.31}$$

Ως ελαχιστοποίηση $L2$ καλείται η ευρέως γνωστή μέθοδος Ελαχίστων Τετραγώνων (Least Squares), η οποία είναι ίσως η πιο θεμελιώδης έννοια στη στατιστική και τη μηχανική μάθηση. Η λύση που προσφέρει ο PAVA είναι L2-βέλτιστη, καθώς ελαχιστοποιεί το συνολικό τετραγωνικό σφάλμα από τις αρχικές προβλέψεις. Αυτή η προσέγγιση είναι ριζικά ανώτερη από την απλοϊκή επιβολή του κανόνα του μεγίστου $\hat{y}_h^{mono} = \max(\hat{y}_h^{raw}, \hat{y}_{h+1}^{mono})$, διότι ο απλοϊκός κανόνας:

- Δεν είναι L2-βέλτιστος: Μπορεί να «ανεβάσει» αδικαιολόγητα την πρόβλεψη των 90 ημερών, ώστε να ξεπεράσει μια ήδη υπερεκτιμημένη πρόβλεψη των 30 ημερών, παγιώνοντας το σφάλμα αντί να το εξομαλύνει.
- Είναι ασύμμετρος: Διορθώνει μόνο τις προς τα κάτω παραβιάσεις, αφήνοντας τις προς τα πάνω υπερβολές ανέπαφες.
- Δημιουργεί τεχνητά «σκαλοπάτια»: Μια μικρή παραβίαση προκαλεί απότομες αναπηδήσεις (*discrete jumps*) στην πρόβλεψη.

Ο PAVA, αντίθετα, συγχωνεύει (*pools*) τις γειτονικές παραβιάσεις στη μέση τιμή τους, κατανέμοντας ομαλά τη διόρθωση σε όλα τα εμπλεκόμενα σημεία. Η πρακτική επίδραση αποτυπώνεται στο ακόλουθο παράδειγμα:

- **Ακατέργαστες προβλέψεις** ($h \in \{15, 30, 60, 90, 180, 365\}$): `[12.0, 10.5, 14.0, 13.2, 25.0, 50.0]` *(Παραβιάσεις σε 30d & 90d)*
- **Απλοϊκός κανόνας μεγίστου (naive max-rule)**: `[12.0, 12.0, 14.0, 14.0, 25.0, 50.0]` *(Χάνει πληροφορία, διατηρεί τα λάθη)*
- **PAVA (L2-βέλτιστο)**: `[11.25, 11.25, 13.6, 13.6, 25.0, 50.0]` *(Συγχωνεύει τα ζεύγη [15,30] και [60,90])*

Υλοποίηση και Αντίκτυπος

Η διαδικασία υλοποιείται με τη συνάρτηση `_enforce_monotone_cumulative_predictions()` και εφαρμόζεται σε όλα τα παραγόμενα μοντέλα (`base hurdle models`, `Ensemble_Weighted`, `Ensemble_Median`, `Ensemble_Meta`) πριν από την τελική τους αποθήκευση. Παρατηρείται ότι η μέση μείωση του `WAPE` από την εφαρμογή του `PAVA` είναι μικρή (περίπου 0,5–1,5% σε απόλυτες μονάδες). Παρόλα αυτά, όμως, η επίτευξη απόλυτης **διαχρονικής συνέπειας (*temporal consistency*)** είναι καθοριστικής σημασίας για την αξιοπιστία των προβλέψεων κατά τον μεταγενέστερο σχεδιασμό αποθεμάτων.

5.6.7 Κλασικά μοντέλα αναφοράς (Baselines: Croston-SBA, TSB)

Με σκοπό τη διασφάλιση της μεθοδολογικής αρτιότητας και τη συγκριτική αξιολόγηση των αποτελεσμάτων, το σύστημα (*pipeline*) ενσωματώνει και παράγει προβλέψεις από δύο κλασικές μεθόδους της βιβλιογραφίας:

- **Croston-SBA (Syntetos-Boylan Approximation) [1], [5]:** Αποτελεί διόρθωση της αρχικής μεθόδου Croston [4] με στόχο την εξάλειψη της ασυμπτωτικής μεροληψίας (*asymptotic bias*) που οδηγούσε σε συστηματική υπερεκτίμηση. Μέχρι σήμερα, συνιστά το *de facto* πρότυπο αναφοράς (*benchmark*) για την πρόβλεψη της διαλείπουσας ζήτησης.
- **TSB [2]:** Η μέθοδος αυτή αποτελεί εξέλιξη του Croston, καθώς αποσυνδέει την εκθετική εξομάλυνση της πιθανότητας ζήτησης από την εξομάλυνση του μεγέθους της παραγγελίας. Αυτή η διαφοροποίηση προσφέρει βελτιωμένη συμπεριφορά σε περιβάλλοντα, όπου τα προϊόντα τείνουν προς απαξίωση (*obsolescence*) ή οριστική κατάργηση, καθώς ενημερώνει την πιθανότητα ακόμα και σε περιόδους μηδενικής ζήτησης.

Υλοποίηση και Σκοπιμότητα

Η αλγοριθμική υλοποίηση πραγματοποιείται μέσω της ενιαίας συνάρτησης `_croston_tsb_forecast()`. Οι παραγόμενες προβλέψεις καταγράφονται ως διακριτές εγγραφές στο συγκεντρωτικό αρχείο `summary_sparse_true_forecast.csv`.

Ο πρωταρχικός σκοπός της συμπερίληψης αυτών των μεθόδων δεν είναι ο άμεσος ανταγωνισμός τους με τα προηγμένα μοντέλα μηχανικής μάθησης. Αντιθέτως, χρησιμεύουν ως σταθερά σημεία αναφοράς για την αποφυγή του κινδύνου αξιολογικής πλάνης, δηλαδή με αυτό τον τρόπο διασφαλίζεται ότι η υπεροχή των μοντέλων ML τεκμηριώνεται με ρητή αντιπαραβολή έναντι καθιερωμένων βιβλιογραφικών προτύπων.

Όπως καταδεικνύεται στην ανάλυση των αποτελεσμάτων (βλ. ενότητα 5.6.14), τόσο η μέθοδος Croston-SBA όσο και η TSB υπολείπονται δραματικά σε απόδοση έναντι των μοντέλων Hurdle (με σφάλμα

`WAPE` > *30* έναντι `WAPE` ~ *0.5 - 1.1*). Το εύρημα αυτό είναι απολύτως αναμενόμενο, διότι τα κλασικά στατιστικά μοντέλα αποτελούν αυστηρά μονομεταβλητές (*univariate*) προσεγγίσεις. Κατά συνέπεια, αγνοούν πλήρως τα διαστρωματικά σήματα (*cross-sectional signals*), όπως είναι η πλούσια πληροφορία που αντλείται από τα χαρακτηριστικά (*features*) και η συνδυαστική συμπεριφορά άλλων κωδικών, και περιορίζονται, αποκλειστικά, στο απομονωμένο ιστορικό του εκάστοτε προϊόντος.

5.6.8 Κατασκευή Συνδυαστικών Μοντέλων (*Ensembles*)

Στο προτεινόμενο σύστημα υλοποιούνται συνολικά τρία διαφορετικά συνδυαστικά μοντέλα (*ensembles*). Τα δύο εξ' αυτών δημιουργούνται μετά την ολοκλήρωση της βαθμονόμησης αξιοποιώντας τις ήδη παραγόμενες προβλέψεις χωρίς να απαιτείται επιπρόσθετη εκπαίδευση. Το τρίτο παράγεται μέσω του μοντέλου μέτα-μάθησης και διαθέτει σαφώς υψηλότερη πολυπλοκότητα. Στην τρέχουσα ενότητα αναλύονται οι μηχανισμοί των δύο πρώτων προσεγγίσεων.

5.6.8.1 Σταθμισμένο Συνδυαστικό Μοντέλο (`Ensemble_Weighted`)

Η πρόβλεψη αυτού του μοντέλου προκύπτει ως ο σταθμισμένος μέσος όρος των τεσσάρων επιμέρους μοντέλων, όπου τα βάρη υπολογίζονται αντιστρόφως ανάλογα του σφάλματος βαθμονόμησης στο παρακρατηθέν δείγμα επικύρωσης (`holdout WAPE`):

$$w_m = \frac{1/(\mathrm{WAPE}_m + 0.01)}{\sum_j 1/(\mathrm{WAPE}_j + 0.01)} \tag{5.32}$$

Αξίζει να σημειωθεί ότι τα βάρη (w_m) υπολογίζονται διακριτά για κάθε χρονικό ορίζοντα. Κατά συνέπεια, εάν ένα μοντέλο παρουσιάσει μειωμένη απόδοση σε έναν συγκεκριμένο ορίζοντα (π.χ. το `Hurdle_RF` στις 15 ημέρες), θα λάβει αναλογικά μικρότερο βάρος σε αυτόν το σημείο, χωρίς ωστόσο να τιμωρηθεί ή να επηρεαστεί η συνεισφορά του στους υπόλοιπους, μακροπρόθεσμους ορίζοντες. Η προσθήκη της σταθεράς *0.01* στον παρονομαστή διασφαλίζει την αποφυγή διαίρεσης με το μηδέν στην περίπτωση της τέλειας πρόβλεψης.

5.6.8.2 Συνδυαστικό Μοντέλο Διαμέσου (`Ensemble_Median`)

Η πρόβλεψη υπολογίζεται ως η διάμεσος των επιμέρους προβλέψεων των τεσσάρων μοντέλων για κάθε κωδικό (SKU) ξεχωριστά. Η χρήση της διαμέσου καθιστά το τελικό αποτέλεσμα σημαντικά πιο ανθεκτικό (*robust*) στις ακραίες τιμές (*outliers*) σε σύγκριση με τον απλό ή τον σταθμισμένο μέσο όρο.

Η ιδιότητα αυτή αποδεικνύεται εξαιρετικά χρήσιμη στην πράξη. Δηλαδή, σε περιπτώσεις, όπου ένα μεμονωμένο μοντέλο «απορρυθμίζεται» και παράγει μια υπερβολικά αποκλίνουσα πρόβλεψη λόγω ενός ασυνήθιστου ιστορικού προτύπου (π.χ. μια ακραία υπερεκτίμηση του `Hurdle_RF` για ένα συγκεκριμένο SKU), η διάμεσος απομονώνει και, πρακτικά, αγνοεί την ανώμαλη τιμή διαφυλάσσοντας τη σταθερότητα της πρόβλεψης.

Και τα δύο αυτά συνδυαστικά μοντέλα ενσωματώνονται και καταγράφονται ως διακριτές εγγραφές στο τελικό αρχείο αναφοράς `summary_sparse_true_forecast.csv` και, ουσιαστικά, αποτελούν τη βάση σύγκρισης για το πιο προηγμένο `Ensemble_Meta` που αναλύεται στη συνέχεια.

5.6.9 Αξιολόγηση μέσω αναδρομικού ελέγχου (*Backtesting*)

Για την εκπαίδευση του μοντέλου μέτα-μάθησης (*meta-learner*) και την εξαγωγή αξιόπιστων συμπερασμάτων σχετικά με την απόδοση του συστήματος, το *pipeline* εκτελεί μια σειρά αναδρομικών δοκιμών σε ιστορικά δεδομένα.

Ο αναδρομικός έλεγχος (*backtesting*) αποτελεί μια αυστηρή διαδικασία προσομοίωσης, η οποία αναπαριστά ακριβώς τον τρόπο με τον οποίο θα λειτουργούσε το σύστημα, εάν είχε κληθεί να εκτελέσει προβλέψεις σε μια παρελθοντική χρονική στιγμή αγνοώντας πλήρως τη μετέπειτα εξέλιξη των πωλήσεων.

Η διαδικασία βασίζεται στην έννοια των χρονικών σημείων αποκοπής (*cutoffs*). Κάθε σημείο αποκοπής λειτουργεί ως μια υποθετική "σημερινή ημέρα" (`forecast origin`):

- Ορισμός σημείων αποκοπής: Μέσω της συνάρτησης `propose_training_cutoffs()` αναγνωρίζονται και ορίζονται τα επιθυμητά σημεία τομής στον άξονα του χρόνου.
- Προσομοίωση τυφλής πρόβλεψης: Για κάθε ένα από αυτά τα σημεία, το μοντέλο "τυφλώνεται" για οτιδήποτε συνέβη μετά από αυτή την ημερομηνία.

Για κάθε cutoff, ο αλγόριθμος χωρίζει τα δεδομένα ως εξής:

- Δεδομένα εκπαίδευσης (*Training set*): Όλα τα στιγμιότυπα δεδομένων (*snapshots*) με ημερομηνία προγενέστερη του εκάστοτε cutoff. Το μοντέλο εκπαιδεύεται να μαθαίνει τη σχέση των χαρακτηριστικών (features) με τη ζήτηση που έλαβε χώρα στο παρελθόν.
- Δεδομένα αξιολόγησης (*Test set*): Το στιγμιότυπο που αντιστοιχεί ακριβώς στην ημερομηνία αποκοπής. Αυτό το στιγμιότυπο περιέχει τα πραγματικά δεδομένα ζήτησης που συνέβησαν τις επόμενες 15, 30, 90 κ.λπ. ημέρες (`Targets`), τα οποία χρησιμοποιούνται για τη σύγκριση με τις προβλέψεις.

Μετά την παραγωγή των προβλέψεων για κάθε SKU και ορίζοντα, το σύστημα υπολογίζει μια σειρά από δείκτες απόδοσης, όπως ακριβώς συμβαίνει και στην κύρια φάση εκπαίδευσης – αξιολόγησης.

Η εν λόγω αξιολόγηση μέσω αναδρομικού ελέγχου είναι σημαντική για τους κάτωθι λόγους:

- **Επιλογή μοντέλου**: Το σύστημα αξιολογεί ποιο βασικό μοντέλο (π.χ. `Hurdle_RF`, `Hurdle_CAT` ή τα `Ensembles`) απέδωσε καλύτερα ιστορικά σε κάθε ορίζοντα.
- **Βαθμονόμηση (Calibration)**: Προσφέρει τον απαραίτητο, μη-μολυσμένο (*leakage-safe*) χώρο για την εύρεση των βέλτιστων παραμέτρων διόρθωσης της κλίμακας, όπως αναλύθηκε διεξοδικά στην Ενότητα 5.5.6.3.
- **Εκπαίδευση μοντέλου μετα-μάθησης:** Τα σφάλματα που προκύπτουν (`Out-of-Sample predictions`) αποθηκεύονται για να εκπαιδεύσουν ένα μοντέλο μετά-μάθησης που μαθαίνει πώς να συνδυάζει τις προβλέψεις των επιμέρους μοντέλων με τον βέλτιστο τρόπο.

5.6.9.1 Σχεδιασμός αναδρομικού ελέγχου

Ο σχεδιασμός του αναδρομικού ελέγχου δεν βασίζεται σε στατικά, προκαθορισμένα σημεία, αλλά προσδιορίζεται δυναμικά μέσω των ρυθμίσεων του συστήματος (αρχείο `cfg`). Συγκεκριμένα, ο μέγιστος αριθμός των παραθύρων αξιολόγησης ελέγχεται από την παράμετρο `backtest_max_cutoffs`.

Όταν η διαδικασία είναι ενεργοποιημένη (`load_backtest = True`), η συνάρτηση `propose_training_cutoffs()` αντλεί τον κατάλογο όλων των διαθέσιμων σημείων αποκοπής και επιλέγει τα *N* πλέον πρόσφατα εξ αυτών. Για παράδειγμα, στην προεπιλεγμένη εκτέλεση της παρούσας μελέτης, κατά την οποία επιλέχθηκε `backtest_max_cutoffs = 3` (και τα σημεία παραγωγής εξαμηνιαία), το σύστημα αυτοματοποιημένα απομονώνει τα εξής τρία ιστορικά σημεία αποκοπής:

- 2022-07-01
- 2023-01-01
- 2023-07-01

Για κάθε επιλεγμένο σημείο αποκοπής c_b, ακολουθείται αυστηρά η κάτωθι ροή εκτέλεσης:

- **Εκπαίδευση:** Τα τέσσερα βασικά μοντέλα (Hurdle models) εκπαιδεύονται αποκλειστικά στο υποσύνολο των δεδομένων που διαθέτουν ημερομηνία στιγμιοτύπου προγενέστερη του c_b.
- **Πρόβλεψη:** Παράγονται προβλέψεις για τους ορίζοντες ενδιαφέροντος εντός του διαστήματος $(c_b, c_b + h]$.
- **Σύγκριση:** Οι προβλέψεις αντιπαραβάλλονται με τις πραγματικές αθροιστικές ποσότητες ζήτησης, προκειμένου να υπολογιστούν οι μετρικές σφάλματος (`WAPE`, `Ratio`, `Score`, `MASE`) ανά συνδυασμό μοντέλου, ορίζοντα και σημείου αποκοπής.

Αυτός ο σχεδιασμός προσομοιάζει τη διασταυρούμενη επικύρωση επεκτεινόμενου παραθύρου (*expanding window cross-validation*) για χρονοσειρές [39], βελτιστοποιημένη στις ιδιαίτερες απαιτήσεις ενός επιχειρησιακού πλαισίου που βασίζεται σε περιοδικά στιγμιότυπα δεδομένων.

5.6.9.2 Επιτάχυνση της διαδικασίας εκπαίδευσης

Δεδομένου ότι τα πλήρη μοντέλα εκπαιδεύονται σε εκατοντάδες χιλιάδες δείγματα φέροντας άνω των 500 χαρακτηριστικών (μαζί με τα embeddings), η πλήρης επανεκπαίδευσή τους για τρεις διαφορετικές πτυχές (*folds*) θα αύξανε δραματικά τον συνολικό χρόνο εκτέλεσης. Για την αντιμετώπιση αυτού του τεχνικού περιορισμού, εφαρμόζεται ένας ελεγχόμενος μηχανισμός παράκαμψης παραμέτρων (`backtest_model_params_override`), ο οποίος μειώνει την πολυπλοκότητα των μοντέλων κατά τον αναδρομικό έλεγχο:

- `Hurdle_CAT`: Οι επαναλήψεις (`iterations`) μειώνονται από 1200 σε 600.
- `Hurdle_RF` / `Hurdle_ET`: Ο αριθμός των δέντρων (`n_estimators`) μειώνεται από 500 σε 200.

Η χρήση «ελαφρύτερων» μοντέλων κατά τον αναδρομικό έλεγχο δεν υποβαθμίζε τη χρησιμότητά τους για την εκπαίδευση του μετα-μοντέλου. Ο αλγόριθμος μετα-μάθησης βασίζεται σε προβλέψεις εκτός δείγματος (OOS *predictions*) και η αποτελεσματικότητά του εξαρτάται από τη σταθερότητα της σχετικής αξιολόγησης (*relative ratings*) των μοντέλων και όχι από την απόλυτη ταυτότητα των υπερπαραμέτρων τους.

5.6.9.3 Ενδιάμεση μνήμη (*Cache*) με ψηφιακό αποτύπωμα

Λόγω του υψηλού υπολογιστικού φόρτου, τα αποτελέσματα του αναδρομικού ελέγχου αποθηκεύονται τοπικά σε αρχείο μορφής `CSV`, συνοδευόμενα από ένα μοναδικό ψηφιακό αποτύπωμα (*fingerprint*) μορφής `JSON`. Το αποτύπωμα αυτό ενσωματώνει τις κρίσιμες μεταβλητές της παραμετροποίησης, όπως την έκδοση του *pipeline*, τη διαδρομή των δεδομένων, την αρχή της πρόβλεψης (`forecast origin`), τους ορίζοντες, τα σημεία αποκοπής και τον αρχικό σπόρο τυχαιότητας (`seed`).

Σε κάθε νέα εκτέλεση του κώδικα, το σύστημα (μέσω της συνάρτησης `_build_backtest_cache_metadata()`) ανακατασκευάζει το τρέχον αποτύπωμα βάσει των ενεργών ρυθμίσεων του αρχείου `cfg` και το συγκρίνει με το ήδη αποθηκευμένο:

- Εάν τα αποτυπώματα ταυτίζονται απόλυτα, τα αποτελέσματα φορτώνονται ακαριαία από την ενδιάμεση μνήμη (cache) παρακάμπτοντας την εκπαίδευση του *backtest*.
- Εάν υπάρχει διαφωνία (έστω και σε μία παράμετρο), ολόκληρος ο αναδρομικός έλεγχος εκτελείται εξ ολοκλήρου από την αρχή.

Αυτός ο αρχιτεκτονικός μηχανισμός διασφαλίζει την απόλυτη ακεραιότητα των πειραμάτων επιτυγχάνοντας, ταυτόχρονα, σημαντική εξοικονόμηση υπολογιστικών πόρων κατά τη φάση της ανάπτυξης.

5.6.10 Συσσωρευτική γενίκευση μοντέλου μετα-μάθησης (*Meta-learner stacking*)

Ένα δεύτερο, ανώτερο επίπεδο συνδυασμού μοντέλων (*ensemble layer*) τοποθετείται πάνω από τα βασικά μοντέλα Hurdle, υιοθετώντας την αρχιτεκτονική της συσσωρευτικής γενίκευσης (*stacked* generalization) [58]. Σε αυτή τη διάταξη, ένας μετα-συνδυαστής (*meta-blender*) εκπαιδεύεται πάνω στις προβλέψεις εκτός δείγματος (*OOS predictions*) του αναδρομικού ελέγχου, με σκοπό να διδαχθεί τον βέλτιστο συνδυασμό των επιμέρους μοντέλων.

Η επιλογή αυτή εδράζεται στις παρατηρήσεις της Ενότητας 3.5. Τα νικητήρια σχήματα του διαγωνισμού M5 [21], [114] και οι σύγχρονες προσεγγίσεις μετα-μάθησης για διαλείπουσα ζήτηση [108] κατέδειξαν ότι ο επιμελής συνδυασμός διαφορετικών (*heterogeneous*) μοντέλων βάσης προσφέρει σταθερά καλύτερη ικανότητα γενίκευσης από κάθε μεμονωμένο μοντέλο.

Στην τρέχουσα υλοποίηση, ο μετα-συνδυαστής τροφοδοτείται και από τα τέσσερα βασικά μοντέλα `Hurdle_RF`, `Hurdle_ET`, `Hurdle_HGB` και `Hurdle_CAT`. Από την άλλη, οι μέθοδοι `Croston_SBA` και `TSB` παραμένουν εκτός λειτουργώντας, αποκλειστικά, ως στατιστικά μέτρα αναφοράς.

5.6.10.1 Επιλογή περιορισμένης βελτιστοποίησης (*constrained optimization*) έναντι RidgeCV

Η πρώτη, προφανής βιβλιογραφική επιλογή για γραμμική στοίβαξη (*stacking*) είναι το μοντέλο Παλινδρόμησης Κορυφογραμμής με διασταυρούμενη επικύρωση (`RidgeCV`) [38]. Ωστόσο, η χρήση του απορρίφθηκε για τρεις λόγους:

- Αρνητικά βάρη: Το `RidgeCV` παρήγε συχνά αρνητικούς συντελεστές, γεγονός που στερείται επιχειρησιακής λογικής (είναι άτοπο ένα μοντέλο να «αφαιρεί» ποσότητα ζήτησης από ένα άλλο).
- Ασυμβατότητα κριτηρίων: Το `RidgeCV` βελτιστοποιεί το τετραγωνικό σφάλμα (`MSE`), ενώ ο βασικός στόχος του συστήματος είναι η ακρίβεια συνολικής μάζας (`Ratio`) σε συνδυασμό με την ακρίβεια σε επίπεδο δείγματος (`WAPE`).
- Κατάρρευση ποικιλομορφίας: Σε περιπτώσεις υψηλής συσχέτισης των μοντέλων βάσης, τείνει να συγκεντρώνει το βάρος σε ένα μόνο μοντέλο προβλέψεων ακυρώνοντας τα οφέλη του *ensemble*.

Για τους ανωτέρω λόγους, ο *meta-learner* υλοποιείται ως **Περιορισμένος Κυρτός Συνδυαστής (Constrained Convex Blender)**μέσω του αλγορίθμου `L-BFGS-B` που υλοποιείται στη συνάρτηση `scipy.optimize.minimize`.

5.6.10.2 Διατύπωση της Συνάρτησης Βελτιστοποίησης

Έστω M το πλήθος των βασικών μοντέλων που συμμετέχουν στον συνδυασμό. Αναζητείται το διάνυσμα βαρών $\mathrm{w} \in \mathbb{R}^{\mathbb{M}}$ που ελαχιστοποιεί την κάτωθι τριπλή συνάρτηση σύνθετης ποινής $\mathcal{L}(\mathrm{w})$:

$$\mathcal{L}(\mathrm{w}) \;=\; \lambda_{\text{ratio}} \cdot \left| \frac{\sum_i \hat{y}_i}{\sum_i y_i} - 1 \right| \;+\; \lambda_{\text{wape}} \cdot \mathrm{WAPE}(\hat{y}, y) \;+\; \lambda_{\text{div}} \cdot \sum_{m=1}^{M} \left(w_m - \frac{1}{M} \right)^2 \qquad (5.33)$$

όπου $\hat{y} = X_{\text{meta}} \cdot w$ αντιπροσωπεύει τη συνολική συσσωρευμένη (*stacked*) πρόβλεψη.

Η συνάρτηση αποτελείται από τρεις όρους:

- **Ευθυγράμμιση λόγου (*Ratio-alignment*):** Τιμωρεί την απόκλιση της συνολικής προβλεπόμενης μάζας από την πραγματική ωθώντας το σύστημα σε μη-μεροληπτικές προβλέψεις ($\tau\iota\mu\acute{\eta}\ \sigma\tau\alpha\theta\epsilon\rho\acute{\alpha}\varsigma\ \lambda_{\text{ratio}} = 0{,}80$).

- **Ακρίβεια (WAPE):** Το κλασικό μέτρο σφάλματος ανά δείγμα για την τιμωρία των ατομικών αποκλίσεων (*τιμή σταθεράς* $\lambda_{\text{wape}} = 0{,}20$).
- **Ποικιλομορφία L2 (L2 *Diversity*):** Λειτουργεί ως όρος κανονικοποίησης (*Bayesian shrinkage*) που τιμωρεί την απομάκρυνση των βαρών από την ομοιόμορφη κατανομή $1/\text{M}$ εμποδίζοντας την κυριαρχία ενός μόνο μοντέλου (*τιμή σταθεράς* $\lambda_{\text{div}} = 0{,}10$).

5.6.10.3 Η σημασία του κυρτού συνδυασμού (*convex combination*)

Η επιβολή των περιορισμών $w_m \geq 0$ και $\sum w_m = 1$ εξασφαλίζει ότι η τελική πρόβλεψη αποτελεί κυρτό συνδυασμό (*convex combination*) των αρχικών προβλέψεων. Αυτό προσφέρει τρία κρίσιμα πλεονεκτήματα:

- Φυσική ερμηνευσιμότητα: Κάθε μοντέλο συνεισφέρει θετικά ή αγνοείται ($w_m = 0$). Τα βάρη διαβάζονται απευθείας ως μερίδιο εμπιστοσύνης από τους ενδιαφερόμενους (π.χ., "το 48% της τελικής πρόβλεψης προέρχεται από το `CatBoost`").
- Ευστάθεια: Η τελική τιμή παραμένει εντός του κυρτού περιβλήματος (*convex hull*) των αρχικών προβλέψεων. Αυτό αποτρέπει την κατάρρευση σε παράλογες (αρνητικές ή υπερβολικές) τιμές σε περιπτώσεις ακραίων δεδομένων εισόδου.
- Σταθερότητα: Το μοντέλο παρουσιάζει χαμηλότερη διακύμανση απόδοσης, καθώς εξομαλύνει τις στιγμιαίες αστοχίες των επιμέρους αλγορίθμων.

5.6.10.4 Διαδικασία εκπαίδευσης

Η εκπαίδευση πραγματοποιείται διακριτά για κάθε ορίζοντα *h ως εξής*:

- Σχηματίζεται ο πίνακας μετα-δεδομένων $X_{\text{meta}} \in \mathbb{R}^{N \times M}$ στοιβάζοντας ως στήλες τις `OOS` προβλέψεις $\hat{y}_m$ που παρήγαγε κάθε βασικό μοντέλο κατά τη φάση του αναδρομικού ελέγχου (βλ. ενότητα 5.6.9).
- Τα βάρη αρχικοποιούνται στην ομοιόμορφη κατανομή *1/M*, δηλαδή $w^0 = (1/M, \ldots, 1/M)$.
- Ο αλγόριθμος `L-BFGS-B` εκτελείται μέχρι τη σύγκλιση (συνήθως σε < 50 επαναλήψεις) τηρώντας τους περιορισμούς μη-αρνητικότητας.
- Τα τελικά βάρη κανονικοποιούνται ώστε να αθροίζουν ακριβώς σε $1{,}0$ και χρησιμοποιούνται κατά τη φάση της πρόβλεψης (inference): $\hat{y}^{meta} = X^{inf} \cdot w$.

Όπως καταδεικνύεται στην αξιολόγηση (βλ. ενότητα 5.6.15), το `Ensemble_Meta` επιτυγχάνει την υψηλότερη αξιοπιστία. Προσφέρει τον βέλτιστο συνδυασμό ακρίβειας (`WAPE`) και συνολικής ευθυγράμμισης της ζήτησης (`Ratio`). Κατά συνέπεια, αποτελεί την προεπιλεγμένη έξοδο (`default output`) για την επιχειρησιακή διαχείριση των αποθεμάτων.

5.6.11 Στρωματοποιημένη κλιμάκωση (*Stratified scaling: Top-revenue protection*)

Παρόλο που η καθολική βαθμονόμηση (*global calibration*, βλ. ενότητα 5.6.6) βελτιστοποιεί το συνολικό σφάλμα `WAPE`, δεν διασφαλίζει εγγενώς την ακρίβεια στις προβλέψεις των επιχειρησιακά κρίσιμων κωδικών. Η ομάδες ενδιαφέροντος στην προκειμένη περίπτωση είναι οι Top-20 βάσει τζίρου, κατά πρώτον, και τα είδη της κλάσης Α (*Pareto-A* – 80%), κατά δεύτερον.

Η ανάγκη για μια στρωματοποιημένη (*stratified*) προσέγγιση προκύπτει απευθείας από τα δομικά χαρακτηριστικά του συνόλου δεδομένων (βλ. κεφάλαιο 4). Η κατανομή του τζίρου παρουσιάζει ακραία συγκέντρωση (συντελεστής Gini = 0,742), με τους 20 κορυφαίους κωδικούς να αντιπροσωπεύουν το 6,4–7,2% των συνολικών εσόδων της πενταετίας απασχολώντας μόλις το 2,7% της αξίας των

αποθεμάτων. Επίσης, η κλάση Α (*Pareto-A*) αποτελεί το περίπου 22% των συνολικών κωδικών ειδών και είναι υπεύθυνη για το 80% του συνολικού τζίρου.

Υπό αυτό το πρίσμα, οποιαδήποτε υποεκτίμηση της ζήτησης σε αυτές τις ομάδες (*cohorts*) μεταφράζεται σε ελλείψεις αποθέματος με εξαιρετικά υψηλό επιχειρησιακό κόστος. Η μεθοδολογία ευθυγραμμίζεται με τη βιβλιογραφική παραδοχή ότι η ακρίβεια της πρόβλεψης δεν είναι αυτοσκοπός αλλά αποτελεί εργαλείο για τη βέλτιστη λήψη αποφάσεων αναπλήρωσης [3], [95], [131]. Ιδιαίτερα σε περιπτώσεις, όπου τα κόστη σφάλματος διαφέρουν ριζικά μεταξύ των κρίσιμων ειδών και της «μακράς ουράς» του καταλόγου.

Εμπειρικά, παρατηρήθηκε συστηματική υπό-πρόβλεψη της τάξης του 15–25% στους κορυφαίους κωδικούς για τους σύντομους ορίζοντες (15 και 30 ημερών). Επομένως, η λύση που υιοθετήθηκε είναι η **στρωματοποιημένη κλιμάκωση περιορισμένης ομάδας (*narrow-cohort stratified scaling*)**.

5.6.11.1 Δυναμική επιλογή ομάδας (*Cohort selection*)

Η συνάρτηση `_resolve_stratified_scaling_top_n()` επιλέγει δυναμικά το βέλτιστο μέγεθος *K* της ομάδας των κορυφαίων κωδικών ειδών ακολουθώντας την εξής λογική:

- Εξετάζονται υποψήφια μεγέθη $K \in \{50, 75, 100\}$.
- Κριτήριο επιλογής αποτελεί το μικρότερο *K,* για το οποίο το μερίδιο τζίρου της ομάδας είναι τουλάχιστον 5% επί του συνολικού ιστορικού αξίας πωλήσεων (`Total_Value_History`).
- Εάν κανένα μέγεθος δεν πληροί το κατώφλι, επιλέγεται η μέγιστη τιμή (*K=100*).

Στο τρέχον σύνολο δεδομένων επιλέγεται το *K=50*, καθώς πληροί οριακά τη συνθήκη του 5%.

5.6.11.2 Υπολογισμός συντελεστή ανά επίπεδο (*Tier-specific scale*)

Για κάθε συνδυασμό μοντέλου και ορίζοντα, η διαδικασία εξειδικευμένης βαθμονόμησης εξελίσσεται σε τρία στάδια:

- Υπολογισμός τοπικού συντελεστή: Προσδιορίζεται ο συντελεστής κλίμακας αποκλειστικά για την ομάδα Top-K:

$$\text{scale}_{\text{top}} = \frac{\sum_{i \in \text{top-K}} y_i}{\sum_{i \in \text{top-K}} \widehat{y_i^{\text{raw}}}} \tag{5.34}$$

- Περιορισμός (*Clipping*): Ο συντελεστής περιορίζεται εντός του εύρους $[0{,}85, 1{,}60]$, επιτρέποντας σημαντική ενίσχυση (*lift*), όπου απαιτείται.

- Συνδυασμός (*Blending*): Ο τελικός συντελεστής για την ομάδα προκύπτει από τη στάθμιση του τοπικού με τον καθολικό συντελεστή ($\alpha = 0{,}70$):

$$\text{scale}_{\text{top-eff}} = \alpha \cdot \text{scale}_{\text{top}} + (1 - \alpha) \cdot \text{scale}_{\text{final}} \tag{5.35}$$

Οι κωδικοί που βρίσκονται εκτός της ομάδας Top-K συνεχίζουν να βαθμονομούνται με τον καθολικό συντελεστή (`scale_final`) διασφαλίζοντας ότι η συνολική ισορροπία (`global ratio`) του συστήματος δεν διαταράσσεται.

5.6.11.3 Αξιολόγηση της διόρθωσης

Η εφαρμογή της στρωματοποιημένης κλιμάκωσης οδήγησε σε θεαματική βελτίωση του λόγου πρόβλεψης (`ratio`) για τους 20 κορυφαίους κωδικούς:

- **Ορίζοντας 15 ημερών:** Βελτίωση από **0,79 σε 0,96** (+17 ποσοστιαίες μονάδες).

- **Ορίζοντας 30 ημερών:** Βελτίωση από **0,78 σε 0,90**.
- **Ορίζοντας 45 ημερών:** Βελτίωση από **0,83 σε 0,98**.

Η βελτίωση αυτή επιτεύχθηκε χωρίς να επηρεαστούν οι καθολικές μετρικές του συστήματος (μετατόπιση του `global ratio` < 1%). Έτσι, επιβεβαιώνεται την αποτελεσματικότητα της επιλεκτικής βαθμονόμησης ως μηχανισμού θωράκισης των κρίσιμων προϊόντων της επιχείρησης.

5.6.12 Διαστήματα Πρόβλεψης (*Prediction Intervals* - PI)

Πέραν των κλασικών σημειακών προβλέψεων (*point forecasts*), για κάθε συνδυασμό κωδικού και ορίζοντα ($SKU, horizon$) παράγεται ένα Διάστημα Πρόβλεψης (PI) με ονομαστικό επίπεδο εμπιστοσύνης 80. Οι τιμές εξαγωγής αντιστοιχούν στα ποσοστημόρια 10 και 90 της κατανομής (`interval_quantiles = (0.1, 0.9)`) οριοθετώντας ένα εύρος $\left[q^{\text{low}},\, q^{\text{high}}\right]$.

5.6.12.1 Μεθοδολογία βασισμένη σε κατάλοιπα ποσοστημορίων

Όσον αφορά τον υπολογισμό των διαστημάτων εμπιστοσύνης, η τρέχουσα προσέγγιση αποφεύγει ενσυνείδητα την εκπαίδευση ανεξάρτητων μοντέλων παλινδρόμησης ποσοστημορίων (*quantile regressors*), όπως συμβαίνει στην Προσέγγιση 1 (βλ. Ενότητα 5.4), διότι κάτι τέτοιο θα διπλασίαζε τον χρόνο εκτέλεσης και τον υπολογιστικό φόρτο του pipeline. Αντ' αυτού, εφαρμόζεται ένα πιο ευέλικτο και ελαφρύ εμπειρικό σχήμα (*residual-quantile based*), το οποίο αξιοποιεί πλήρως τα ήδη παραχθέντα αποτελέσματα του αναδρομικού ελέγχου:

- Υπολογίζονται τα κατάλοιπα σφάλματος (*residuals*) $\mathrm{r_i}$ από τις προβλέψεις εκτός δείγματος (`OOS`) που παρήγαγε το μοντέλο κατά τον αναδρομικό έλεγχο (*backtesting*):

$$\mathrm{r_i} = \mathrm{Y_i} - \widehat{\mathrm{Y}_\mathrm{i}} \tag{5.36}$$

- Προσδιορίζονται τα εμπειρικά ποσοστημόρια $Q_{0.1}$ και $Q_{0.9}$ των καταλοίπων αυτών. Ο υπολογισμός πραγματοποιείται διακριτά ανά συνδυασμό μοντέλου, ορίζοντα και κατηγορίας διαλείπουσας ζήτησης (*intermittent SBC class*) διασφαλίζοντας ότι η αβεβαιότητα προσαρμόζεται στη στατιστική συμπεριφορά του εκάστοτε προϊόντος.
- Το τελικό διάστημα πρόβλεψης για μια νέα σημειακή πρόβλεψη $\hat{Y}_{\text{new}}$ διαμορφώνεται ως:

$$PI \;=\; \left[\max\left(0,\, \hat{Y}_{\text{new}} + Q_{0.1}\right),\;\; \hat{Y}_{\text{new}} + Q_{0.9}\right] \tag{5.37}$$

(Σημειώνεται ότι το κατώτατο όριο περιορίζεται στο μηδέν μέσω της συνάρτησης $\max$, καθώς η φυσική ζήτηση δεν μπορεί να λάβει αρνητικές τιμές).

Αυτή η μη-παραμετρική, εμπειρική προσέγγιση δεν βασίζεται σε αυστηρές υποθέσεις κατανομής, αλλά αντανακλά το πραγματικό ιστορικό σφάλμα του συστήματος. Όπως αναλύθηκε θεωρητικά στην Ενότητα 5.2.3.8, τα ποσοστημόρια $Q_{0.1}$ και $Q_{0.9}$ αντιπροσωπεύουν το 10 και το 90 της αθροιστικής κατανομής αποτυπώνοντας, στην ουσία, το «απαισιόδοξο» και το «έντονα αισιόδοξο» σενάριο ζήτησης, αντίστοιχα. Η εμπειρική απόδοση κάλυψης (`PI_Coverage`), όπως αποτυπώνεται στις αναφορές της αξιολόγησης, παραμένει συστηματικά εξαιρετικά κοντά στο ονομαστικό επίπεδο ($80\% \pm 5\%$) επιβεβαιώνοντας τη στατιστική αξιοπιστία του σχήματος.

5.6.12.2 Κανονικοποιημένο Εύρος Διαστήματος (`PI_Width`)

Το σχετικό, κανονικοποιημένο εύρος του διαστήματος (`PI_Width`) χρησιμοποιείται ως η κεντρική μετρική για την αποτίμηση της αβεβαιότητας των προβλέψεων και υπολογίζεται σύμφωνα με τον μαθηματικό τύπο που ορίστηκε στην Ενότητα 5.2.3.8:

$$\text{PI\_Width} = \frac{q^{\text{high}} - q^{\text{low}}}{\max(\hat{Y}, 1)} \tag{5.38}$$

Η χρήση της συνάρτησης $\max(\hat{Y}, 1)$ στον παρονομαστή είναι καθοριστική, διότι διασφαλίζει την αποφυγή μαθηματικού απειρισμού (διαίρεση με το μηδέν) σε περιπτώσεις, κατά τις οποίες η παραγόμενη σημειακή πρόβλεψη είναι μηδενική ($\hat{Y} = 0$).

Στόχος της αρχιτεκτονικής είναι ο ιδανικός συνδυασμός υψηλής εμπειρικής κάλυψης (κοντά στο 80%) με το μικρότερο δυνατό (πιο «σφιχτό») εύρος. Ένα μεγάλο εύρος τιμών υποδηλώνει δομική αβεβαιότητα και χαμηλή εμπιστοσύνη του μοντέλου στη συγκεκριμένη πρόβλεψη.

Όπως επιβεβαιώνεται από τα πειραματικά αποτελέσματα, το εύρος των διαστημάτων αυξάνεται μονοτονικά με την επέκταση του χρονικού ορίζοντα h. Το φαινόμενο αυτό αντικατοπτρίζει απόλυτα φυσιολογικά τη διεύρυνση της στατιστικής αβεβαιότητας, όσο το σύστημα επιχειρεί να προβάλει τις εκτιμήσεις του στο απώτερο μέλλον.

Από επιχειρησιακής σκοπιάς, η επίτευξη ενός στενότερου εύρους (μικρότερο `PI_Width`), με παράλληλη διατήρηση της ονομαστικής κάλυψης, αποτελεί ισχυρή ένδειξη βεβαιότητας. Στο πλαίσιο της διαχείρισης αποθεμάτων, η ιδιότητα αυτή μεταφράζεται άμεσα σε **μικρότερες απαιτήσεις για απόθεμα ασφαλείας** μειώνοντας το δεσμευμένο κεφάλαιο κίνησης χωρίς να αυξάνεται ο κίνδυνος εμφάνισης ελλείψεων.

5.6.13 Μελέτη αποδόμησης και απορριφθείσες τεχνικές (Ablation study and disabled features)

Κατά την ανάπτυξη και εξέλιξη του προτεινόμενου συστήματος, εξετάστηκαν εκτενώς τεχνικές, οι οποίες, μολονότι παραμένουν δομικά υλοποιημένες στον κώδικα (ως επιλογές στο `cfg`), απενεργοποιήθηκαν ρητά στη βασική, παραγωγική παραμετροποίηση. Παρόλα αυτά, η αναφορά τους στο παρόν κεφάλαιο δεν στερείται, διόλου, σημασίας. Αντιθέτως, καταγράφει με διαφάνεια τη μεθοδολογική διαδρομή και αιτιολογεί την επιστημονική απόρριψη λύσεων, οι οποίες, παρά τη βιβλιογραφική ή ερευνητική τους γοητεία, απέτυχαν να προσφέρουν μετρήσιμη επιχειρησιακή στιβαρότητα.

5.6.13.1 Δίκτυα Κατάστασης Ηχούς (Echo State Networks - ESN)

Το Δίκτυο Κατάστασης Ηχούς (ESN) [81] ως μοντέλο Υπολογιστικής Δεξαμενής (Reservoir Computing) αναλύεται θεωρητικά στο Κεφάλαιο 2. Η συγκεκριμένη αρχιτεκτονική δοκιμάστηκε πειραματικά στο παρόν σύστημα, καθώς προτείνεται συχνά στη βιβλιογραφία ως μια ταχύτερη εναλλακτική των κλασικών RNN/LSTM, με πρόσφατες μελέτες να αναφέρουν οφέλη σε δεδομένα βιομηχανικών ανταλλακτικών [138].

Ωστόσο, όπως επισημάνθηκε στο Ερευνητικό Κενό 5 (Ενότητα 3.8), η πραγματική αποτελεσματικότητα αυτών των μοντέλων έναντι ισχυρών δενδρικών αλγορίθμων σε περιβάλλοντα της δευτερογενούς αγοράς αυτοκινήτου, παραμένει έντονα υπό διερεύνηση.

Αιτία απενεργοποίησης (`use_esn_features=False`):

Ο λόγος απόρριψης είναι απλός και ξεκάθαρος: το ESN εισήγαγε «θόρυβο» αντί για σήμα στα μοντέλα Hurdle.

- Υψηλή διακύμανση**:** Σε σύντομους ορίζοντες, η δεξαμενή αντιδρούσε υπερβολικά έντονα σε μεμονωμένες αιχμές πωλήσεων (*spikes*) αναπαράγοντας πληροφορία, την οποία τα δέντρα

ελάμβαναν ήδη (και πολύ πιο καθαρά) μέσω των αυστηρών κυλιόμενων χαρακτηριστικών (*rolling features*).

- Περιορισμένη ιστορικότητα (*Data scarcity*): Η εφαρμογή του reservoir σε επίπεδο μεμονωμένου SKU εντός ενός περιβάλλοντος «μακράς ουράς» (96,5\% διαλείποντα είδη, βλ. Ενότητα 4.6) αποδείχθηκε καταστροφικά μη αποδοτική. Η απουσία πυκνών σημείων δεδομένων ανά χρονοσειρά εμπόδιζε το reservoir από το να αναπτύξει και να διατηρήσει χρήσιμες δυναμικές (*useful dynamics*) οδηγώντας νομοτελειακά σε υπερπροσαρμογή (*overfitting*) επί του στατιστικού θορύβου.

Η διατήρηση του κώδικα επιτρέπει, φυσικά, τη μελλοντική επανεκτίμηση, ιδιαίτερα όσον αφορά την ανάπτυξη εξειδικευμένων, προσαρμοσμένων ανά κατηγορία δεξαμενών (*category-conditioned reservoirs*), όπου η άθροιση των δεδομένων (*aggregation*) ενδέχεται να μετριάσει το πρόβλημα της ακραίας σποραδικότητας.

5.6.13.2 Μείωση διαστασιμότητας μέσω Ανάλυσης Κύριων Συνιστωσών (Principal Component Analysis - PCA)

Η Ανάλυση Κύριων Συνιστωσών (PCA) εφαρμόστηκε στο υποσύνολο των κυλιόμενων στατιστικών (όπως `QtySum`, `QtyMean`) με σκοπό τη συμπίεση της πληροφορίας σε έναν μικρότερο, πυκνό χώρο 50 σημαντικών συνιστωσών.

Γιατί απενεργοποιήθηκε (`use_pca_features=False`):

- Οριακό όφελος: Η μείωση του σφάλματος `WAPE` ήταν αμελητέα ($< 0{,}2\%$) και όχι συστηματική σε όλους τους ορίζοντες.
- Απώλεια ερμηνευσιμότητας (*Interpretability/Explainability*): Σε ένα επιχειρησιακό σύστημα, η δυνατότητα των μετόχων να κατανοούν, γιατί το μοντέλο έδωσε μια συγκεκριμένη πρόβλεψη, είναι κρίσιμη. Από τη μία, χαρακτηριστικά όπως το `SalesDays_180d` έχουν άμεση φυσική ερμηνεία, ενώ από την άλλη, οι συνιστώσες PCA στερούνται νοήματος καθιστώντας τη διάγνωση και την αποδοχή του μοντέλου δυσκολότερη.

5.6.14 Υπολογισμός Αποθέματος Ασφαλείας (*Safety Stock*) και Παραδοτέα Συστήματος

Ως άμεσο επιχειρησιακό παραδοτέο, το σύστημα υπολογίζει το **απόθεμα ασφαλείας (*safety stock*)** και τα σημεία αναπαραγγελίας (*reorder points* – ROP) ανά κωδικό είδους (SKU) βασιζόμενο στις κατανομές πρόβλεψης. Ο υπολογισμός αυτός υλοποιεί την κατεύθυνση που έχει τεθεί βιβλιογραφικά για τα ολοκληρωμένα πλαίσια πρόβλεψης και αποθεμάτων [1], [94], [116] και αποτελεί τη γέφυρα μεταξύ του αλγοριθμικού συστήματος και του ευρύτερου εργαλείου υποστήριξης αποφάσεων της επιχείρησης. Η λογική υλοποιείται στη συνάρτηση `compute_safety_stock_table()` και εκτελείται μετά την ολοκλήρωση της πρόβλεψης και του αναδρομικού ελέγχου (*backtesting*), ώστε να έχει πρόσβαση τόσο στις μελλοντικές εκτιμήσεις όσο και στα εκτός-δείγματος σφάλματα. Με αυτόν τον τρόπο, η Προσέγγιση 3 ενσωματώνει τη φιλοσοφία της πρόβλεψης με επίγνωση αποθέματος (*inventory-aware forecasting*) [2], [131] απαντώντας στο αντίστοιχο βιβλιογραφικό κενό (βλ. ενότητα 3.8).

5.6.14.1 Επίλυση χρόνου προμήθειας (*lead time*) ανά προμηθευτή

Το σύστημα αποφεύγει τη χρήση ενός ενιαίου, αυθαίρετου χρόνου προμήθειας για όλα τα SKUs. Για κάθε κωδικό, αντλείται το αναγνωριστικό (`Supplier_ID`) και αναζητείται ο αντίστοιχος προμηθευτής στο μητρώο μεταδεδομένων (`supplier_meta`). Η αναζήτηση εφαρμόζει ιεραρχικούς κανόνες υποκατάστασης (*hierarchical fallbacks*):

- Ελέγχονται διαδοχικά οι πιθανές στήλες χρόνου προμήθειας (`Lead_Time_Days`, `LeadTimeDays`, `Supplier_Lead_Time_Days`).
- Επιλέγεται η πρώτη αριθμητικά έγκυρη και θετική τιμή.
- Σε περίπτωση απουσίας δεδομένων, το σύστημα καταφεύγει σε μια προεπιλεγμένη τιμή (*default*) 30 ημερών.

Η ακριβής πηγή του χρόνου προμήθειας (`Lead_Time_Source`) καταγράφεται στην τελική έξοδο, διασφαλίζοντας την ιχνηλασιμότητα των δεδομένων.

5.6.14.2 Επιλογή ορίζοντα αναφοράς και κλιμάκωση ζήτησης

Αντί να απαιτείται η εκπαίδευση ενός ξεχωριστού μοντέλου για τον ακριβή χρόνο προμήθειας κάθε προμηθευτή, το σύστημα αξιοποιεί τις ήδη διαθέσιμες προβλέψεις πολλαπλών οριζόντων. Ο πλησιέστερος χρονικός ορίζοντας H^* στο ζητούμενο *lead time* L επιλέγεται ως:

$$H^* = \arg\min_{H \in \mathcal{H}} |H - L| \tag{5.39}$$

όπου $\mathcal{H} = \{15,30,45,60,75,90,180,365\}$ ίναι το σύνολο των διαθέσιμων οριζόντων του μοντέλου. Η πρόβλεψη της ζήτησης κατά τη διάρκεια του χρόνου προμήθειας ($\hat{D}_L$) υπολογίζεται μέσω γραμμικής κλιμάκωσης:

$$\hat{D}_L = \hat{Q}_{H^*} \times \frac{L}{H^*} \tag{5.40}$$

όπου $\hat{Q}_{H^*}$ είναι η τελική (βαθμονομημένη) πρόβλεψη ποσότητας για τον επιλεγμένο ορίζοντα. Η γραμμική αυτή κλιμάκωση θεωρείται ασφαλής και επαρκής, δεδομένου ότι το σύνολο $\mathcal{H}$ καλύπτει με μεγάλη πυκνότητα το τυπικό εύρος των *lead times* (15–90 ημέρες).

5.6.14.3 Πρωτεύουσα μέθοδος: Μη-παραμετρικό Σημείο Αναπαραγγελίας (ROP) από Διάστημα Πρόβλεψης

Η κύρια μέθοδος του συστήματος υλοποιεί άμεσα τη μη-παραμετρική προσέγγιση [103] (όπως αναλύθηκε στο θεωρητικό Κεφάλαιο 2) υπολογίζοντας το Σημείο Αναπαραγγελίας βάσει του εμπειρικού διαστήματος πρόβλεψης:

$$\text{ROP}_{SKU} = \hat{q}_\alpha(D_L \mid \mathbf{x}) \approx \hat{U}_{H^*} \times \frac{L}{H^*} \tag{5.41}$$

όπου $\hat{U}_{H^*}$ είναι το ανώτατο όριο του διαστήματος πρόβλεψης (`Forecasted_Upper`) του βέλτιστου μοντέλου για τον ορίζοντα H^*. Το όριο αυτό αντιστοιχεί στο επιθυμητό α-ποσοστημόριο (π.χ. $\alpha = 0.90$) της κατανομής της ζήτησης. Η μέθοδος αυτή δεν προϋποθέτει κανονικότητα στην κατανομή των σφαλμάτων, που είναι ένα κρίσιμο πλεονέκτημα για τη διαλείπουσα ζήτηση, και εξαρτάται αποκλειστικά από την ποιότητα βαθμονόμησης των διαστημάτων πρόβλεψης.

5.6.14.4 Εναλλακτική μέθοδος: Γκαουσιανή Προσέγγιση (Gaussian Fallback)

Για SKUs όπου το ανώτατο όριο του διαστήματος δεν είναι διαθέσιμο, το σύστημα εφαρμόζει την κλασική παραμετρική (Γκαουσιανή) προσέγγιση [142], [153]:

$$\text{ROP}_{SKU} = \hat{D}_L + z_\alpha \cdot \hat{\sigma}_{lead} \tag{5.42}$$

όπου:

- z_α (Z-Score): Η κρίσιμη τιμή της τυπικής κανονικής κατανομής για το επιθυμητό Επίπεδο Εξυπηρέτησης Πελατών (*Service Level*) a. Για παράδειγμα, για στόχο ικανοποίησης της ζήτησης στο 90, η τιμή ορίζεται ως $z \approx 1{,}282$.
- $\hat{\sigma}_{lead}$: το προβλεπόμενο τυπικό σφάλμα της ζήτησης κατά τον χρόνο προμήθειας, το οποίο υπολογίζεται ως:

$$\hat{\sigma}_{lead} = \text{RMSE}_{base} \times \sqrt{\frac{L}{H^*}} \tag{5.43}$$

Το RMSE_{base} εκτιμάται από τα εκτός-δείγματος (*out-of-sample*) κατάλοιπα του αναδρομικού ελέγχου. Για την αποφυγή ακραίων τιμών σε κωδικούς με ελλιπές ιστορικό, εφαρμόζεται ιεραρχική αναζήτηση με σταθμισμένη συρρίκνωση (*shrinkage blending*):

Πίνακας 5.19: Ιεραρχική αναζήτηση υπολογισμού `RMSE`

Επίπεδο	**Πηγή Σφάλματος**	**Κλειδί Αναζήτησης**
1 (Ακριβέστερο)	Επίπεδο κωδικού (SKU)	`Product_ID`
2	Επίπεδο μητρικού κωδικού	`Product_MasterCode`
3	Επίπεδο προϊοντικής κατηγορίας	`Product_Hierarchy_Level3`
4 (Εφεδρικό)	Καθολικό σφάλμα ορίζοντα	-

Ο τελικός συνδυασμός υπολογίζεται με συντελεστή συρρίκνωσης $c = 3{,}0$:

$$\hat{\sigma}_{blended} = \frac{n}{n+3}\,\hat{\sigma}_{local} + \frac{3}{n+3}\,\hat{\sigma}_{global} \tag{5.44}$$

όπου n είναι ο αριθμός των παρατηρήσεων στο τοπικό επίπεδο. Η τεχνική αυτή λειτουργεί ως Μπεϋζιανή εκ των προτέρων κατανομή (*Bayesian prior*) 3 «εικονικών παρατηρήσεων» αποτρέποντας την υπερπροσαρμογή.

5.6.14.5 Υπολογισμός Αποθέματος Ασφαλείας

Ανεξαρτήτως της μεθόδου που χρησιμοποιήθηκε για το Σημείο Αναπαραγγελίας (ROP), το τελικό απόθεμα ασφαλείας (SS) προκύπτει αφαιρώντας την αναμενόμενη μέση ζήτηση:

$$SS_{SKU} = \max\left(0,\ \text{ROP}_{SKU} - \hat{D}_L\right) \tag{5.45}$$

Ο υπολογισμός αυτός εκτελείται δυναμικά για πολλαπλά επίπεδα εξυπηρέτησης (π.χ. 80%, 90%, 95%), εξάγοντας τα αποτελέσματα στο αρχείο `safety_stock.csv`. Το αρχείο αυτό καταγράφει αναλυτικά τον ορίζοντα αναφοράς, την πηγή του `lead time`, το κλιμακωμένο `RMSE` και τη μέθοδο (Μη-παραμετρική ή Γκαουσιανή).

Αυτή η διαδικασία αποτελεί το κομβικό σημείο σύζευξης της αλγοριθμικής πρόβλεψης με τη διαχείριση αποθεμάτων [24], [95]. Οι υπεύθυνοι προμηθειών (*planners*) λαμβάνουν έτοιμες προτάσεις που ενσωματώνουν τόσο τη βέλτιστη εκτίμηση της ζήτησης (από το μοντέλο συνόλου), όσο και την αξιόπιστη ποσοτικοποίηση της αβεβαιότητας αντιμετωπίζοντας με μαθηματική αυστηρότητα τις ιδιαιτερότητες της διαλείπουσας ζήτησης.

5.6.14.6 Παραδοτέα Συστήματος και Επιχειρησιακή Σημασία

Όπως έχει τονιστεί επανειλημμένα στην παρούσα εργασία, ο απώτερος σκοπός είναι η παραγωγή βέλτιστων προτάσεων για παραγγελίες αναπλήρωσης. Το τρέχον υποσύστημα, το οποίο αναλύθηκε

διεξοδικά στο παρόν κεφάλαιο, αναλαμβάνει την αξιόπιστη πρόβλεψη της μελλοντικής ζήτησης ανά κωδικό προϊόντος και ανά χρονικό ορίζοντα.

Τα αποτελέσματα αυτά προορίζονται να τροφοδοτήσουν, στη συνέχεια, μια ανεξάρτητη **εφαρμογή παραγωγής προτάσεων παραγγελιών**, η οποία αποτελεί το τελικό περιβάλλον αλληλεπίδρασης με τον χρήστη. Ως εκ τούτου, οι έξοδοι του συστήματος πρόβλεψης οφείλουν να ακολουθούν μια τυποποιημένη μορφή, ώστε να μπορούν να ενσωματωθούν ομαλά. Τα τελικά παραγόμενα αρχεία αφορούν τις προβλέψεις μελλοντικής ζήτησης και τα απαιτούμενα αποθέματα ασφαλείας αναλυμένα σε τρία διακριτά ιεραρχικά επίπεδα: (α) επίπεδο τελικού κωδικού προϊόντος, (β) επίπεδο κύριου/μητρικού κωδικού (*Master Code*) και (γ) επίπεδο κατηγορίας προϊόντων. Συνολικά παράγονται έξι (6) αρχεία μορφής `CSV`, τα οποία συνοψίζονται στον Πίνακα 5.20. Σημειώνεται ότι τα αρχεία επιπέδου μητρικού κωδικού και κατηγορίας μέσω άθροισης και ομαδοποίησης τν περιεχομένων των αρχείων επιπέδου κωδικού προϊόντος. Αυτό ισχύει και για τις δύο κατηγορίες παραγόμενων αρχείων.

Πίνακας 5.20: Τα τελικά παραδοτέα αρχεία του συστήματος πρόβλεψης

Τύπος Δεδομένων	**Επίπεδο Κωδικού Προϊόντος (SKU)**	**Επίπεδο *Master Code***	**Επίπεδο Κατηγορίας (*Hierarchy*)**
Πρόβλεψη Μελλοντικής Ζήτησης	`by_product_forecast.csv`	`by_master_forecast.csv`	`by_hierarchy_forecast.csv`
Αποθέματα Ασφαλείας	`safety_stock.csv`	`safety_stock_by_master.csv`	`safety_stock_by_hierarchy.csv`

Η παραγωγή των συγκεκριμένων συνόλων δεδομένων αποτελεί το σημείο σύγκλισης και ολοκλήρωσης του συστήματος πρόβλεψης. Η εξωτερική εφαρμογή παραγγελιών θα λαμβάνει αυτά τα παραδοτέα (πρόβλεψη ζήτησης και αποθέματα ασφαλείας) και θα τα συνδυάζει δυναμικά με τα μητρώα της επιχείρησης, τα οποία περιλαμβάνουν λίστες ειδών, καταλόγους προμηθευτών και μοναδικά ζεύγη «είδους-προμηθευτή» (λαμβάνοντας υπόψη τον βασικό προμηθευτή καθώς και τυχόν εφεδρικούς-εναλλακτικούς).

Μέσω της τελικής διεπαφής, ο χρήστης θα έχει τη δυνατότητα να επιλέγει τον επιθυμητό ορίζοντα κάλυψης φιλτράροντας τα δεδομένα ανά κωδικό, κατηγορία ή συγκεκριμένο προμηθευτή, ώστε να παράγονται άμεσα οι βέλτιστες προτάσεις αναπλήρωσης. Η πλήρης αρχιτεκτονική και η αναλυτική περιγραφή λειτουργίας αυτής της εφαρμογής παρουσιάζονται εκτενώς στο επόμενο κεφάλαιο (Κεφάλαιο 6).

5.6.15 Αποτελέσματα αξιολόγησης (Evaluation results)

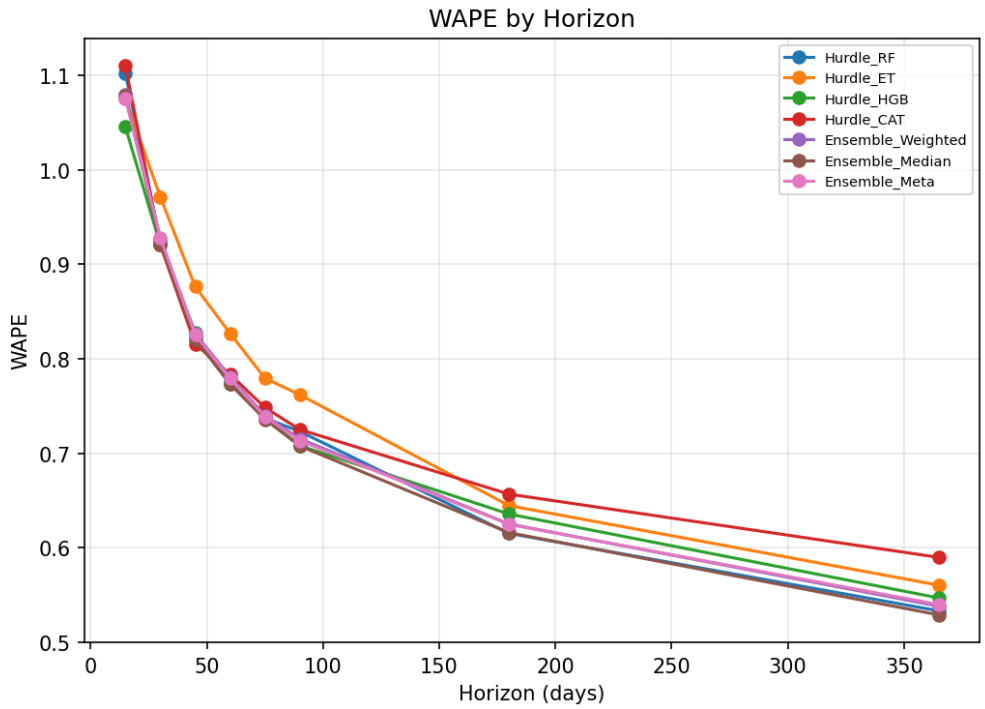
Στην ενότητα αυτή παρουσιάζονται και αναλύονται τα αποτελέσματα της αξιολόγησης του συστήματος στο σύνολο δοκιμής (*test set*) για το έτος 2025. Η αξιολόγηση πραγματοποιείται σε πολλαπλά επίπεδα και εξετάζει τόσο τις καθολικές μετρικές όσο και την εξειδικευμένη επίδοση του συστήματος σε κρίσιμες επιχειρησιακές υπό-ομάδες προϊόντων.

5.6.15.1 Συνολικές μετρικές αξιολόγησης

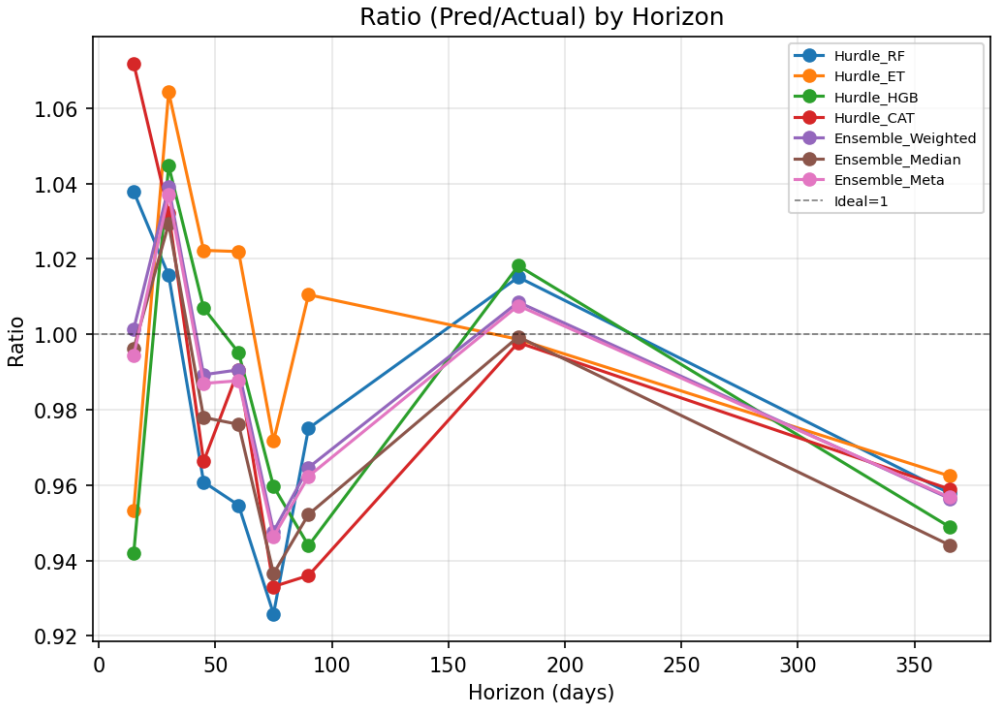
Ο Πίνακας 5.21 συγκεντρώνει τα κύρια αποτελέσματα για το σφάλμα `WAPE` και τον λόγο συνολικής μάζας (`Ratio`) ανά χρονικό ορίζοντα και μοντέλο. Επίσης, στα Σχήματα 5.8 και 5.9. παρουσιάζονται γραφικά τα περιεχόμενα του πίνακα.

Πίνακας 5.21: Συνολική απόδοση των μοντέλων (WAPE / Ratio) στο σύνολο δοκιμής του 2025.

Horizon	RF WAPE / Ratio	ET WAPE / Ratio	HGB WAPE / Ratio	CAT WAPE / Ratio	Ensemble_ Weighted WAPE / Ratio	Ensemble_ Meta WAPE / Ratio	Croston_ SBA WAPE / Ratio	TSB WAPE / Ratio
15d	1,102 / 1,038	1,075 / 0,953	1,045 / 0,942	1,111 / 1,072	1,079 / 1,001	**1,076 / 0,994**	36,5 / 36,6	44,6 / 45,3
30d	0,924 / 1,016	0,972 / 1,064	0,921 / 1,045	0,921 / 1,032	0,927 / 1,039	**0,928 / 1,037**	32,3 / 32,6	39,6 / 40,5
45d	0,824 / 0,961	0,876 / 1,022	0,828 / 1,007	0,816 / 0,966	0,827 / 0,989	**0,826 / 0,987**	30,6 / 31,0	37,6 / 38,4
60d	0,775 / 0,955	0,827 / 1,022	0,773 / 0,995	0,784 / 0,991	0,781 / 0,991	**0,779 / 0,988**	30,2 / 30,8	37,2 / 38,1
75d	0,737 / 0,926	0,779 / 0,972	0,738 / 0,960	0,748 / 0,933	0,740 / 0,948	**0,739 / 0,946**	28,4 / 28,9	34,9 / 35,8
90d	0,723 / 0,975	0,762 / 1,011	0,708 / 0,944	0,725 / 0,936	0,715 / 0,965	**0,713 / 0,962**	29,0 / 29,6	35,8 / 36,6
180d	0,615 / 1,015	0,645 / 0,999	0,636 / 1,020	0,656 / 1,000	0,625 / 1,009	**0,625 / 1,009**	29,6 / 30,2	36,6 / 37,5
365d	0,533 / 0,958	0,560 / 0,962	0,547 / 0,949	0,590 / 0,959	0,538 / 0,956	**0,540 / 0,957**	30,1 / 30,8	37,3 / 38,1



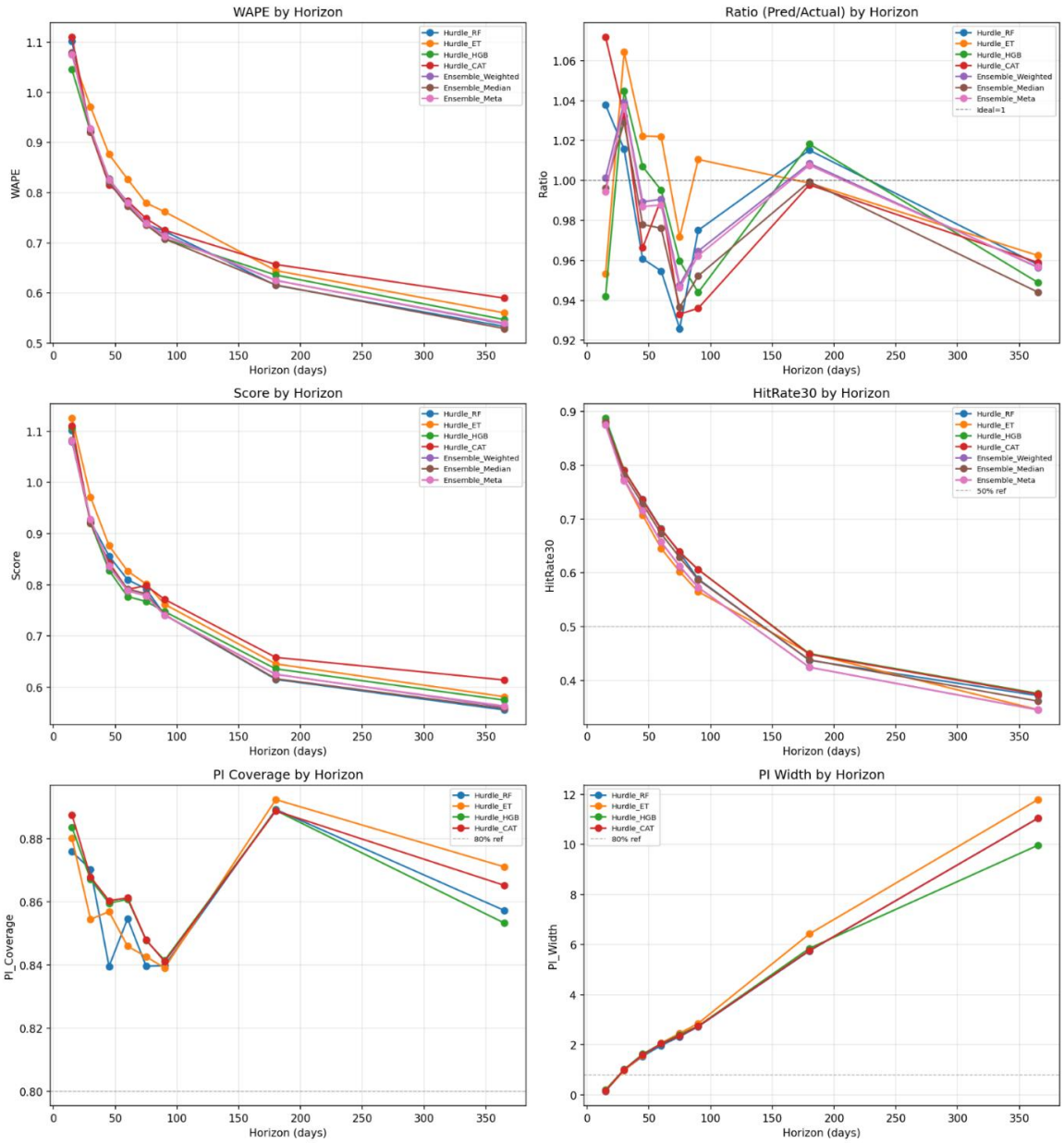
Σχήμα 5.8: Σφάλμα WAPE ανά ορίζοντα



Σχήμα 5.9: `Ratio` ανά ορίζοντα

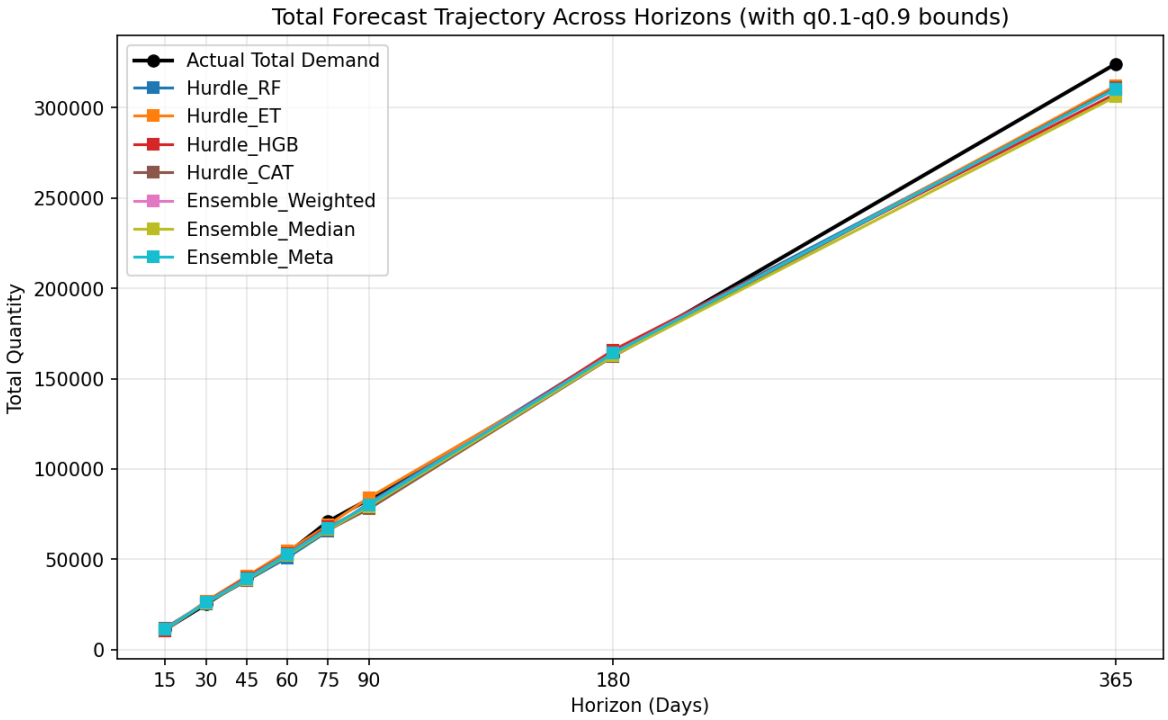
Κατόπιν ανάλυσης των ανωτέρω αποτελεσμάτων, προκύπτουν τα κάτωθι κρίσιμα συμπεράσματα:

- **Φθίνουσα πορεία του WAPE:** Η συστηματική μείωση του σφάλματος από ~1,08 στις 15 ημέρες σε ~0,54 στις 365 ημέρες αναδεικνύει την ευεργετική επίδραση της χρονικής άθροισης (*temporal aggregation*). Η «εξομάλυνση» της διαλείπουσας ζήτησης σε μεγαλύτερα χρονικά παράθυρα μειώνει αισθητά τη διακύμανση καθιστώντας το μακροπρόθεσμο σήμα πιο προβλέψιμο.
- **Επιτυχία Βαθμονόμησης:** Όλα τα προτεινόμενα μοντέλα διατηρούν τον λόγο μάζας (`Ratio`) αυστηρά εντός του εύρους [0,92, 1,07], γεγονός που επιβεβαιώνει περίτρανα την ευστάθεια και την αναγκαιότητα των σταδίων *calibration* και *meta-learning*.
- **Χαώδης Υπεροχή έναντι Κλασικών Μοντέλων:** Τα εξαιρετικά υψηλά, μη-αποδεκτά σφάλματα των μεθόδων `Croston_SBA` και `TSB` οφείλονται σε μεγάλο βαθμό στο φαινόμενο των «αδρανών κωδικών» (*zombie-SKUs*). Τα μοντέλα Hurdle αντιμετωπίζουν αυτό το φαινόμενο αποτελεσματικά μέσω του αυστηρού σταδίου ταξινόμησης (*classifier*).



Σχήμα 5.10: Συγκεντρωτική απόδοση των μοντέλων ανά ορίζοντα

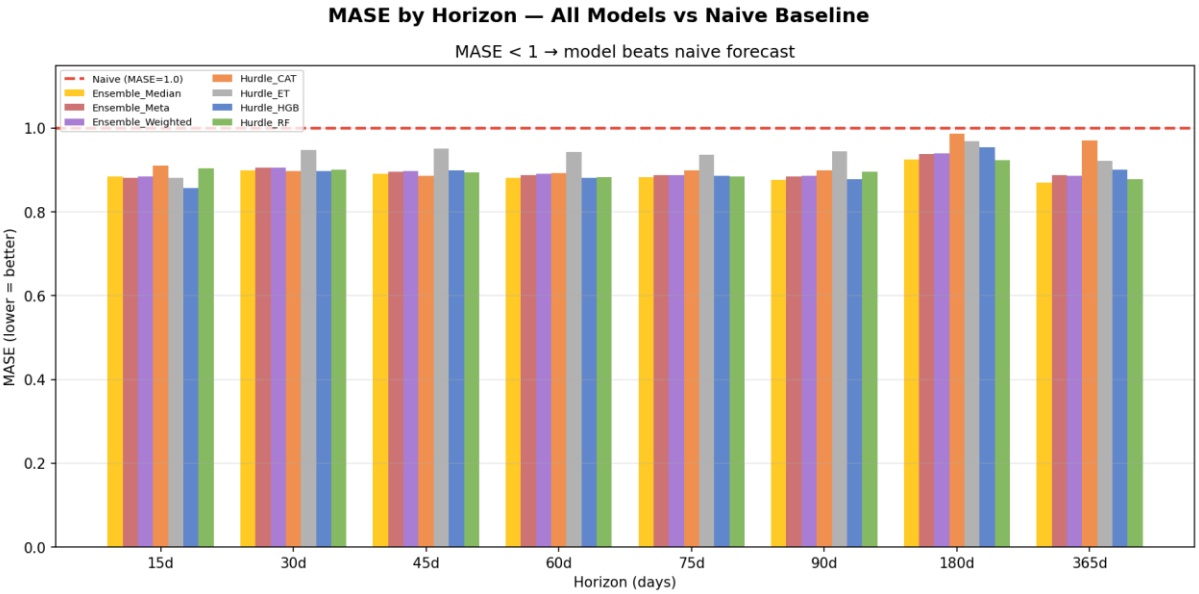
Στο Σχήμα 5.11 φαίνεται η συγκριτική απόδοση κάθε μοντέλου σε σχέση με τις πραγματικές τιμές πωλήσεων.



Σχήμα 5.11: Συγκριτική απόδοση κάθε μοντέλου σε σχέση με τις πραγματικές τιμές

Ανάλυση `MASE` (Mean Absolute Scaled Error)

Η μετρική `MASE` κανονικοποιεί το απόλυτο σφάλμα (`MAE`) ως προς αυτό μιας αφελούς (naive) εποχικής πρόβλεψης. Ως αφελή θεωρούμε την προσέγγιση, κατά την οποία η εκάστοτε τρέχουσα τιμή ζήτησης ισούται με την πλέον πρόσφατη αντίστοιχη ιστορική τιμή. Σκορ μετρικής $< 1{,}0$ υποδηλώνει υπεροχή του τρέχοντος συστήματος έναντι της αφελούς βάσης. Όπως φαίνεται στο Σχήμα 5.12, όλα τα μοντέλα hurdle επιτυγχάνουν MASE < 1,0, με το `Ensemble_Meta` να κυμαίνεται μεταξύ 0,882 και 0,939.



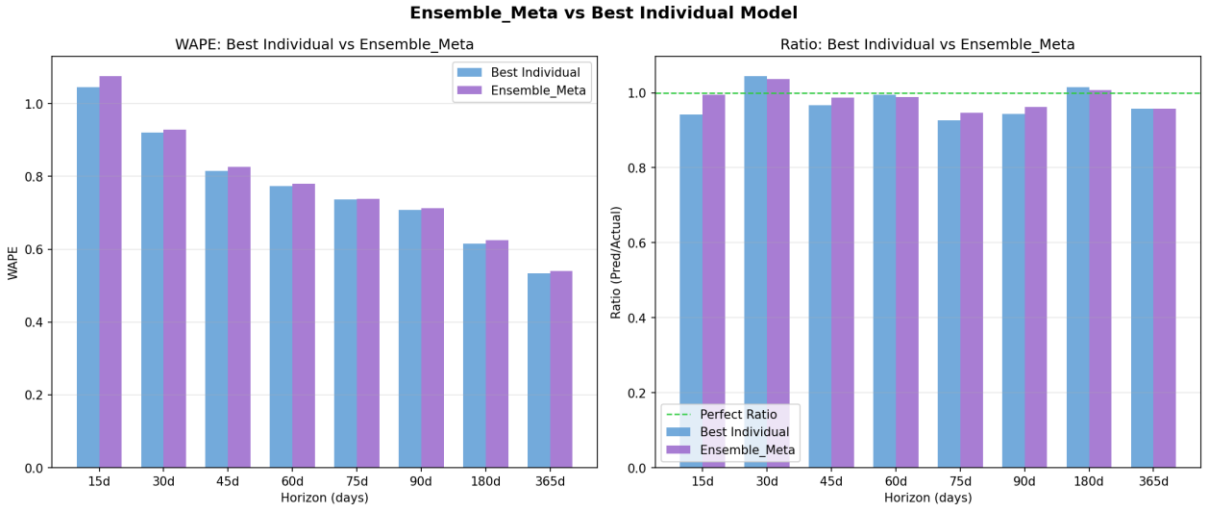
Σχήμα 5.12: Απόδοση `MASE` ανά ορίζοντα

5.6.15.2 Μοντέλο Μέτα-Μάθησης (*Meta-Learner*) έναντι Βασικών Μοντέλων

Η συγκριτική αξιολόγηση του `Ensemble_Meta` έναντι του εκάστοτε βέλτιστου βασικού μοντέλου αποκαλύπτει την πραγματική επιχειρησιακή αξία της συσσωρευτικής γενίκευσης. Όπως αποτυπώνεται στα γραφήματα του Σχήματος 5.13, το μοντέλο μετα-μάθησης δεν αναδεικνύεται απαραίτητα ως το απολύτως κορυφαίο σε κάθε μεμονωμένο ορίζοντα, υπολειπόμενο ελαφρώς σε `WAPE` έναντι του εκάστοτε τοπικού βέλτιστου (*local optimum*) αλγορίθμου.

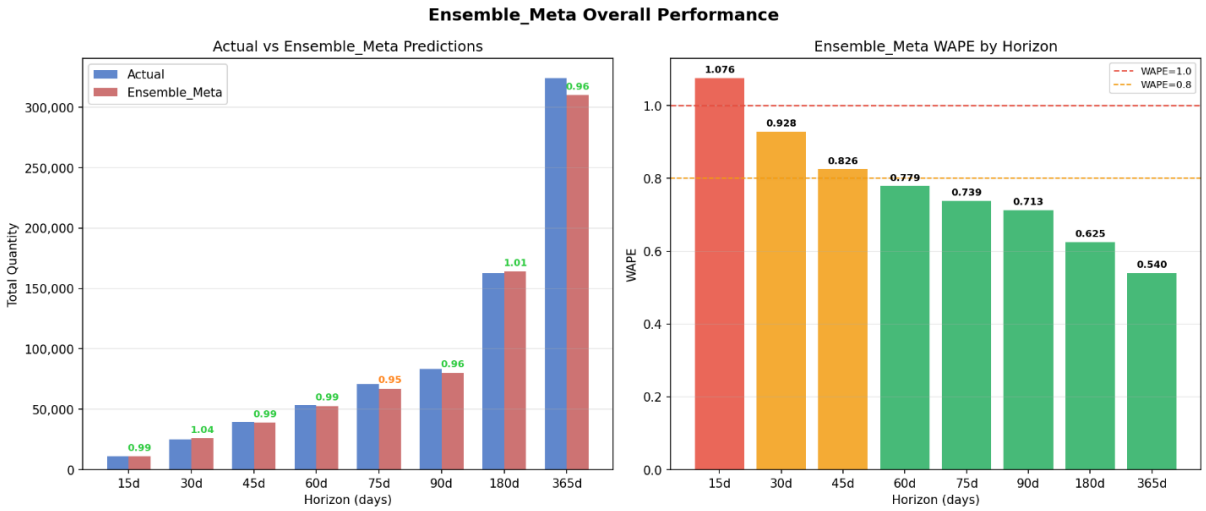
Ωστόσο, το ουσιαστικό του όφελος δεν εντοπίζεται στην επίτευξη της απόλυτης μέγιστης ακρίβειας σε ένα συγκεκριμένο χρονικό παράθυρο, αλλά στην **απαράμιλλη διαχρονική του ευστάθεια**. Ενώ τα μεμονωμένα βασικά μοντέλα (όπως το `Random Forest` ή το `CatBoost`) παρουσιάζουν έντονες διακυμάνσεις στην απόδοσή τους ανάλογα με τον ορίζοντα πρόβλεψης, το `Ensemble_Meta` καταφέρνει να παραμένει σταθερά αξιόπιστο («πάντα καλό») σε όλες τις περιπτώσεις.

Εάν εξεταστεί ο συνολικός μέσος όρος του σφάλματος σε όλους τους ορίζοντες, το μοντέλο μετα-μάθησης αναδεικνύεται αδιαμφισβήτητα ως η κορυφαία επιλογή. Αποφεύγοντας την απόλυτη εξάρτηση από μία μόνο αρχιτεκτονική, η πρόβλεψη «διασώζεται» σε περίπτωση τοπικής ή συγκυριακής αστοχίας ενός μεμονωμένου αλγορίθμου. Παράλληλα, πετυχαίνει εξαιρετική διόρθωση της συνολικής μάζας ζήτησης, διατηρώντας τον δείκτη `Ratio` σταθερά κοντά στην ιδανική τιμή της μονάδας (1,0). Κατά συνέπεια, προσφέρει την πλέον ισορροπημένη και ασφαλή λύση για ένα παραγωγικό σύστημα διαχείρισης αποθεμάτων.



Σχήμα 5.13: Σύγκριση `Ensemble_Meta` με το εκάστοτε βέλτιστο μοντέλο

Επίσης, στο Σχήμα 5.14 παρουσιάζεται η συνολική απόδοση του μοντέλου μετα-μάθησης ως προς τις πραγματικές τιμές του συνόλου δεδομένων.



Σχήμα 5.14: Απόδοση μοντέλου μετα-μάθησης

5.6.15.3 Προϊόντα Υψηλού Τζίρου (Top-20 *by Revenue*)

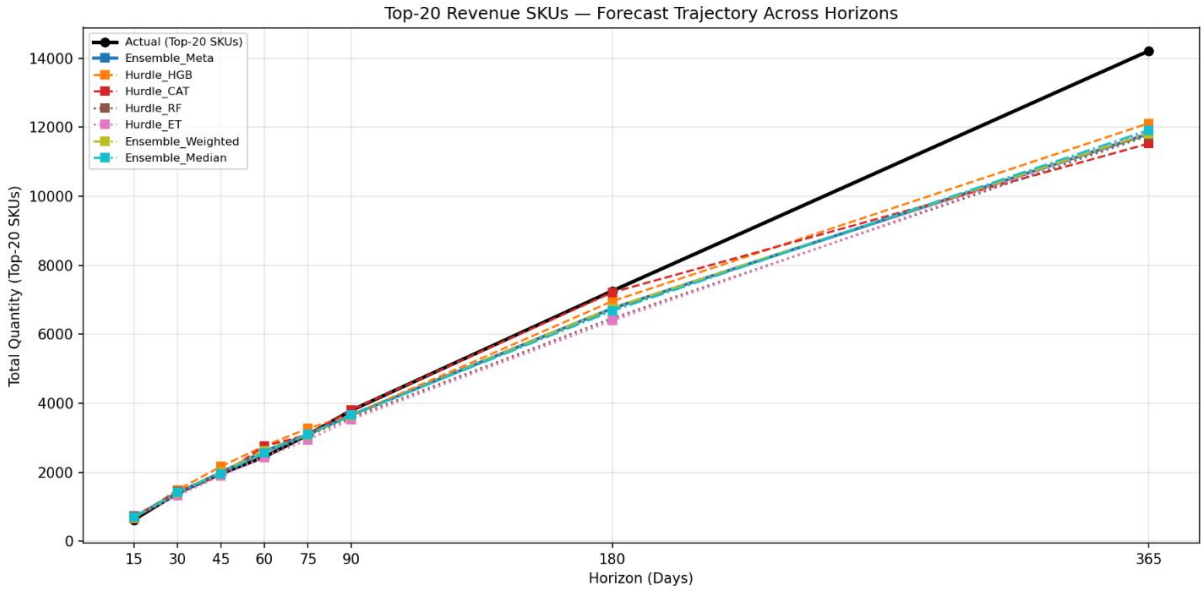
Στην τρέχουσα ενότητα παρατίθεται μια εκτενής ανάλυση της απόδοσης των πλέον κρίσιμων από επιχειρησιακής άποψης κωδικών ειδών. Ο Πίνακας 5.22 παρουσιάζει την απόδοση του συστήματος (`Ensemble_Meta`) αποκλειστικά όσον αφορά τα 20 κορυφαία προϊόντα βάσει ιστορικού τζίρου.

Πίνακας 5.22: Απόδοση του `Ensemble_Meta` στο υποσύνολο Top-20 (Έτος 2025)

Ορίζοντας	**Πραγματικός Τζίρος Top-20 (€)**	**Πρόβλεψη `Ensemble_Meta` (Τεμ.)**	**Πραγματική Ζήτηση (Τεμ.)**	**`Ratio` (Meta/Actual)**
15d	13.211	635	648	**0,980**
30d	29.104	1.113	1.236	**0,900**
45d	44.447	1.663	1.687	**0,986**
60d	54.684	2.426	2.405	**1,009**
75d	65.518	2.898	3.001	**0,966**
90d	78.490	3.547	3.755	**0,944**
180d	148.588	6.498	7.196	**0,903**
365d	296.284	10.931	13.981	**0,782**

Η παρατηρούμενη υπο-πρόβλεψη στον ετήσιο ορίζοντα (`Ratio` $= 0{,}782$) οφείλεται, κυρίως, σε έναν κωδικό (`103-A103`), ο οποίος εμφάνισε εκρηκτική αύξηση πωλήσεων το 2025 ($\times 2{,}7$ έναντι του ιστορικού). Χωρίς αυτόν, το `Ratio` των υπολοίπων 19 προϊόντων διαμορφώνεται στο $\mathbf{1,099}$.

Στο Σχήμα 5.15 αποτυπώνεται η εξέλιξη του συνολικού `ratio` των top-20 ειδών του 2025 βάσει των πραγματικών πωλήσεων.



Σχήμα 5.15: Εξέλιξη συνολικού `ratio` των Top-20 ειδών εντός του 2025

Στο σημείο αυτό, πρέπει να τονιστεί ότι η τρέχουσα προσέγγιση είναι αμιγώς ερευνητική. Δηλαδή, το μοντέλο δεν έλαβε ουδεμία εξωτερική πληροφορία σχετικά με επιχειρηματικές αποφάσεις (π.χ. διαφημιστικές καμπάνιες, ακυρώσεις προμηθευτών). Σε ένα παραγωγικό σύστημα, τέτοια γεγονότα κωδικοποιούνται ως επιπλέον χαρακτηριστικά. Στην ανάλυση που ακολουθεί, συμπεραίνεται εύκολα ότι όλοι οι «προβληματικοί» κωδικοί ανάμεσα στους Top-20 είναι είδη που είτε εισήχθηκαν στον κατάλογο εντός του 2025 είτε παρατηρήθηκε πάρα πολύ έντονη αυξομείωση στις πωλήσεις τους, η οποία ευθυνόταν σε εμπορικές συμφωνίες.

Η πρόκληση της μεταβλητότητας (Σύνδεση με το κεφάλαιο 4)

Προτού προχωρήσουμε στην ανάλυση των μετρικών κατάταξης, είναι απαραίτητο να ανακαλέσουμε την Εικόνα 4.7 (*heatmap* ετήσιου τζίρου Top-20) από την Ενότητα 4.3.4, στην οποία αποτυπώνεται η ιστορική κατανομή τζίρου των κορυφαίων προϊόντων. Όπως καθίσταται σαφές στο εν λόγω γράφημα, η σύσταση της ομάδας των κορυφαίων κωδικών παρουσιάζει ακραία ετήσια μεταβλητότητα και ανακύκλωση (churn).

Είναι χαρακτηριστικό ότι κωδικοί που κυριάρχησαν απόλυτα το 2025 (όπως π.χ. οι `156-05005DPF` και `156-04004`) κατέγραφαν κυριολεκτικά μηδενικές πωλήσεις καθ' όλη τη διάρκεια της περιόδου 2021-2024. Αντίστοιχα, άλλα παραδοσιακά ισχυρά είδη εμφάνισαν απότομη πτώση. Αυτή η δομική αστάθεια του καταλόγου μετατρέπει τον εντοπισμό των "νέων" κορυφαίων ειδών (*cold-start* SKUs) σε ένα εξαιρετικά απαιτητικό έργο για κάθε προγνωστικό σύστημα, καθώς το μοντέλο καλείται να αναγνωρίσει τάσεις χωρίς την ύπαρξη ισχυρού ιστορικού σήματος.

Μετρικές ποιότητας κατάταξης (Ranking quality metrics)

Λαμβάνοντας υπόψη τη δυσκολία που περιγράφηκε άνωθεν, η αξιολόγηση επεκτείνεται πέραν της απλής ακρίβειας πρόβλεψης όγκου. Η ικανότητα του συστήματος να αναγνωρίζει και να ιεραρχεί σωστά τα σημαντικότερα προϊόντα αξιολογείται μέσω εξειδικευμένων δεικτών ιεράρχησης (*Ranking Accuracy*):

Επικάλυψη κωδικών (SKU Overlap (%))

Το ποσοστό των κωδικών που ταυτίστηκαν μεταξύ των πραγματικών και των προβλεπόμενων Top-20.

Κάλυψη Τζίρου (*Revenue Capture* (%))

Το ποσοστό του πραγματικού τζίρου που καλύπτουν τα επιλεγμένα από το μοντέλο ως Top-20 προϊόντα. Αυτός ο δείκτης είναι πιο σημαντικός από επιχειρηματικής σκοπιάς σε σχέση με το SKU overlap. Το γεγονός ότι ένας κωδικός έχει προβλεφθεί στο Top-20 αλλά δεν ανήκει στην πραγματικότητα, δεν συνεπάγεται αυτομάτως ότι έχει χαμηλό τζίρο. Π.χ. *Revenue Capture* 92% σημαίνει ότι, τα 20 SKU που προτείνει ο αλγόριθμος ως κορυφαία, καλύπτουν το 92% του τζίρου των 20 πραγματικά σημαντικότερων.

Κανονικοποιημένο Απομειωμένο Αθροιστικό Κέρδος (Normalized Discounted Cumulative Gain – NDCG@20)

Πρόκειται για μια καθιερωμένη μετρική από το πεδίο της Ανάκτησης Πληροφορίας (Information Retrieval) [171] που αξιολογεί την ποιότητα της σειράς κατάταξης. Η φιλοσοφία της βασίζεται στη λογαριθμική απομείωση (*logarithmic discounting*) της αξίας κάθε θέσης, γεγονός που σημαίνει ότι οι κωδικοί (SKU) που τοποθετούνται στις πρώτες θέσεις της πρόβλεψης έχουν πολλαπλάσια βαρύτητα («μετρούν» πολύ περισσότερο) σε σχέση με αυτούς που βρίσκονται χαμηλότερα στην κατάταξη. Οι μαθηματικοί τύποι υπολογισμού ορίζονται ως εξής:

$$DCG@k = \sum_{i=1}^{k} \frac{rel_i}{log_2(i+1)} \tag{5.46}$$

$$NDCG@k = \frac{DCG@k}{IDCG@k} \tag{5.47}$$

Στην παρούσα υλοποίηση, η μεταβλητή σχετικότητας (rel_i) του κάθε κωδικού ισούται με τον πραγματικό ετήσιο τζίρο του σε Ευρώ (€) για το έτος 2025. Λόγω της φύσης της μεταβλητής (συνεχείς τιμές μεγάλου εύρους), επιλέγεται η γραμμική διατύπωση του αριθμητή αντί της εκθετικής προς αποφυγή υπολογιστικού σφάλματος υπερχείλισης (*numerical overflow*).

Το *IDCG@k* αντιπροσωπεύει το ιδανικό απομειωμένο αθροιστικό κέρδος (Ideal DCG), δηλαδή το μέγιστο δυνατό σκορ που θα επιτυγχανόταν εάν το μοντέλο είχε προβλέψει την απόλυτα τέλεια σειρά κατάταξης βάσει των πραγματικών πωλήσεων.

Η μετρική λαμβάνει τιμές στο κλειστό διάστημα $[0,1]$. Τιμή $\text{NDCG@20} = 1{,}0$ υποδηλώνει ότι τα 20 προϊόντα με τον υψηλότερο πραγματικό τζίρο εντοπίστηκαν από το μοντέλο και κατατάχθηκαν **ακριβώς** στη σωστή φθίνουσα σειρά. Σύμφωνα με τη βιβλιογραφία [172], τιμές άνω του $0{,}8$ θεωρούνται εξαιρετικές για παραγωγικά συστήματα συστάσεων και κατάταξης.

Συντελεστής συσχέτισης κατάταξης Kendall τ

Μη-παραμετρική συσχέτιση κατάταξης που μετρά πόσα ζεύγη SKU έχουν συνεπή σειρά μεταξύ πραγματικής και προβλεπόμενης κατάταξης. Προτιμάται σε σχέση με τον δείκτη συσχέτισης Pearson, γιατί είναι πιο ανθεκτικός σε ακραίες τιμές και κατάλληλος για δεδομένα με συγκεκριμένη κατάταξη, π.χ. χρονική, όπως στην προκειμένη περίπτωση. Υπολογίζεται από τη σχέση

$$\tau = \frac{(\text{concordant pairs}) - (\text{discordant pairs})}{\binom{n}{2}} \tag{5.48}$$

Παίρνει τιμές στο διάστημα *[-1, 1]* και, μάλιστα, με την εξής ερμηνεία:

τ = 1: τέλεια συμφωνία σειράς,

τ = 0: καμία συσχέτιση (τυχαία),

$\tau = -1$: τέλεια αντίστροφη σειρά.

Συντελεστής συσχέτισης κατάταξης Spearman ρ

Μη-παραμετρική συσχέτιση για τη συνοχή της εσωτερικής ιεράρχησης. Υπολογίζεται από τη σχέση

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{5.49}$$

Όπου d_i , η διαφορά κατάταξης μεταξύ πραγματικότητας και πρόβλεψης για κάθε κωδικό προϊόντος i . Παίρνει, και αυτός, τιμές στο διάστημα *[-1, 1]* με παρόμοια ερμηνεία με τον Kendall. Τιμές $\rho > 0.6$ θεωρούνται ισχυρή συσχέτιση κατάταξης.

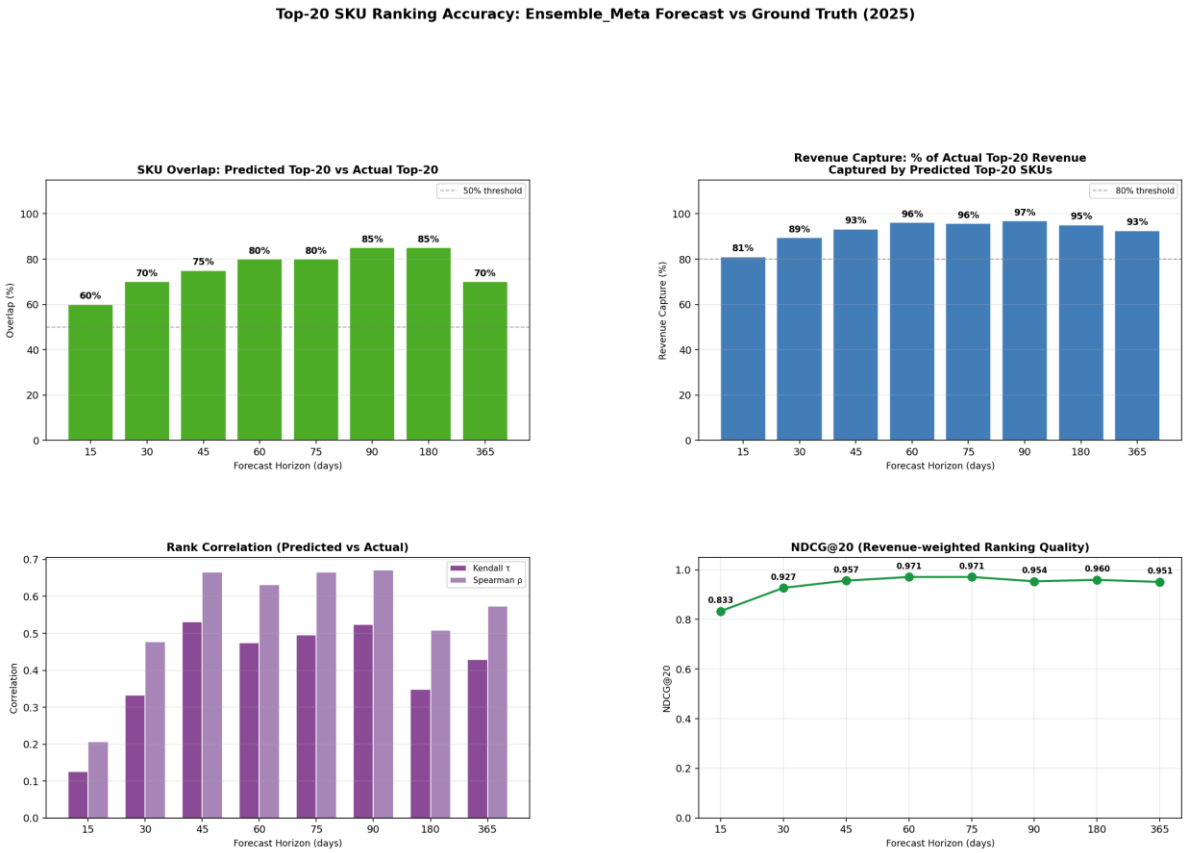
Στον Πίνακα 5.23 παρουσιάζονται οι τιμές των δεικτών αξιολόγησης και στο Σχήμα 5.16. παρατίθενται σχετικά γραφήματα.

Πίνακας 5.23: Μετρικές Ποιότητας Κατάταξης για την Ομάδα Top-20

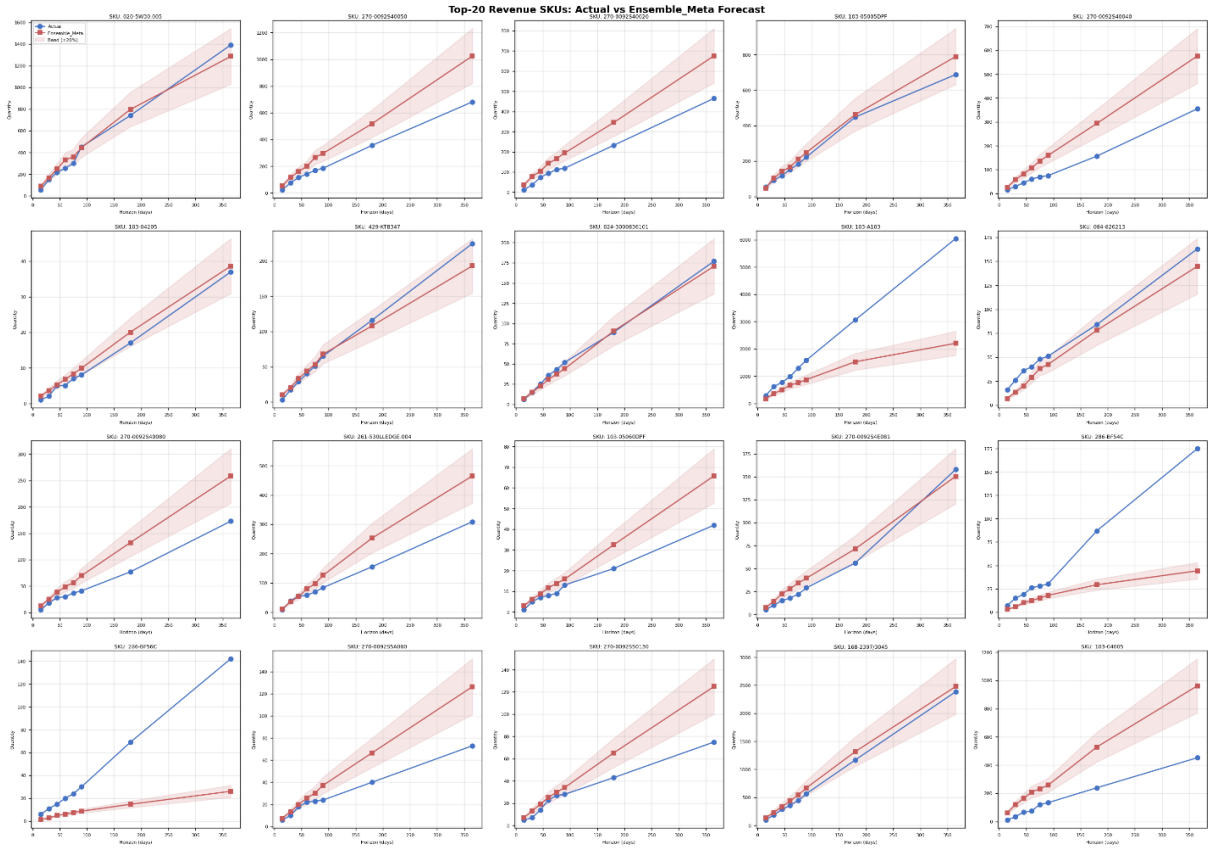
Ορίζοντας	SKU Overlap	Revenue Capture	Kendall τ	Spearman ρ	NDCG@20
15d	12/20 (60%)	81,0%	0,126	0,207	0,833
30d	14/20 (70%)	89,4%	0,332	0,478	0,928
45d	15/20 (75%)	93,2%	0,531	0,666	0,957
60d	16/20 (80%)	96,1%	0,474	0,632	0,971
75d	16/20 (80%)	95,7%	0,496	0,666	0,971
90d	17/20 (85%)	96,8%	0,524	0,672	0,954
180d	17/20 (85%)	95,1%	0,348	0,508	0,960
365d	14/20 (70%)	92,6%	0,429	0,574	0,951

Ερμηνεύοντας τα αποτελέσματα, παρατηρείται ότι η ζώνη με τους βέλτιστους ορίζοντες είναι αυτή μεταξύ 45 και 90 ημερών. Αυτό το διάστημα αναδεικνύεται ως ο «χρυσός κανόνας» για αποφάσεις ανεφοδιασμού επιτυγχάνοντας `NDCG` > *0,95* και `Revenue Capture` > *93%*. Σε αυτό το εύρος, το μοντέλο αξιοποιεί επαρκώς τα ιστορικά χαρακτηριστικά χωρίς να συσσωρεύει υπερβολικό σφάλμα.

Δύο ακόμη σημαντικές παρατηρήσεις είναι ότι η χαμηλότερη ακρίβεια εντοπίζεται στον ορίζοντα των 15 ημερών (60% επικάλυψη) και οφείλεται, κυρίως, σε μεμονωμένες μεγάλες παραγγελίες (*spikes*) που ανεβάζουν τυχαία κάποια SKU στην κορυφή (π.χ. `286-BF54C`: 1 παραγγελία €459, `188-600016800`: 1 παραγγελία €458), τα οποία το μοντέλο ορθώς δεν προβλέπει ως συστηματικά υψηλά. Επίσης, Στον μακροπρόθεσμο ορίζοντα, η πτώση του overlap στο 70% οφείλεται σε κωδικούς με εκρηκτική αύξηση ζήτησης το 2025 που δεν προϋπήρχε στο ιστορικό (π.χ. `325-0065-F0153`), αποτελώντας το φυσικό όριο κάθε προγνωστικού συστήματος. Παρόλα αυτά, βέβαια, αξίζει να σημειωθεί ότι τα 6 κορυφαία είδη έχουν προβλεφθεί με σειρά κατάταξης: `2, 1, 3, 4, 6, 5`. Τυπικά υπάρχει μόλις 33% επικάλυψη αλλά, πρακτικά, η πρόβλεψη θεωρείται επιτυχημένη.



Σχήμα 5.16: Μετρικές Ποιότητας Κατάταξης για την Ομάδα Top-20

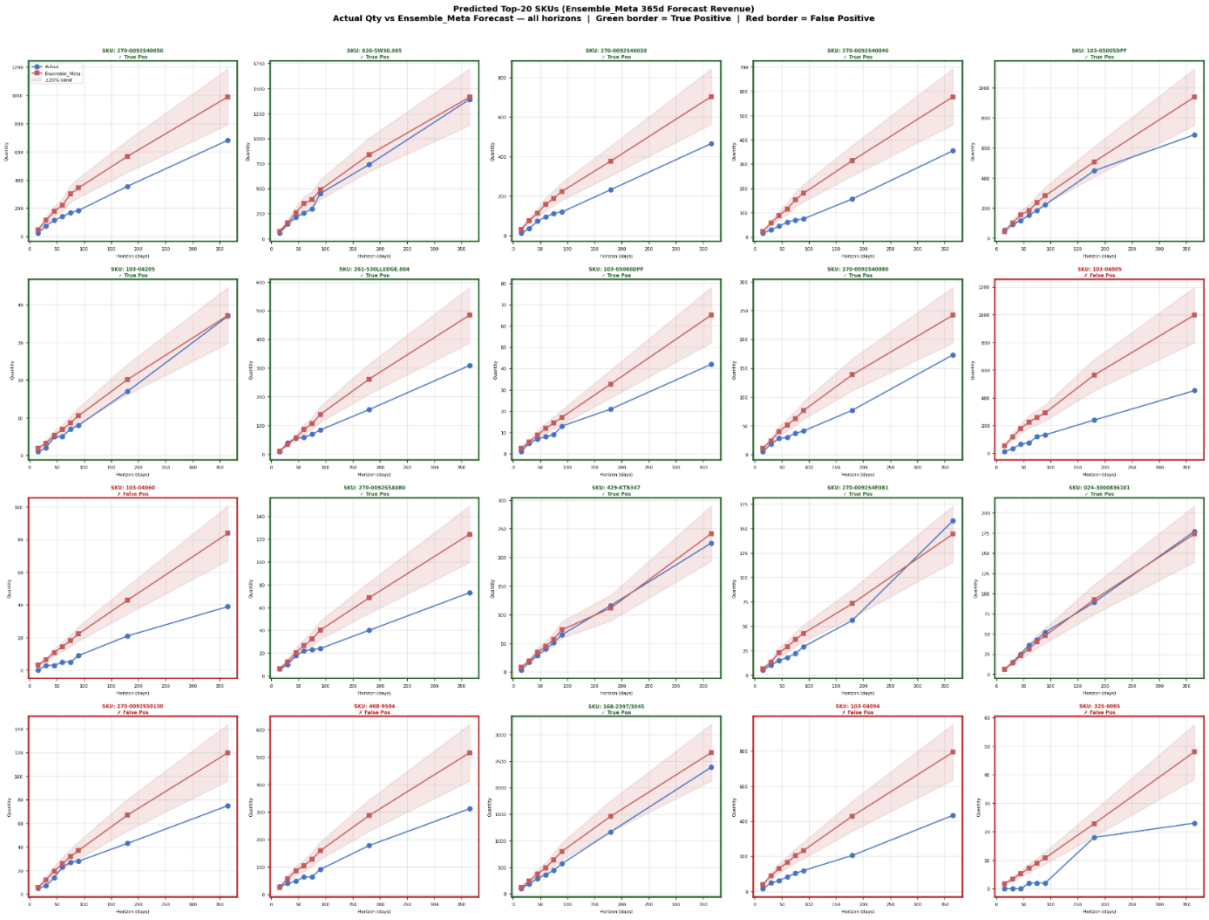


Σχήμα 5.17: Συγκριτική απόδοση των πραγματικών Top-20 κωδικών ειδών

Τα Σχήματα 5.17 και 5.18 παρουσιάζουν αναλυτικά τα Top-20 είδη από δύο διαφορετικές οπτικές γωνίες. Στην πρώτη παρουσιάζονται τα πραγματικά 20 κορυφαία προϊόντα και, παράλληλα, συγκρίνεται η εκάστοτε πρόβλεψη.

Επισημαίνεται ότι όλα τα είδη που αποτυγχάνουν σημαντικά στην πρόβλεψη, σημειώνουν πολύ μεγάλη διαφοροποίηση στο τζίρο του 2025 σε σχέση με αυτόν των προηγούμενων ετών.

Από την άλλη, στη δεύτερη εικόνα φαίνονται τα 20 κορυφαία βάσει πρόβλεψης είδη και γίνεται σε αυτά σύγκριση με τις πραγματικές πωλήσεις. Όσα βρίσκονται σε πράσινο πλαίσιο έχουν καταταχθεί ορθώς στα top-20 (*true positives*), ενώ αυτά με το κόκκινο πλαίσιο είναι λανθασμένες προβλέψεις (*false positives*).



Σχήμα 5.18: Συγκριτική απόδοση των προβλεφθέντων Top-20 κωδικών ειδών

Συνοψίζοντας, η ανάλυση ακρίβειας πρόβλεψης αποκαλύπτει ότι το σύστημα πρόβλεψης επιτελεί αξιόπιστα τη λειτουργία εντοπισμού των ειδών υψηλού τζίρου επιτυγχάνοντας `Revenue Capture` **81–97%** σε όλους τους ορίζοντες. Οι αποκλίσεις δεν οφείλονται σε συστηματικά σφάλματα του μοντέλου αλλά, κατά κύριο λόγο, σε δύο δομικούς παράγοντες:

- **Στιγμιαίες μεγάλες παραγγελίες** (βραχυπρόθεσμοι ορίζοντες, H15d) που δεν αποτελούν επαναλαμβανόμενο μοτίβο.
- **Κωδικοί με ελάχιστο ή μηδενικό ιστορικό (**μακροπρόθεσμοι ορίζοντες, H365d) που κανένα προγνωστικό σύστημα δεν μπορεί να εντοπίσει εκ των προτέρων (*cold-start* SKU).

Αυτά τα ευρήματα συνάδουν με τη σχετική βιβλιογραφία, σύμφωνα με την οποία τα μοντέλα μηχανικής μάθησης που χρησιμοποιούνται για πρόβλεψη διαλείπουσας ζήτησης επιτυγχάνουν ισχυρή πρόβλεψης

κατάταξης (*ranking*) (NDCG > 0.85) αλλά παραμένουν ευάλωτα στα νέο-εισερχόμενα προϊόντα [93], [99].

5.6.15.4 Ομάδα Pareto-A

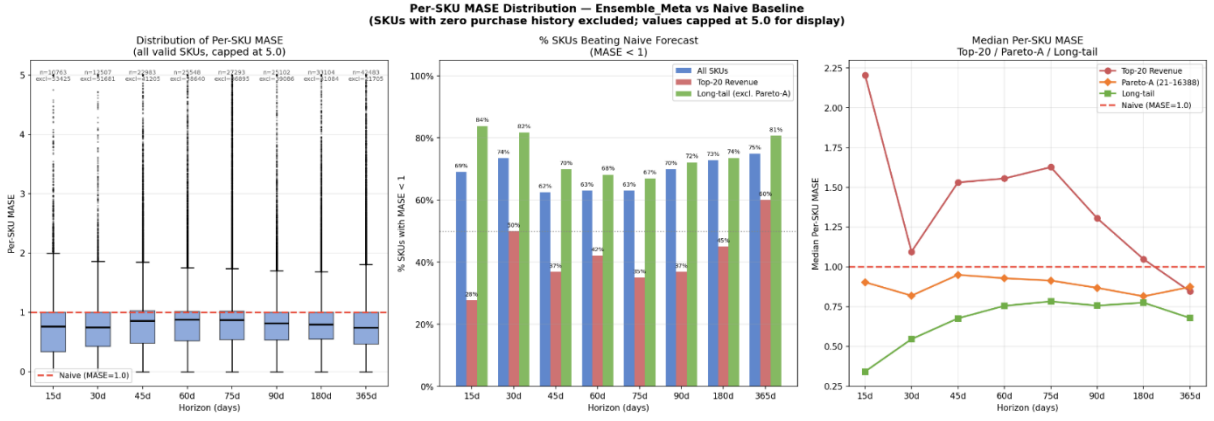
Η ομάδα *Pareto-A*, αποτελούμενη περίπου 16.400 κωδικούς, αποτελεί την κύρια πρόκληση. Με `Ratio` μεταξύ 0,66 και 0,86 παρατηρείται δομική υπό-πρόβλεψη στους κωδικούς μεσαίας δυναμικής, η οποία, κατά κύριο λόγο, οφείλεται στην έντονη διαφοροποίηση του καταλόγου ανά έτος. Στον κάτωθι πίνακα παρατίθενται τα στοιχεία αναλυτικότερα:

Πίνακας 5.24: Απόδοση ειδών ομάδας *Pareto-A*

Ορίζοντας	Πραγματικός τζίρος ομάδας *Pareto-A* (€)	Πραγματική Ζήτηση (Τεμ.)	Πρόβλεψη `Meta` (Τεμ.)	`Ratio`
15d	139.594	11.214	7.448	0,664
30d	316.649	25.129	18.916	0,753
45d	491.591	39.688	30.315	0,764
60d	653.490	53.352	41.831	0,784
75d	822.444	69.233	52.999	0,766
90d	971.394	78.929	62.653	0,794
180d	1.850.439	144.230	123.253	0,854
365d	3.618.431	279.094	239.080	0,857

5.6.15.5 Κατανομή `MASE` ανά κωδικό είδους

Μια κρίσιμη προσθήκη στην ανάλυση είναι η κατανομή του MASE σε επίπεδο SKU. Στο Σχήμα 5.19 παρουσιάζεται η εν λόγω επίδοση ανά κατηγορία ειδών βάσει εσόδων.



Σχήμα 5.19: Κατανομή `MASE` ανά κατηγορία ειδών

Παρατηρείται ότι η ομάδα Pareto-A επιτυγχάνει οριακά καλύτερη απόδοση από την naive προσέγγιση, παρότι παρουσιάζει συστηματική υπό-πρόβλεψη. Επίσης, η καλύτερη σχετική βελτίωση έναντι της αφελούς πρόβλεψης επιτυγχάνεται από την ομάδα της «μακράς ουράς». Όπως αναφέρεται και στο κεφάλαιο 4, η εν λόγω ομάδα καλύπτει περίπου το 20% του συνολικού τζίρου δεσμεύοντας το 50% της συνολικής αξίας αποθήκης. Με τη μέση τιμή ημερών κάλυψης να κυμαίνεται στις 323 ημέρες, κρίνεται

πολύ σημαντική η παροχή βελτιωμένων προβλέψεων και, κατ' επέκταση, μείωση της αξίας αποθήκης. Στο σημείο αυτό, η εξαιρετική επίδοση του συστήματος σε αυτό το φάσμα μεταφράζεται σε δραστική μείωση του δεσμευμένου κεφαλαίου.

5.6.15.6 Διαστήματα Πρόβλεψης (Prediction Intervals)

Τα βασικά μοντέλα παράγουν διαστήματα πρόβλεψης εμπιστοσύνης $[Q_{0.1}, Q_{0.9}]$, τα οποία υπερβαίνουν σταθερά τον ονομαστικό στόχο του 80 (επιτυγχάνοντας κάλυψη $84\% - 89\%$). Στην εικόνα 5.10 παρουσιάζονται οι ακριβείς τιμές κάλυψης (`PI_Coverage`) και εύρους (`PI_Width`). Αυτή η συντηρητική εκτίμηση κινδύνου (*over-coverage*) είναι στρατηγικά επιθυμητή, καθώς μια υπερβολικά στενή πολιτική μπορεί μεν να μειώνει τα αποθέματα, αλλά αυξάνει εκθετικά τον κίνδυνο ελλείψεων (*stock-outs*).

5.6.15.7 Σύνοψη αποτελεσμάτων

Συνοψίζοντας τα ανωτέρω αποτελέσματα, συμπεραίνονται τα κάτωθι:

- Ικανοποιητική επίδοση στις καθολικές μετρικές (`WAPE` 0,54–1,08, `Ratio` 0,95–1,04).
- Καλή επίδοση στα Top-20 προϊόντα που συνεπάγεται την προστασία των ειδών υψηλού τζίρου και διασφαλίζοντας τον πυρήνα των εσόδων της εταιρείας (`Revenue Capture` > *93%*. Στην πλειονότητα των οριζόντων).
- Θεαματική βελτίωση στα είδη της «μακριάς ουράς» (κλάση B/C) σε σχέση με την αντίστοιχη επίδοση της αφελούς προσέγγισης, ώστε να μειωθεί η συνολική αξία των αργά κινούμενων προϊόντων της αποθήκης και, παράλληλα, να εξασφαλιστεί μεγαλύτερη ρευστότητα.
- Η συστηματική υπό-πρόβλεψη των προϊόντων της κλάσης A (*Pareto-A*) παραμένει ένα ανοιχτό ζήτημα αναδεικνύοντας το φαινόμενο των ειδών «ψυχρής εκκίνησης» (*cold-start* SKUs) ως την κύρια πρόκληση για μελλοντική βελτίωση.

Κεφάλαιο 6ο: Επιχειρησιακή εφαρμογή του συστήματος

6.1 Εισαγωγή και επιχειρησιακός στόχος της εφαρμογής

Η μετάβαση από τη θεωρητική διατύπωση μιας μεθοδολογίας πρόβλεψης ζήτησης στην πρακτική της εφαρμογή αποτελεί το κρισιμότερο βήμα για την επιβεβαίωση της χρησιμότητάς της. Η ακαδημαϊκή έρευνα συχνά παράγει μοντέλα με εξαιρετική στατιστική επίδοση, τα οποία, ωστόσο, παραμένουν εγκλωβισμένα σε πειραματικά περιβάλλοντα, υπό μορφή αυτόνομων σεναρίων εντολών (*scripts*) στην Python και μεμονωμένων αρχείων δεδομένων. Η πραγματική αξία ενός τέτοιου συστήματος αναδεικνύεται αποκλειστικά, όταν αυτό μετασχηματιστεί σε ένα λειτουργικό εργαλείο, προσβάσιμο από χρήστες χωρίς εξειδίκευση στην Επιστήμη Δεδομένων.

Στο πλαίσιο αυτό, αναπτύχθηκε η διαδικτυακή εφαρμογή FM_App (*Forecasting and Management Application*), η οποία αποτελεί τη μοναδική διεπαφή με τον τελικό χρήστη. Ο πρωταρχικός σκοπός της εφαρμογής είναι να λειτουργήσει ως ένα ολοκληρωμένο Σύστημα Υποστήριξης Αποφάσεων (*Decision Support System*) γεφυρώνοντας το χάσμα ανάμεσα στην παραγωγή των αλγοριθμικών προβλέψεων και στη λήψη άμεσα εκτελέσιμων αποφάσεων παραγγελίας.

Ενώ ένα μοντέλο Μηχανικής Μάθησης παράγει αμιγώς αριθμητικές εκτιμήσεις μελλοντικής ζήτησης, η εφοδιαστική αλυσίδα απαιτεί κάτι περισσότερο. Αυτό είναι η ενσωμάτωση κρίσιμων εμπορικών παραμέτρων, όπως το τρέχον απόθεμα, ο χρόνος παράδοσης του προμηθευτή (*lead time*) και το επιθυμητό επίπεδο εξυπηρέτησης (*service level*). Η εφαρμογή αναλαμβάνει ακριβώς αυτή τη σύνθεση. Αντί ο υπεύθυνος αναπλήρωσης (*planner*) να εκτελεί το σύστημα (*pipeline*) πρόβλεψης μέσω γραμμής εντολών, η FM_App του παρέχει πλήρη έλεγχο μέσω ενός φιλικού, παραμετρικού περιβάλλοντος. Μέσα από τη διεπαφή, ο χρήστης μπορεί να ρυθμίσει όλες τις παραμέτρους εκπαίδευσης, όπως το χρονικό σημείο αναφοράς της πρόβλεψης (*forecast origin*), την επιλογή μοντέλων όπως Random Forest ή XGBoost και την επιλογή υπερπαραμέτρων, και να εκκινήσει την αυτοματοποιημένη εκτέλεση της ροής εργασιών.

Το τελικό ζητούμενο της εφαρμογής δεν είναι απλώς η παρουσίαση της στατιστικής ακρίβειας, αλλά η μετάφραση της πρόβλεψης σε μια στρατηγική απόφαση, η οποία εξισορροπεί διαρκώς το ρίσκο της έλλειψης (*stock-out*) με το κόστος της υπερβολικής δέσμευσης κεφαλαίου. Με αυτόν τον τρόπο, η εφαρμογή αποδεικνύει την επιχειρησιακή σκοπιμότητα του συστήματος *Sparse True Forecast Pipeline* απέναντι στην πρόκληση της διαλείπουσας ζήτησης.

6.2 Ολοκληρωμένη αρχιτεκτονική πολλαπλών επιπέδων και τεχνολογική στοίβα

Η επιλογή της αρχιτεκτονικής και των τεχνολογικών εργαλείων για τη δημιουργία της εφαρμογής FM_App δεν έγινε με βάση τη γρήγορη ανάπτυξη ενός πειραματικού πρωτοτύπου, αλλά με στόχο την ασφαλή μετάβαση σε περιβάλλον παραγωγής. Για τον λόγο αυτό, η εφαρμογή αναπτύχθηκε ως ένα σύστημα ολοκληρωμένης αρχιτεκτονικής πολλαπλών επιπέδων (*full-stack*), βασισμένο στο πρότυπο του αυστηρού διαχωρισμού ευθυνών (*separation of concerns*). Το σύστημα εξυπηρέτησης (*backend*) και το σύστημα παρουσίασης (*frontend*) αποτελούν δύο διακριτές οντότητες, οι οποίες επικοινωνούν αποκλειστικά μέσω ενός καλά ορισμένου στρώματος διεπαφών REST API και συνδέσεων WebSocket. Στον Πίνακα 6.1 παρατίθεται συνοπτικά η τεχνολογική στοίβα που επιλέχθηκε για κάθε επιμέρους στρώμα της εφαρμογής.

Πίνακας 6.1: Συγκεντρωτικός πίνακας τεχνολογικής στοίβας ανά αρχιτεκτονικό στρώμα της εφαρμογής FM_App

Στρώμα Εφαρμογής	**Τεχνολογία / Εργαλείο**
Σύστημα Εξυπηρέτησης (*Backend*)	FastAPI, SQLAlchemy 2 (async), Alembic, Pydantic 2
Βάση Δεδομένων	Microsoft SQL Server, SQLite (για *tests*)
Ροή Μηχανικής Μάθησης (*ML Pipeline*)	XGBoost, LightGBM, CatBoost, scikit-learn (Hurdle Models)
Βελτιστοποίηση Υπερπαραμέτρων	Optuna 4
Σύστημα Παρουσίασης (*Frontend*)	Vue 3, Vite 6, TypeScript 5.6
Συστατικά Διεπαφής (UI) & Μορφοποίηση	PrimeVue 4, TailwindCSS
Διαχείριση Κατάστασης (*State*)	Pinia, TanStack Vue Query 5
Οπτικοποίηση Δεδομένων & Γραφήματα	Apache ECharts 5, vue-echarts 7
Πολυγλωσσική Υποστήριξη (*i18n*)	vue-i18n 9 (με στρώμα υπερκάλυψης από τη βάση)

6.2.1 Το σύστημα εξυπηρέτησης (*backend*) και η στρατηγική Βάσεων Δεδομένων

Στο επίπεδο του *backend*, ο πυρήνας της εφαρμογής αναπτύχθηκε με το πλαίσιο FastAPI. Η επιλογή του συγκεκριμένου πλαισίου έναντι παλαιότερων λύσεων τεκμηριώνεται από την εγγενή υποστήριξη ασύγχρονης λειτουργίας (*asyncio*). Σε μια εφαρμογή, όπου η εκπαίδευση αλγορίθμων Μηχανικής Μάθησης (όπως τα Random Forests και CatBoost) και η βελτιστοποίηση παραμέτρων (μέσω της βιβλιοθήκης `Optuna`) απαιτούν έντονους και χρονοβόρους υπολογισμούς, η ασύγχρονη λειτουργία εξασφαλίζει ότι η εξυπηρέτηση των υπόλοιπων αιτημάτων του χρήστη δεν παραμένει αποκλεισμένη (*non-blocking* I/O). Παράλληλα, η επικύρωση των δεδομένων εισόδου και εξόδου υλοποιείται μέσω του Pydantic 2, το οποίο εγγυάται την αυστηρή τυποποίηση της πληροφορίας και δημιουργεί αυτόματα τα συμβόλαια διασύνδεσης (OpenAPI).

Η αλληλεπίδραση με τα δεδομένα διαχειρίζεται μέσω του ασύγχρονου SQLAlchemy 2. Κρίσιμη, επίσης, σχεδιαστική επιλογή αποτελεί η στρατηγική διπλής συμβατότητας στη Βάση Δεδομένων. Συγκεκριμένα:

- **Περιβάλλον παραγωγής:** Ως κύριο σύστημα διαχείρισης χρησιμοποιείται ο Microsoft SQL Server (με ασύγχρονο οδηγό `aioodbc`). Η επιλογή αυτή υπαγορεύθηκε από την ευρεία εγκατάσταση του SQL Server σε βιομηχανικά επιχειρησιακά περιβάλλοντα, γεγονός που ελαχιστοποιεί τις τριβές και δυσχέρειες κατά την επιχειρησιακή ενσωμάτωση της εφαρμογής σε υπάρχοντα συστήματα Διαχείρισης Επιχειρησιακών Πόρων (ERP).
- **Περιβάλλον ανάπτυξης και δοκιμών:** Για τις ανάγκες των αυτοματοποιημένων ελέγχων αξιοποιείται η SQLite σε λειτουργία μνήμης (*in-memory*). Αυτό επιτρέπει την εκτέλεση εκατοντάδων δοκιμών (*unit και integration tests*) σε ελάχιστα δευτερόλεπτα επιταχύνοντας δραστικά τον κύκλο ανάπτυξης, δίχως την ανάγκη εξωτερικού διακομιστή.

6.2.2 Μεθοδολογία Κάθετων Τομών (*Vertical Slices*)

Σε αντίθεση με την παραδοσιακή αρχιτεκτονική οριζόντιων επιπέδων, η ανάπτυξη της FM_App βασίστηκε στη σύγχρονη μεθοδολογία των Κάθετων Τομών (*Vertical Slices*). Αυτό σημαίνει ότι κάθε νέα λειτουργικότητα της εφαρμογής (για παράδειγμα, η δημιουργία μιας παραγγελίας ή η ρύθμιση των οριζόντων πρόβλεψης) υλοποιήθηκε ως μια αυτοτελής μονάδα που διαπερνά όλα τα επίπεδα του συστήματος.

Συνολικά, το σύστημα αποτελείται από δεκαοκτώ τέτοιες κάθετες τομές. Κάθε φορά που εισάγεται μια νέα δυνατότητα, η ανάπτυξή της γίνεται ολιστικά. Αυτό σημαίνει ότι υλοποιούνται παράλληλα οι απαιτούμενες αλλαγές σε όλα τα επίπεδα, δηλαδή από την προσαρμογή του σχήματος της ΒΔ (με τη χρήση του εργαλείου Alembic για τα *database migrations*) και την επιχειρησιακή λογική στο *backend*, έως τα τελικά σημεία του API, τα στοιχεία διεπαφής στο *frontend* και τους αντίστοιχους αυτοματοποιημένους ελέγχους λογισμικού (*tests*). Αυτή η αρχιτεκτονική στεγανοποιεί τον κώδικα, μειώνει τις εξαρτήσεις και επιτρέπει την ασφαλή κλιμάκωση του συστήματος. Η σταδιακή εξέλιξη του σχήματος της βάσης δεδομένων αποτυπώνεται ρητά στο ιστορικό των μεταβάσεων (*migrations*). Κάθε μετάβαση αντιστοιχεί συνήθως σε μία νέα κάθετη τομή, προσθέτοντας τους πίνακες και τις σχέσεις που απαιτεί η εκάστοτε νέα λειτουργικότητα. Η τήρηση ενός τέτοιου ιστορικού επιτρέπει την αναπαραγώγιμη και ασφαλή αναβάθμιση του συστήματος. Η ακολουθία αυτής της σταδιακής ανάπτυξης αποτυπώνεται στον Πίνακα 6.2.

Πίνακας 6.2 Ιστορικό μεταβάσεων (0001–0017) μέσω του εργαλείου Alembic, ως ένδειξη της σταδιακής εξέλιξης του σχήματος της βάσης δεδομένων.

Α/Α Μετάβασης	**Περιγραφή Υλοποίησης / Οντότητες**
0001	Αρχικοί πίνακες συστήματος.
0002	Ρυθμίσεις, οπτικά θέματα, και σύστημα μεταφράσεων (*i18n*).
0003 – 0012	Είδη, προμηθευτές, ιεραρχίες, σύνολα δεδομένων, εργασίες παρασκηνίου, ροή Μηχανικής Μάθησης, Optuna και μετρικές.
0013	Εκτελέσεις του Οδηγού Πρόβλεψης (wizard_runs).
0014	Στοιχεία δυναμικού υπολογισμού (Βήμα 2 του Οδηγού).
0015	Κεφαλίδες και επιμέρους γραμμές παραγγελιών.
0016	Εύρη υπερπαραμέτρων μοντέλων.
0017	Διόρθωση περιορισμού ελέγχου ISJSON στη βάση δεδομένων.

6.2.3 Το σύστημα παρουσίασης (*frontend*) και η διαχείριση κατάστασης

Το γραφικό περιβάλλον (*frontend*) αναπτύχθηκε με το πλαίσιο Vue 3 και τη γλώσσα TypeScript. Η χρήση του TypeScript αποτελεί προέκταση της φιλοσοφίας αυστηρής τυποποίησης που εφαρμόζεται και στο *backend* με το Pydantic. Δηλαδή, τα συμβόλαια των δεδομένων αποτυπώνονται ρητά εντοπίζοντας ασυμφωνίες κατά τη μεταγλώττιση και όχι κατά την εκτέλεση.

Μια από τις πιο κρίσιμες τεχνικές προκλήσεις σε μια εφαρμογή διαχείρισης δεδομένων μεγάλης κλίμακας είναι η διαχείριση της κατάστασης (*state management*). Η FM_App υιοθετεί την πρακτική του διαχωρισμού μεταξύ της κατάστασης του διακομιστή και της τοπικής κατάστασης της διεπαφής:

- Η βιβλιοθήκη TanStack Vue Query αναλαμβάνει την επικοινωνία με τον διακομιστή, την αποθήκευση στην προσωρινή μνήμη (*caching*), την αυτόματη επαναφόρτωση και την ακύρωση παρωχημένων δεδομένων στο *frontend*. Αυτό απαλλάσσει το σύστημα από την υπερφόρτωση με περιττά αιτήματα HTTP για δεδομένα πρόβλεψης που έχουν ήδη ανακτηθεί.
- Η βιβλιοθήκη Pinia διαχειρίζεται αποκλειστικά την τοπική κατάσταση της εφαρμογής (π.χ. ποιο βήμα του οδηγού πρόβλεψης παρακολουθεί αυτή τη στιγμή ο χρήστης).

Η οπτικοποίηση των δεδομένων, όπως οι καμπύλες ζήτησης και τα διαγράμματα των μετρικών MASE και WAPE, υλοποιήθηκε μέσω της βιβλιοθήκης `Apache ECharts` προσφέροντας υψηλή απόδοση ακόμα και κατά την απεικόνιση χιλιάδων σημείων δεδομένων.

6.2.4 Ενιαίος χειρισμός σφαλμάτων και χρονοβόρων διεργασιών

Η λειτουργική σταθερότητα του συστήματος βασίζεται στην υιοθέτηση ενός καθολικού μηχανισμού διαχείρισης σφαλμάτων. Στο *backend*, τα σφάλματα οργανώνονται σε ιεραρχία εξαιρέσεων που μεταφράζονται αυτόματα σε τυποποιημένους κωδικούς κατάστασης HTTP, δηλαδή `400` για σφάλματα επικύρωσης, `404` για μη εύρεση πόρου, `409` για επιχειρησιακές συγκρούσεις και `500` για αστοχίες στο μοντέλο Μηχανικής Μάθησης. Ένας κεντρικός διερμηνέας (*interceptor*) στο *frontend* λαμβάνει αυτούς τους τυποποιημένους φακέλους σφαλμάτων και τους μετατρέπει αυτομάτως σε δίγλωσσες, κατανοητές προς τον χρήστη ειδοποιήσεις.

Τέλος, προκειμένου να διασφαλιστεί η απρόσκοπτη λειτουργία και η άμεση απόκριση της διεπαφής κατά τη διάρκεια της εκπαίδευσης των μοντέλων Μηχανικής Μάθησης, η οποία μπορεί να διαρκέσει από αρκετά λεπτά έως ώρες, υλοποιήθηκε σύστημα εργασιών παρασκηνίου (*background jobs*). Η κατάσταση της εκάστοτε εργασίας καταγράφεται στη Βάση Δεδομένων, ενώ η πρόοδος διεργασιών, όπως η τρέχουσα δοκιμή του αλγορίθμου HistGradientBoosting ή το στάδιο του αναδρομικού ελέγχου (*backtesting*), μεταδίδεται δυναμικά στο περιβάλλον του χρήστη μέσω συνδέσεων WebSocket, ώστε να εξασφαλίζεται μια ομαλή και επαγγελματική εμπειρία χρήσης.

6.3 Λειτουργικότητα και διεπαφή χρήστη (Ενσωμάτωση της υπολογιστικής ροής του συστήματος προβλέψεων)

Η βασική αρχή σχεδιασμού της διεπαφής χρήστη (User Interface – UI) ήταν η απόκρυψη της τεχνικής πολυπλοκότητας διατηρώντας, παράλληλα, τον πλήρη έλεγχο επί της ροής εργασιών (*pipeline*) της Μηχανικής Μάθησης. Για να τεθεί σε παραγωγική λειτουργία ένας τόσο σύνθετος μηχανισμός, ήταν απαραίτητο να καταργηθεί η ανάγκη χειροκίνητης παραμετροποίησης αρχείων ρυθμίσεων (όπως τα αρχεία κώδικα ή διαμόρφωσης `cfg`) από τον τελικό χρήστη. Επομένως, αντικαταστάθηκε από ένα διαδραστικό, καθοδηγούμενο περιβάλλον.

Η εφαρμογή παρέχει κεντρικές οθόνες διαχείρισης των βασικών δεδομένων και επιτρέπει την εύκολη επιλογή παραμέτρων, προμηθευτών, χρόνων παράδοσης, ειδών και κατηγοριών. Ωστόσο, η καρδιά της λειτουργικότητας βρίσκεται στον διαδραστικό Οδηγό Πρόβλεψης (*Forecasting Wizard*). Μέσω αυτού του περιβάλλοντος, ο χρήστης μπορεί να εκτελέσει ολόκληρη τη ροή εργασιών (*pipeline*) της πρόβλεψης ζήτησης, από την εκπαίδευση των μοντέλων έως τον υπολογισμό των αποθεμάτων ασφαλείας και σημείων αναπαραγγελίας, ορίζοντας παραμέτρους μέσα από γραφικά στοιχεία ελέγχου.

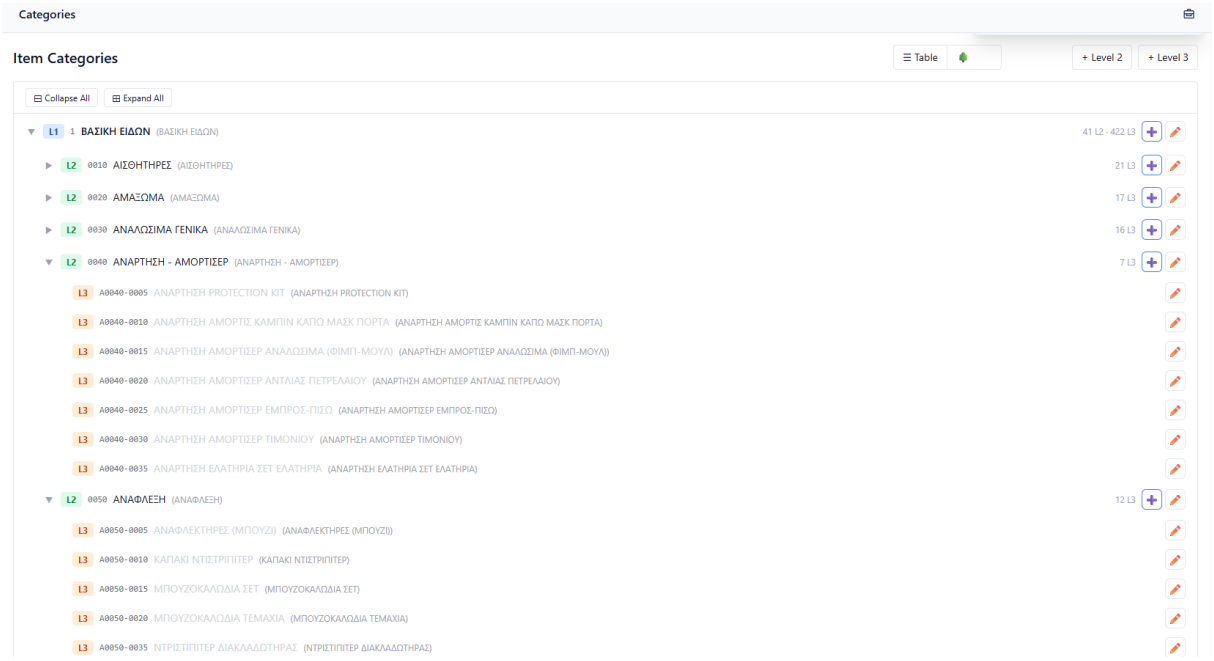
6.3.1 Οδηγός Αρχικής Εγκατάστασης (*Onboarding Wizard*) και Κύρια Δεδομένα (*Master Data*)

Προκειμένου να διασφαλιστεί ότι ο χρήστης δεν θα βρεθεί ποτέ σε ένα μη λειτουργικό περιβάλλον εξαιτίας πιθανής απουσίας δεδομένων, η εφαρμογή υλοποιεί έναν έξυπνο μηχανισμό αρχικοποίησης. Κατά την πρώτη εκκίνηση, και εφόσον δεν ανιχνευθεί εγγεγραμμένο σύνολο δεδομένων (*dataset*) στη ΒΔ, οι φύλακες πλοήγησης (*navigation guards*) του Vue Router ανακατευθύνουν αυτόματα τον χρήστη στον Οδηγό Αρχικής Εγκατάστασης στο τελικό σημείο `/onboarding`.

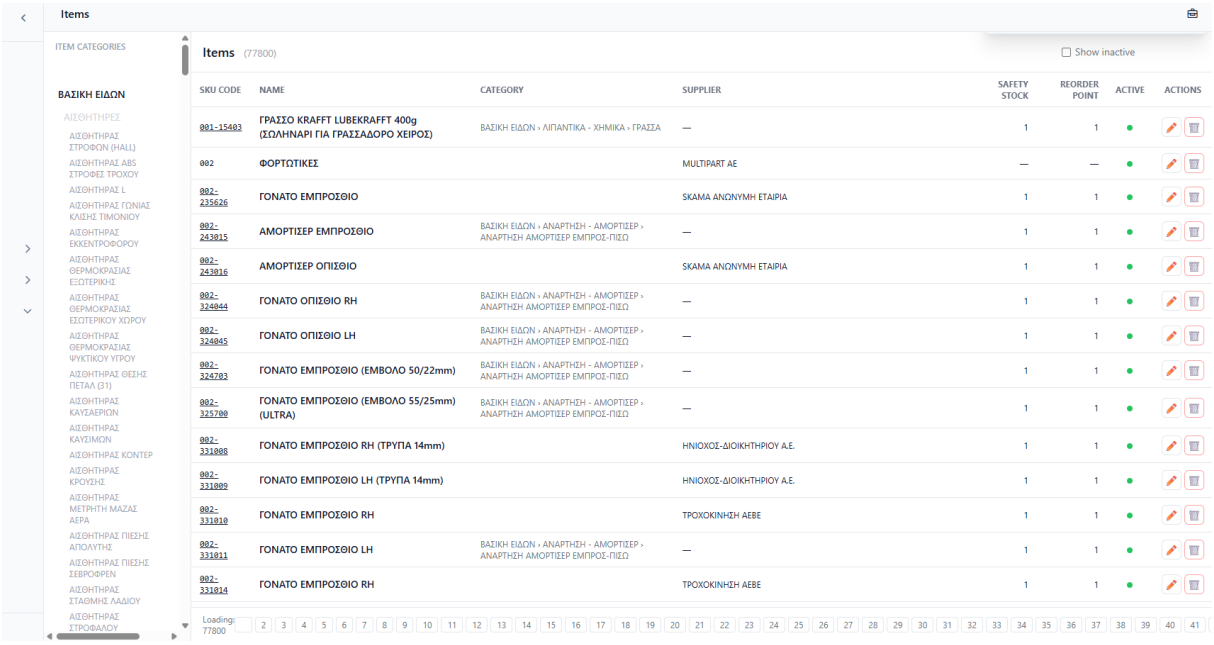
Ο οδηγός αυτός (*onboarding wizard*) καθοδηγεί τον χρήστη διαδοχικά στον έλεγχο της σύνδεσης με τη Βάση Δεδομένων, στην εφαρμογή των απαραίτητων δομικών αλλαγών του σχήματος (`migrations`), στην εισαγωγή των προεπιλεγμένων ρυθμίσεων και τέλος, στη φόρτωση του ιστορικού πωλήσεων και παραγγελιών.

Πέραν της αρχικής εγκατάστασης, το σύστημα προσφέρει πλήρεις οθόνες διαχείρισης των Κύριων Δεδομένων (*Master Data*), δηλαδή των προμηθευτών, των ειδών και της ιεραρχίας κατηγοριών. Ιδιαίτερη έμφαση έχει δοθεί στον μηχανισμό μαζικής εισαγωγής δεδομένων. Μέσω ενός επαναχρησιμοποιήσιμου οδηγού εισαγωγής, το σύστημα ανιχνεύει αυτόματα τη μορφή του διαθέσιμου αρχείου `CSV` (προσδιορίζοντας το διαχωριστικό και την κωδικοποίηση χαρακτήρων βάσει της σήμανσης `BOM`) και προσφέρει στον χρήστη μια προεπισκόπηση για την αντιστοίχιση των στηλών (*column mapping*) πριν την οριστική αποθήκευση στη ΒΔ.

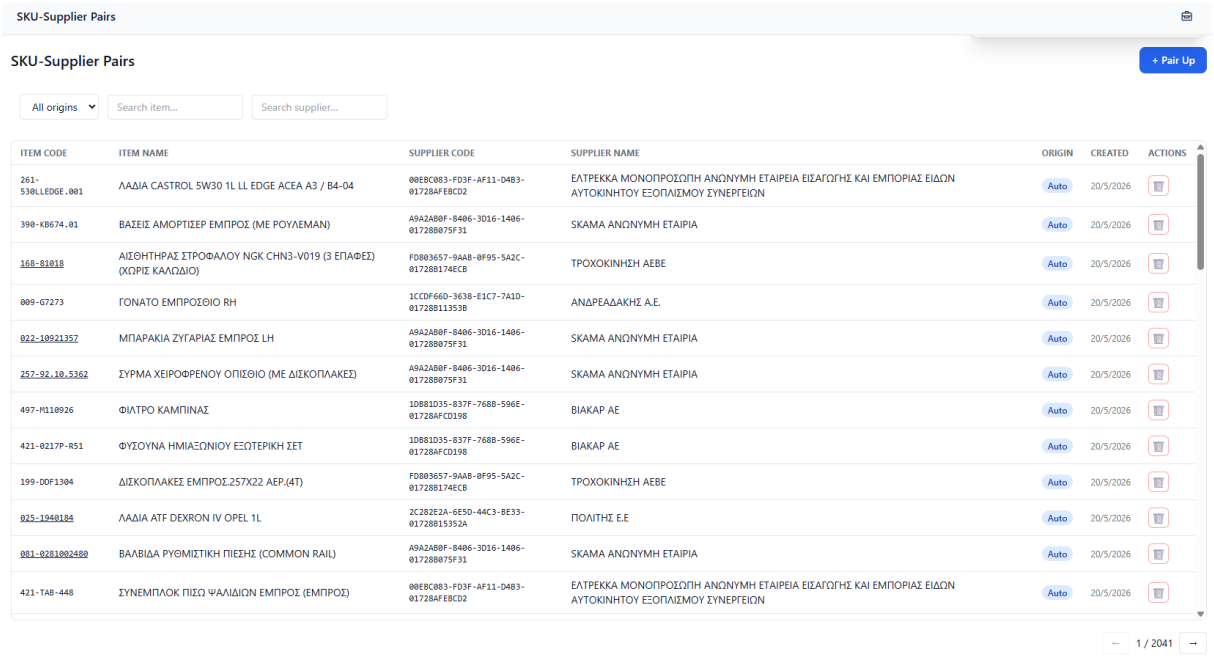
Στα σχήματα 6.1 – 6.3 παρουσιάζονται ενδεικτικές όψεις των κύριων δεδομένων:



Σχήμα 6.1: Κατηγορίες ειδών



Σχήμα 6.2: Καταχωρημένοι κωδικοί ειδών



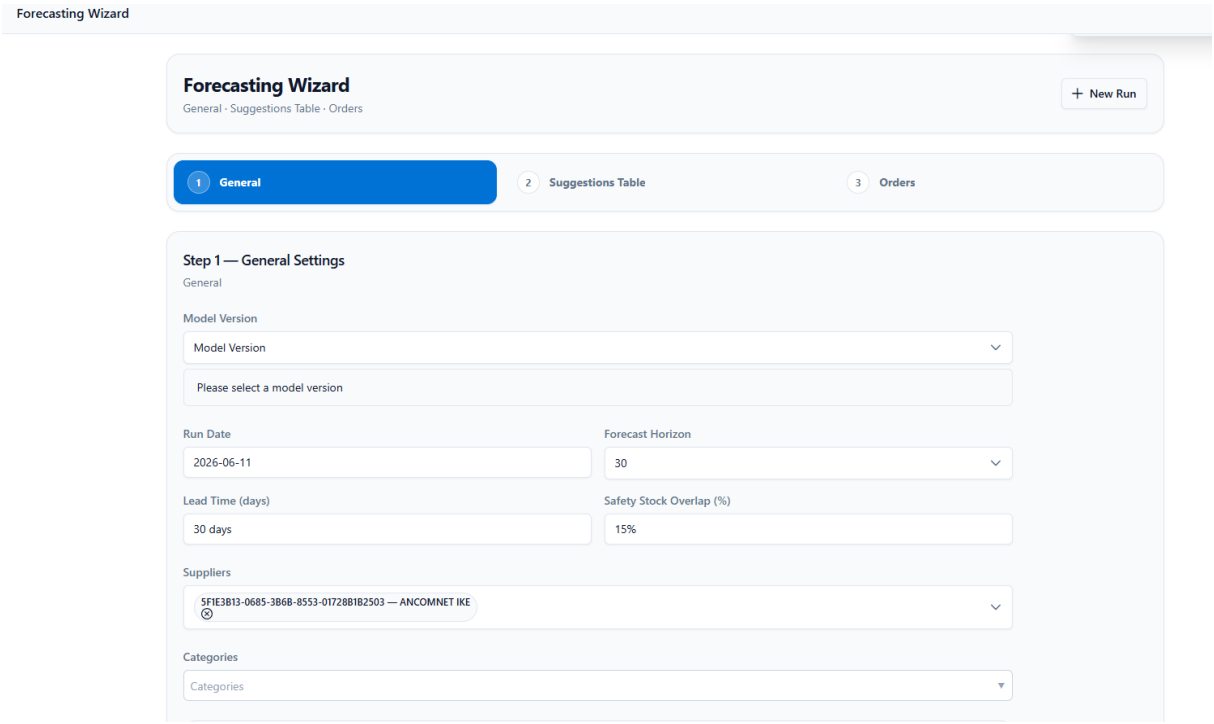
Σχήμα 6.3: Μοναδικά ζεύγη κωδικού προϊόντος και προμηθευτή

6.3.2 Ο Οδηγός Πρόβλεψης (*Forecasting Wizard*) και η παραγωγή προτάσεων

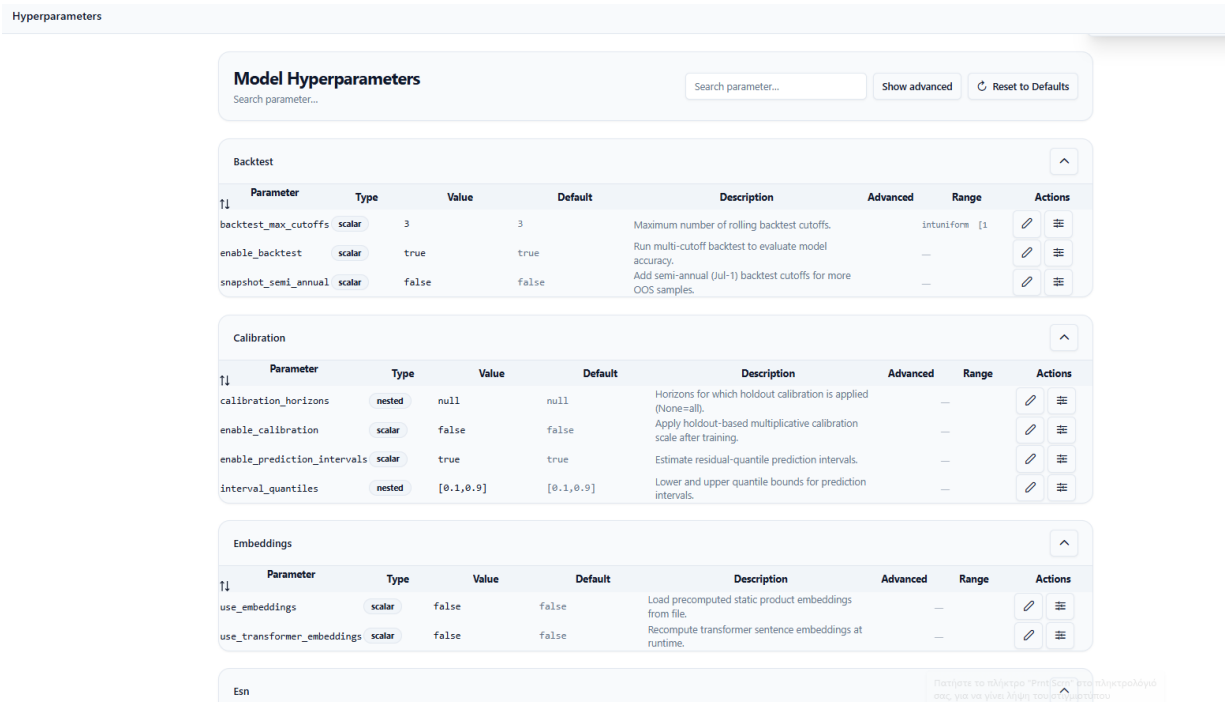
Ο πυρήνας της επιχειρησιακής αξίας της εφαρμογής είναι ο Οδηγός Πρόβλεψης, ο οποίος μεταφράζει τις στατιστικές προβλέψεις σε εκτελέσιμες αποφάσεις αναπλήρωσης μέσα από τρία σαφώς ορισμένα βήματα:

Βήμα 1: Στρατηγική παραμετροποίηση. Στο πρώτο βήμα, ο υπεύθυνος προμηθειών ορίζει τις επιχειρησιακές παραμέτρους. Επιλέγει το σημείο αναφοράς πρόβλεψης (*forecast origin*), τον χρονικό ορίζοντα εξυπηρέτησης, την έκδοση του εκπαιδευμένου μοντέλου που επιθυμεί να αξιοποιήσει (π.χ. ένα σταθμισμένο *ensemble* των μοντέλων Random Forests και XGBoost), τον χρόνο παράδοσης (*lead time*) του προμηθευτή και το επιθυμητό επίπεδο εξυπηρέτησης πελατών (*service level*). Οι επιλογές

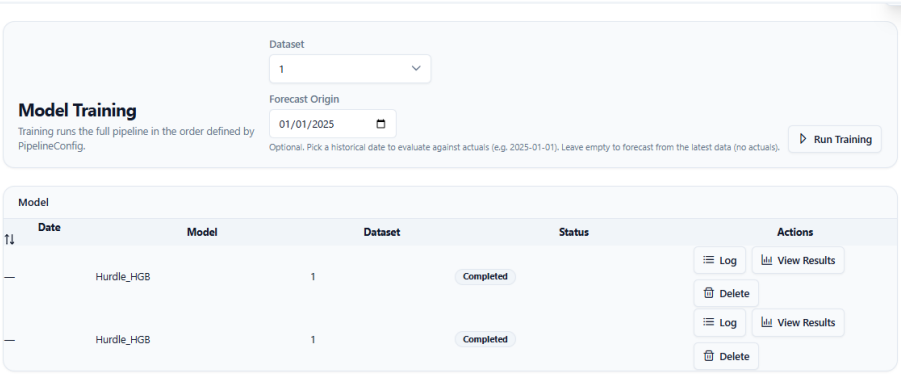
αποθηκεύονται αυτόματα στο παρασκήνιο με τη χρήση τεχνικής χρονικής καθυστέρησης (debounce), βελτιστοποιώντας την εμπειρία χρήσης. Στο σχήμα 6.4 παρουσιάζεται η αρχική σελίδα του οδηγού με τι βασικές ρυθμίσεις και στα σχήματα 6.5 – 6.7 παρατίθενται στιγμιότυπα από την εκπαίδευση του μοντέλου πρόβλεψης μέσω της εφαρμογής.



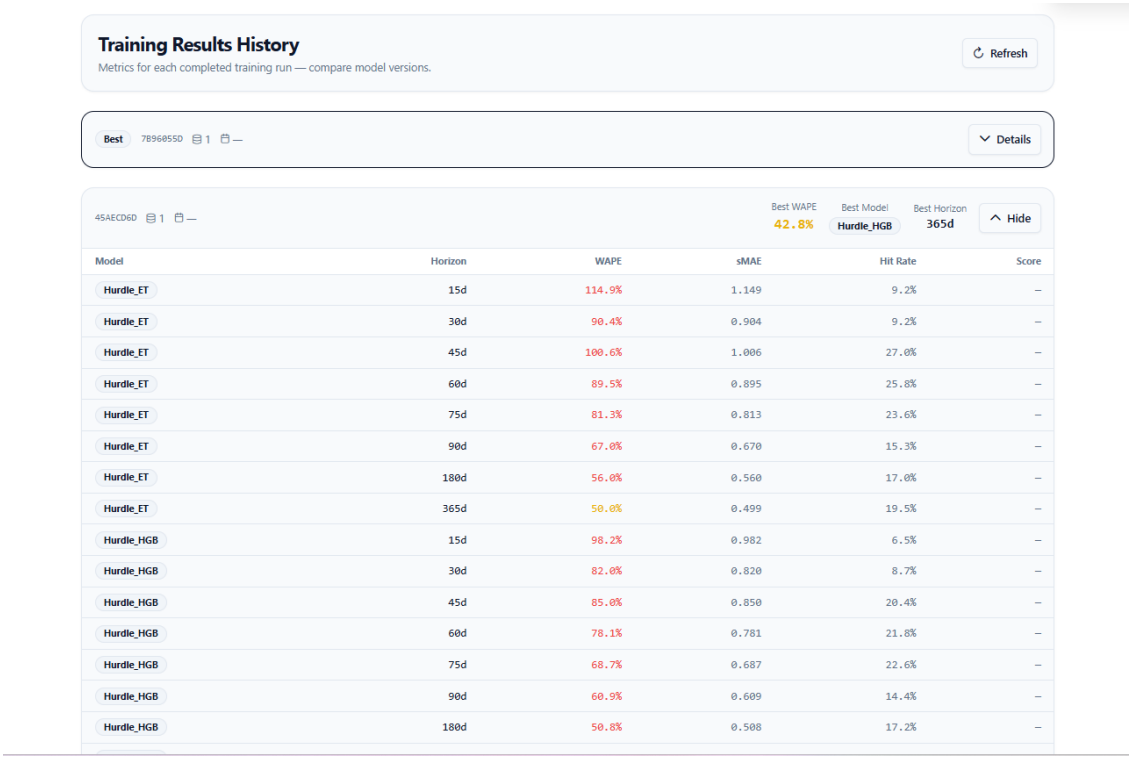
Σχήμα 6.4: Αρχική σελίδα οδηγού εφαρμογής FM_App



Σχήμα 6.5: Ρύθμιση υπερπαραμέτρων εκπαίδευσης



Σχήμα 6.6: Εκπαίδευση μοντέλου μέσω της εφαρμογής



Σχήμα 6.7: Αποτελέσματα εκπαίδευσης

Βήμα 2: Πίνακας προτάσεων και δυναμικοί υπολογισμοί. Στο δεύτερο βήμα ενεργοποιείται ο πυρήνας των υπολογισμών (`Recompute`). Η υπηρεσία του *backend* αντλεί τις προβλέψεις ζήτησης και εκτελεί τους απαραίτητους υπολογισμούς για κάθε κωδικό προϊόντος (SKU):

- Το υπολειμματικό σφάλμα της πρόβλεψης (*residual* RMSE) χρησιμοποιείται για τον υπολογισμό του απαραίτητου αποθέματος ασφαλείας (*safety stock*).
- Το σύστημα προσδιορίζει το τρέχον απόθεμα αθροίζοντας τις ποσότητες ανά αποθήκη.
- Υπολογίζεται η προβλεπόμενη κατάσταση αποθέματος (*projected stock*) αφαιρώντας τη σωρευτική προβλεπόμενη ζήτηση από το τρέχον απόθεμα.
- Η τελική προτεινόμενη ποσότητα παραγγελίας προκύπτει ως η μη-αρνητική διαφορά μεταξύ του αναπροσαρμοσμένου αποθέματος ασφαλείας και της προβλεπόμενης κατάστασης αποθέματος.

Για τη διευκόλυνση του χρήστη, ο πίνακας προτάσεων χρησιμοποιεί οπτική κωδικοποίηση: το κόκκινο χρώμα επισημαίνει αρνητικό προβλεπόμενο απόθεμα (άμεσος κίνδυνος *stock-out*), ενώ το κίτρινο χρώμα υποδεικνύει τις εγγραφές με θετική πρόταση παραγγελίας.

Βήμα 3: Οριστικοποίηση και ενσωμάτωση της ανθρώπινης κρίσης. Στο τελευταίο στάδιο παρέχεται μια σαφής προεπισκόπηση των προτεινόμενων ποσοτήτων, ομαδοποιημένων ανά προμηθευτή. Εδώ, εφαρμόζεται η αρχή της ενσωμάτωσης της ανθρώπινης κρίσης (*human-in-the-loop*), καθώς ο χρήστης μπορεί να τροποποιήσει τις ποσότητες πριν την τελική επιβεβαίωση λαμβάνοντας υπόψη αστάθμητους παράγοντες που ενδέχεται να αγνοεί ο αλγόριθμος

6.3.3 Διαχείριση παραγγελιών και Κύκλος Ζωής (*State Machine*)

Με την ολοκλήρωση του Οδηγού Πρόβλεψης, το σύστημα δημιουργεί αυτόματα τις κεφαλίδες και τις γραμμές των παραγγελιών. Η διεπαφή διαχείρισης αυτών των παραγγελιών οργανώνεται αρχιτεκτονικά σε τρεις στήλες: τη λίστα των εκτελέσεων (*runs*), τις κεφαλίδες παραγγελιών και τις λεπτομέρειες των επιμέρους γραμμών. Αυτή η διάταξη υλοποιεί ένα πρότυπο σταδιακής εμβάθυνσης (*drill-down*) και επιτρέπει την ομαλή πλοήγηση από τη συνολική εικόνα στη λεπτομέρεια.

Για τη διασφάλιση της επιχειρησιακής ακεραιότητας, κάθε παραγγελία διέπεται από έναν αυστηρό κύκλο ζωής που υλοποιείται μέσω μιας μηχανής καταστάσεων (*state machine*). Η παραγγελία δημιουργείται αρχικά ως προσχέδιο (`DRAFT`). Από εκεί μπορεί να επιβεβαιωθεί (`CONFIRMED`), να σημανθεί ως απεσταλμένη προς τον προμηθευτή (`SENT`) και τελικά να κλείσει (`CLOSED`). Το σύστημα απορρίπτει οποιαδήποτε μη έγκυρη μετάβαση κατάστασης αποτρέποντας, για παράδειγμα, την αλλαγή ποσοτήτων σε μια ήδη απεσταλμένη παραγγελία.

Τέλος, η εξαγωγή των επιβεβαιωμένων παραγγελιών υποστηρίζεται τόσο σε μορφή `CSV` όσο και σε Excel με αξιοποίηση της βιβλιοθήκης `openpyxl` της Python, προσφέροντας έτσι ένα τυποποιημένο αρχείο που είναι άμεσα αξιοποιήσιμο από τα εταιρικά συστήματα ERP.

6.4 Μοντελοποίηση και ροή Μηχανικής Μάθησης (*Machine Learning pipeline*)

Το υποσύστημα Μηχανικής Μάθησης αποτελεί τον υπολογιστικό πυρήνα της εφαρμογής FM_App και ενσωματώνει το σύνολο της μεθοδολογίας που αναπτύχθηκε στα προηγούμενα κεφάλαια. Η εκπαίδευση των μοντέλων και η παραγωγή των προβλέψεων υλοποιείται μέσω μιας εκτενούς ροής εργασιών (*pipeline*), της οποίας η ορχήστρωση γίνεται διαφανώς στο παρασκήνιο (μέσω του *backend*). Οι υποενότητες που ακολουθούν, έχουν αναλυθεί διεξοδικά στα προηγούμενα κεφάλαια. Στο σημείο αυτό γίνεται μια πολύ σύντομη αναφορά για λόγους πληρότητας, καθότι εκτελούνται και μέσω της εφαρμογής.

6.4.1 Αρχιτεκτονική μοντέλων Εμποδίου δύο σταδίων (*Two-stage* Hurdle *models*)

Η θεμελιώδης προσέγγιση μοντελοποίησης που υιοθετεί η εφαρμογή βασίζεται στα μοντέλα Εμποδίου (Hurdle *models*), τα οποία είναι ειδικά σχεδιασμένα για να αντιμετωπίζουν τη φύση της ακραία αραιής ζήτησης (*sparse demand*). Όπως αναλύθηκε και στα προηγούμενα κεφάλαια, η κλασική ενιαία παλινδρόμηση τείνει να αποτυγχάνει σε τέτοια σύνολα δεδομένων, καθώς η συντριπτική πλειοψηφία των μηδενικών παρατηρήσεων παρασύρει την πρόβλεψη προς το μηδέν.

Το σύστημα επιλύει αυτό το πρόβλημα διαχωρίζοντας την εκτίμηση σε δύο ανεξάρτητα στάδια:

- **Στάδιο Ταξινόμησης (*Classification*):** Ένας ταξινομητής εκτιμά την πιθανότητα (p) εμφάνισης έστω και μίας πώλησης (μη μηδενικής ζήτησης) στον επιλεγμένο χρονικό ορίζοντα.

- **Στάδιο Παλινδρόμησης (*Regression*):** Ένας παλινδρομητής εκτιμά το μέγεθος της ζήτησης (μ), υπό την αυστηρή προϋπόθεση ότι η ζήτηση θα είναι θετική.

Η τελική αναμενόμενη ζήτηση εξάγεται μαθηματικά ως το γινόμενο των δύο εκτιμήσεων ($\mathbb{E}[\text{Demand}] = p \cdot \mu$). Για την εκτέλεση αυτών των δύο σταδίων, η εφαρμογή ενσωματώνει ως βασικούς εκτιμητές (*base estimators*) προηγμένους αλγορίθμους δέντρων αποφάσεων και ενίσχυσης με κλίση (*gradient boosting*) υλοποιημένους μέσω των αντίστοιχων κλάσεων της Python (όπως `RandomForestClassifier`, `ExtraTreesRegressor`, `HistGradientBoostingRegressor`).

6.4.2 Διαχείριση λογοκριμένης ζήτησης (*censored demand*) και στάθμιση

Μία από τις σημαντικότερες καινοτομίες της ενσωματωμένης ροής είναι ο δυναμικός μηχανισμός διαχείρισης της λογοκριμένης ζήτησης (*demand censoring*). Σε ένα πραγματικό περιβάλλον εφοδιαστικής αλυσίδας, πολλές ημέρες με μηδενικές πωλήσεις δεν οφείλονται σε απουσία αγοραστικού ενδιαφέροντος, αλλά σε έλλειψη διαθεσιμότητας του προϊόντος στο ράφι (*out-of-stock*).

Η εφαρμογή ανακατασκευάζει εσωτερικά το ιστορικό του επιπέδου αποθέματος από τις επιμέρους συναλλαγές. Όταν εντοπίζει περιόδους, κατά τις οποίες η μηδενική ζήτηση οφειλόταν αποδεδειγμένα σε έλλειψη αποθέματος, μειώνει το βάρος (*sample weight*) αυτών των εγγραφών κατά τη διάρκεια της εκπαίδευσης, αντί να τις απορρίψει πλήρως. Με αυτόν τον τρόπο αποτρέπεται η συστηματική υποεκτίμηση της ζήτησης. Επιπλέον, για την αντιμετώπιση της ανισορροπίας των κλάσεων, εφαρμόζεται προσαρμοστική στάθμιση της θετικής κλάσης ανά ορίζοντα.

6.4.3 Βελτιστοποίηση υπερπαραμέτρων και μνημόνευση μοντέλων (*model caching*)

Η αναζήτηση του βέλτιστου συνδυασμού παραμέτρων διαχειρίζεται πλήρως μέσα από τη διεπαφή της εφαρμογής, αξιοποιώντας στο *backend* τη βιβλιοθήκη Optuna. Ο χρήστης μπορεί να ορίσει εύρη τιμών για διάφορες παραμέτρους και το σύστημα εκτελεί διαδοχικές μελέτες (*studies*) και δοκιμές (*trials*). Τα βέλτιστα αποτελέσματα καταγράφονται στη Βάση Δεδομένων και μπορούν να προαχθούν σε μόνιμες προεπιλεγμένες ρυθμίσεις (`promote-to-default`) δημιουργώντας έναν διαρκή βρόχο βελτίωσης.

Καθοριστικής σημασίας για την αποδοτικότητα της επαναλαμβανόμενης εκπαίδευσης είναι ο ενσωματωμένος μηχανισμός μνημόνευσης (*model cache*). Κάθε εκπαιδευμένο μοντέλο αποθηκεύεται σε ένα κοινόχρηστο *cache* με ένα μοναδικό κλειδί κατακερματισμού (*hash*), το οποίο συντίθεται από τον χρονικό ορίζοντα, τον αλγόριθμο, τα χαρακτηριστικά και μια ψηφιακή υπογραφή των δεδομένων εκπαίδευσης. Όταν ζητηθεί ένας ίδιος συνδυασμός, το σύστημα ανακτά άμεσα το αποθηκευμένο μοντέλο μειώνοντας δραστικά τον χρόνο υπολογισμού.

6.4.4 Αναδρομικός έλεγχος (*backtesting*) και συσσωρευτική γενίκευση

Για να διασφαλιστεί η αξιοπιστία πριν την παραγωγή των παραγγελιών, η υπολογιστική ροή εκτελεί αυτοματοποιημένο αναδρομικό έλεγχο (*backtesting*). Το σύστημα επαναλαμβάνει τη διαδικασία πρόβλεψης σε παλαιότερα χρονικά σημεία αποκοπής (*cutoffs*) και συγκρίνει τις προβλέψεις με τις πραγματικές πωλήσεις. Η αξιολόγηση παράγει κρίσιμες μετρικές, όπως το σταθμισμένο απόλυτο σφάλμα (WAPE), το κλιμακωμένο σφάλμα (MASE) και τον ρυθμό επιτυχούς εντοπισμού (`HitRate_Binary`).

Προκειμένου να ελαχιστοποιηθεί η διασπορά και να αξιοποιηθεί η συμπληρωματικότητα των διαφορετικών αλγορίθμων, οι προβλέψεις συντίθενται σε συνολικά μοντέλα (*ensembles*) με

σημαντικότερο το συσσωρευτικό σύνολο του μοντέλου μετα-μάθησης (*meta-learner stacking ensemble ή* `ensemble_meta`).

6.5 Παραγωγή προτάσεων αναπλήρωσης και επιχειρησιακή αξία

Το σημείο όπου η αλγοριθμική ακρίβεια μετασχηματίζεται σε άμεση και μετρήσιμη επιχειρησιακή αξία είναι η παραγωγή προτάσεων αναπλήρωσης. Σε αυτή τη φάση, η εφαρμογή FM_App υπερβαίνει τον ρόλο ενός απλού στατιστικού εργαλείου και λειτουργεί ως ένα έξυπνο Σύστημα Υποστήριξης Αποφάσεων, το οποίο αντλεί τις προβλέψεις από τη Βάση Δεδομένων και τις συνδυάζει με τα εμπορικά κριτήρια της επιχείρησης. Η παραγωγή παραγγελιών δομείται σε τρία σαφή επιχειρησιακά στάδια.

6.5.1 Στρατηγική παραμετροποίηση (Επιλογή κριτηρίων) και δυναμικοί υπολογισμοί

Ο υπεύθυνος αναπλήρωσης επιλέγει τους προμηθευτές ή τις κατηγορίες προϊόντων που επιθυμεί να αναλύσει. Για κάθε σενάριο, καθορίζει τον εκτιμώμενο χρόνο παράδοσης (*lead time*) του εκάστοτε προμηθευτή και το επιθυμητό επίπεδο εξυπηρέτησης πελατών (*service level*). Στο παρασκήνιο, το σύστημα έχει ήδη αναλύσει το ιστορικό χρησιμοποιώντας τεχνικές αντιμετώπισης της λογοκρισίας της ζήτησης (*demand censoring*) διασφαλίζοντας ότι οι ιστορικές μηδενικές πωλήσεις λόγω έλλειψης αποθέματος δεν θα ξεγελάσουν τον αλγόριθμο σε προτάσεις υπο-αναπλήρωσης.

Με βάση τις παραμέτρους του χρήστη, η εφαρμογή υπολογίζει δυναμικά και προβάλλει τα εξής κρίσιμα μεγέθη ανά κωδικό:

- Τη συνολική προβλεπόμενη ζήτηση για τον επιλεγμένο ορίζοντα.
- Το απαιτούμενο **απόθεμα ασφαλείας (SS)**, το οποίο προσαρμόζεται ανάλογα με την αβεβαιότητα της πρόβλεψης.
- Το **σημείο αναπαραγγελίας (ROP)**, δηλαδή το ακριβές κατώφλι αποθέματος που ενεργοποιεί την ανάγκη για παραγγελία, προκειμένου να καλυφθεί η ζήτηση κατά τη χρονική διάρκεια που μεσολαβεί, έως ότου ικανοποιηθεί η παραγγελία.
- Την τελική προτεινόμενη ποσότητα παραγγελίας.

6.5.2 Πρόβλεψη με γνώμονα το απόθεμα (*Inventory-aware forecasting*)

Σε αυτό το στάδιο ενσωματώνεται πλήρως η λογική της πρόβλεψης με γνώμονα το απόθεμα (*inventory-aware forecasting*). Το σύστημα διαφοροποιεί τη στρατηγική του βάσει της εμπορικής βαρύτητας των ειδών:

- Για τους **κωδικούς υψηλού τζίρου (Top-20)**, το σύστημα διατηρεί αυστηρά όρια αποθέματος, ώστε να αποτρέπεται δραστικά ο κίνδυνος των ελλείψεων. Στη διεπαφή, το χρώμα κόκκινο προειδοποιεί τον χρήστη για επερχόμενη έλλειψη, αν δεν ληφθεί άμεσα δράση. Εξάλλου, με, μόλις, 55 ημέρες μέση αναμονή στο ράφι (βλ. ενότητα 4.5) δεν υφίσταται ιδιαίτερος κίνδυνος για υπερβολικό απόθεμα. Οπότε, οι αναπληρώσεις ακολουθούν σταθερά αισιόδοξο σενάριο.
- Αντιθέτως, για τα **είδη αραιής ζήτησης (long-tail SKUs)**, το σύστημα αξιοποιεί την ικανότητα των μοντέλων Hurdle να αναγνωρίζουν την αληθινή μηδενική ζήτηση προτείνοντας γενναίες μειώσεις στο απόθεμα ασφαλείας και απελευθερώνοντας πολύτιμο κεφάλαιο (working capital) για την επιχείρηση. Σε αυτή την κατηγορία, η μέση διάρκεια κάλυψης είναι 323 ημέρες (βλ. ενότητα 4.5) και, άρα, το βασικό μέλημα είναι η μείωση της αξίας αποθήκης.

6.5.3 Ενσωμάτωση της ανθρώπινης κρίσης (*human-in-the-loop*)

Αναγνωρίζοντας ότι κανένα αλγοριθμικό μοντέλο δεν μπορεί να συλλάβει πλήρως εξωγενείς, αστάθμητους εμπορικούς παράγοντες, όπως ξαφνικές αλλαγές στην πολιτική ενός προμηθευτή ή έκτακτες προσφορές αγοράς, το σύστημα σχεδιάστηκε με βάση την αρχή της συμμετοχής του ανθρώπινου παράγοντα (*human-in-the-loop*).

Στο τελευταίο στάδιο της ροής, η εφαρμογή παρουσιάζει μια καθαρή προεπισκόπηση των προτάσεων. Ο χρήστης δεν εκτελεί τυφλά τις εντολές του συστήματος, αλλά διατηρεί την ευχέρεια να επεμβαίνει χειροκίνητα τροποποιώντας τις ποσότητες πριν την τελική έγκριση της παραγγελίας, όταν αυτό κρίνεται αναγκαίο. Μέσα από αυτή τη διαδικασία, η εφαρμογή ενορχηστρώνει μια πλήρη και μετρήσιμη επιχειρησιακή ροή συνδυάζοντας την αντικειμενικότητα της Μηχανικής Μάθησης με την επιχειρηματική οξυδέρκεια του υπευθύνου προμηθειών.

6.6 Διαφάνεια αποφάσεων (*Explainability*) και εμπιστοσύνη στο σύστημα

Η επιτυχής **μετάβαση** ενός συστήματος Τεχνητής Νοημοσύνης **σε περιβάλλον παραγωγής** και η υιοθέτησή του από το προσωπικό μιας επιχείρησης (*user adoption*) δεν εξαρτάται μόνο από τη στατιστική του ακρίβεια, αλλά και από την ερμηνευσιμότητα των αποτελεσμάτων του. Ένα από τα συχνότερα προβλήματα των προηγμένων μοντέλων Μηχανικής Μάθησης είναι ότι λειτουργούν ως «μαύρα κουτιά» (*black boxes*) δίνοντας έτοιμες απαντήσεις χωρίς να εξηγούν τον συλλογισμό τους, γεγονός που δημιουργεί διστακτικότητα στους τελικούς χρήστες.

Για να αντιμετωπιστεί αυτή η πρόκληση, η FM_App σχεδιάστηκε με γνώμονα τη διαφάνεια αποφάσεων και την ερμηνευσιμότητα (*explainability*) των αποτελεσμάτων του. Αυτό επιτυγχάνεται προσφέροντας την απαραίτητη πληροφορία σε δύο διακριτά επίπεδα, τα οποία εδραιώνουν την εμπιστοσύνη (*Trust in* AI) του χρήστη προς το σύστημα:

Αναλυτικό επίπεδο (Οπτικοποίηση επίδοσης).

Η εφαρμογή παρέχει ειδικές οθόνες (*dashboards*), στις οποίες μπορεί ο χρήστης να δει την πραγματική επίδοση των μοντέλων (όπως το Random Forests ή το CatBoost) που συμμετείχαν στην πρόβλεψη. Μέσω διαδραστικών γραφημάτων, παρουσιάζονται οι ίδιες μετρικές αξιολόγησης που χρησιμοποιήθηκαν και κατά την ανάπτυξη του συστήματος (όπως το WAPE, το MASE και ο Ρυθμός Επιτυχίας – Hit Rate). Ο σκοπός εδώ δεν είναι η εις βάθος στατιστική ανάλυση από πλευράς του υπευθύνου αναπλήρωσης, αλλά η παροχή ορατότητας και διαφάνειας. Με άλλα λόγια, ο χρήστης μπορεί να εντοπίσει με μια ματιά σε ποιες κατηγορίες προϊόντων ή σε ποιους χρονικούς ορίζοντες το σύστημα αποδίδει με τη μεγαλύτερη ακρίβεια και πού υπάρχει μεγαλύτερη αβεβαιότητα προσαρμόζοντας αντίστοιχα το επιθυμητό επίπεδο εξυπηρέτησης.

Διαγνωστικό επίπεδο (Αρχεία καταγραφής – *System logs*).

Σε ένα βαθύτερο τεχνικό επίπεδο, το σύστημα διαθέτει ενσωματωμένη οθόνη καταγραφής συμβάντων σε πραγματικό χρόνο (*system logs*). Ο χρήστης (ή ο διαχειριστής του συστήματος) έχει πλήρη εικόνα των διεργασιών που εκτελούνται στο παρασκήνιο, από την εξαγωγή των χαρακτηριστικών και την εκπαίδευση του *Meta-learner*, μέχρι την εφαρμογή των τεχνικών διόρθωσης.

Παρέχοντας αυτή τη διπλή διαφάνεια, το σύστημα παύει να είναι ένας απρόσωπος αυτοματισμός και λειτουργεί ως ένας αξιόπιστος σύμβουλος, του οποίου οι προτάσεις είναι τεκμηριωμένες, ελέγξιμες και βασισμένες σε ιστορικά δεδομένα που ο ίδιος ο χρήστης έχει στη διάθεσή του.

6.7 Σύνοψη κεφαλαίου

Στο τρέχον κεφάλαιο παρουσιάστηκε η επιχειρησιακή εφαρμογή FM_App, η οποία αποτελεί το πρακτικό απόσταγμα της μεθοδολογίας που αναπτύχθηκε σε αυτή τη διπλωματική εργασία. Στόχος ήταν να καταδειχθεί πώς η ακαδημαϊκή έρευνα πάνω στη διαλείπουσα ζήτηση μπορεί να μετασχηματιστεί σε ένα πλήρως λειτουργικό λογισμικό (με ολοκληρωμένη αρχιτεκτονική πολλαπλών επιπέδων), το οποίο είναι έτοιμο να τεθεί σε παραγωγική λειτουργία.

Αποκρύπτοντας την πολυπλοκότητα των αλγορίθμων Μηχανικής Μάθησης και πίσω από ένα φιλικό, παραμετρικό περιβάλλον χρήστη, η εφαρμογή επιτρέπει σε στελέχη χωρίς τεχνικό υπόβαθρο να εκτελούν προηγμένες ροές δεδομένων. Η ενσωμάτωση της πρόβλεψης με γνώμονα το απόθεμα (*inventory-aware forecasting*) αποδείχθηκε καθοριστική, διότι το σύστημα δεν παράγει απλώς στατιστικά νούμερα, αλλά μεταφράζει τις προβλέψεις σε άμεσα εκτελέσιμες προτάσεις αναπλήρωσης υπολογίζοντας δυναμικά το απόθεμα ασφαλείας και το σημείο αναπαραγγελίας.

Εν κατακλείδι, η εφαρμογή μετατοπίζει τον ρόλο του υπευθύνου προμηθειών από τον χειροκίνητο υπολογισμό παραγγελιών στη λήψη στρατηγικών αποφάσεων. Εξισορροπώντας διαρκώς τον κίνδυνο των ελλείψεων με το κόστος του πλεονάζοντος αποθέματος, η FM_App επιβεβαιώνει ότι οι σύγχρονες τεχνικές (όπως τα μοντέλα Εμποδίου) μπορούν να παραγάγουν μετρήσιμη επιχειρησιακή αξία στο άκρως απαιτητικό τμήμα της δευτερογενούς αγοράς ανταλλακτικών αυτοκινήτων.

Κεφάλαιο 7ο: Αποτίμηση και μελλοντικές επεκτάσεις

7.1 Εισαγωγή

Το παρόν καταληκτικό κεφάλαιο επιτελεί έναν διττό ρόλο. Αφενός, προβαίνει σε μια συνολική, κριτική αποτίμηση της ερευνητικής διαδρομής και των αποτελεσμάτων που παρουσιάστηκαν αναλυτικά στα προηγούμενα κεφάλαια επιχειρώντας να απαντήσει με τεκμηριωμένο τρόπο στα ερευνητικά ερωτήματα που τέθηκαν κατά την εισαγωγή της εργασίας. Αφετέρου, αναγνωρίζοντας ότι κανένα ψηφιακό σύστημα δεν είναι απαλλαγμένο από περιορισμούς, καταγράφει με διαφάνεια τα δομικά όρια της προτεινόμενης λύσης και χαρτογραφεί συγκεκριμένες, ρεαλιστικές κατευθύνσεις για τη μελλοντική της εξέλιξη. Η αφήγηση που ακολουθεί δεν αποτελεί μια απλή επανάληψη των τεχνικών βημάτων, αλλά επιχειρεί τη σύνθεσή τους σε ένα ενιαίο πλαίσιο.

7.2 Απαντήσεις στα ερευνητικά ερωτήματα

Η παρούσα διπλωματική εργασία σχεδιάστηκε με γνώμονα την αντιμετώπιση ενός ρεαλιστικού και εξαιρετικά απαιτητικού επιχειρησιακού προβλήματος. Μέσα από την επαναληπτική ανάπτυξη της μεθοδολογίας και την τελική υλοποίηση της εφαρμογής FM_App, η έρευνα παρέχει πλέον τεκμηριωμένες απαντήσεις στα τέσσερα κεντρικά ερευνητικά ερωτήματα που τέθηκαν κατά την εισαγωγή:

Ερώτημα 1: Ποιες αρχιτεκτονικές Μηχανικής Μάθησης είναι καταλληλότερες για *άμεση* (*direct*) πρόβλεψη σε πολλαπλούς χρονικούς ορίζοντες υπό ενός περιβάλλοντος μεγάλης κλίμακας με ακραία διαλείπουσα και λογοκριμένη ζήτηση;

Η έρευνα απέδειξε ότι οι παραδοσιακές αρχιτεκτονικές ενιαίας παλινδρόμησης αποτυγχάνουν σε τέτοια περιβάλλοντα, καθώς η πληθώρα μηδενικών πωλήσεων παρασύρει την πρόβλεψη. Η καταλληλότερη αρχιτεκτονική αποδείχθηκε η χρήση μοντέλων Εμποδίου δύο σταδίων (*two-stage* Hurdle *models*) ενσωματωμένων σε μια ροή πραγματικής πρόβλεψης (*Sparse True Forecast Pipeline*). Διαχωρίζοντας την πιθανότητα εμφάνισης ζήτησης (ταξινόμηση) από το μέγεθος αυτής (παλινδρόμηση), αλγόριθμοι δέντρων αποφάσεων και ενίσχυσης με κλίση (όπως οι RandomForests, ExtraTrees και HistGradientBoosting) κατόρθωσαν να απομονώσουν τον θόρυβο. Παράλληλα, η λογοκριμένη ζήτηση διαχειρίστηκε επιτυχώς με τη δυναμική μείωση του βάρους (*sample weights*) των παρατηρήσεων σε ημέρες με έλλειψη αποθέματος (*stock-outs*), αποτρέποντας τη μάθηση εσφαλμένων μηδενικών προτύπων.

Ερώτημα 2: Πώς μπορεί να εκτιμηθεί αξιόπιστα η αβεβαιότητα των προβλέψεων μέσω διαστημάτων εμπιστοσύνης, ώστε να υποστηριχθεί ο μη-παραμετρικός υπολογισμός των σημείων αναπαραγγελίας (ROP);

Η μετάβαση από τη σημειακή πρόβλεψη σε αποφάσεις αναπλήρωσης απαιτεί την ακριβή ποσοτικοποίηση της αβεβαιότητας. Η προσέγγιση που υιοθετήθηκε στηρίχθηκε στον άμεσο, μη-παραμετρικό υπολογισμό του Σημείου Αναπαραγγελίας (ROP) βάσει του εμπειρικού διαστήματος πρόβλεψης (*prediction interval*). Αξιοποιώντας το ανώτατο όριο της πρόβλεψης (*upper bound*) του μοντέλου μετα-μάθησης (*meta-learner*), το οποίο αντιστοιχεί στο επιθυμητό Επίπεδο Εξυπηρέτησης (π.χ. 90%) και κλιμακώνοντάς το γραμμικά στον χρόνο παράδοσης (*lead time*), το σύστημα εξάγει το ROP χωρίς να προϋποθέτει κανονικότητα στην κατανομή των σφαλμάτων, η οποία καταρίπτεται σε

περιπτώσεις διαλείπουσας ζήτησης. Η αφαίρεση της αναμενόμενης μέσης ζήτησης από αυτό το σημείο αποδίδει δυναμικά το απαιτούμενο Απόθεμα Ασφαλείας (*safety stock*).

Ερώτημα 3: Πώς μπορεί η βαθμονόμηση (*calibration*) των προβλέψεων να διατηρηθεί συνεπής τόσο στα κρίσιμα επιχειρησιακά προϊόντα (υψηλός τζίρος) όσο και στη «μακρά ουρά» (*long-tail*) της χαμηλής κίνησης αποφεύγοντας φαινόμενα συστηματικής υπο-εκτίμησης;

Το φαινόμενο της συστηματικής υπο-εκτίμησης αντιμετωπίστηκε μέσω εξειδικευμένων στρατηγικών στάθμισης και συσσωρευτικής γενίκευσης (*stacked generalization*). Για τα κρίσιμα προϊόντα (Top-20 και *Pareto-A*), η εφαρμογή στρωματοποιημένων βαρών (*stratified scaling*) και ο μηχανισμός λογοκρισίας ζήτησης (*demand censoring*) εξασφάλισαν ότι οι αλγόριθμοι δεν επηρεάζονται αρνητικά από ιστορικές ελλείψεις. Παράλληλα, το μοντέλο μετα-μάθησης (*meta-learner*) βαθμονόμησε τις προβλέψεις συνθέτοντας τα επιμέρους μοντέλα διατηρώντας την ικανότητα αναγνώρισης της αληθινής μηδενικής ζήτησης στη «μακρά ουρά», δίχως να θυσιάζεται η επιθετική κάλυψη που απαιτείται για τους κωδικούς υψηλού τζίρου.

Ερώτημα 4: Σε ποιον βαθμό η ενσωμάτωση αυτών των προγνωστικών αλγορίθμων σε μια πλήρη κανονιστική πολιτική αποθεμάτων οδηγεί σε μετρήσιμη επιχειρησιακή βελτίωση εξισορροπώντας το επίπεδο εξυπηρέτησης με τη δεσμευμένη ρευστότητα;

Η απάντηση αναδείχθηκε στην πράξη μέσω της διεπαφής FM_App. Η ενσωμάτωση της πρόβλεψης με γνώμονα το απόθεμα (*inventory-aware forecasting*) οδήγησε σε μετρήσιμη επιχειρησιακή βελτίωση διαχωρίζοντας στρατηγικά τον κατάλογο. Στα είδη αραιής ζήτησης (με ιστορική μέση κάλυψη 323 ημερών) το σύστημα προτείνει δραστικές μειώσεις του αποθέματος ασφαλείας απελευθερώνοντας κρίσιμη ρευστότητα (*working capital*). Αντιθέτως, στους κωδικούς που ανήκουν στις ομάδες Top-20 και *Pareto-A*, εγγυάται σταθερά υψηλό επίπεδο εξυπηρέτησης. Μετατρέποντας τα στατιστικά μεγέθη σε έτοιμες προτάσεις παραγγελίας, η εργασία αποδεικνύει ότι τα προηγμένα μοντέλα συνιστούν ισχυρούς μηχανισμούς επιχειρησιακής εξισορρόπησης.

7.3 Επιχειρησιακή συνεισφορά

Πέρα από την ακαδημαϊκή τεκμηρίωση, η παρούσα ΔΕ προσφέρει άμεση και μετρήσιμη αξία στο πραγματικό επιχειρησιακό περιβάλλον της εφοδιαστικής αλυσίδας και, ειδικότερα, στον κλάδο της δευτερογενούς αγοράς αυτοκινήτου (*automotive aftermarket*). Η επιχειρησιακή συνεισφορά του συστήματος συνοψίζεται στους ακόλουθους άξονες:

- **Απελευθέρωση δεσμευμένου κεφαλαίου (*Working capital*):** Όπως καταδείχθηκε από την ανάλυση του συνόλου δεδομένων, η συντριπτική πλειονότητα των κωδικών (περίπου 77%) ανήκει στη «μακρά ουρά» (*long-tail*) και εμφανίζει εξαιρετικά αραιή ζήτηση. Ωστόσο, δεσμεύει δυσανάλογα μεγάλο μέρος του κεφαλαίου. Το προτεινόμενο σύστημα αναγνωρίζει με ακρίβεια την αληθινή μηδενική ζήτηση μέσω των μοντέλων Εμποδίου (Hurdle *models*) και προτείνει δραστικές μειώσεις στα περιττά αποθέματα ασφαλείας και απελευθερώνοντας ρευστότητα.
- **Ενίσχυση του επιπέδου εξυπηρέτησης (Αποφυγή ελλείψεων):** Για τους κρίσιμους κωδικούς υψηλού τζίρου, Top-20, οι οποίοι εμφανίζουν μέση αναμονή στο ράφι μόλις 55 ημερών, το σύστημα αποτρέπει την υποεκτίμηση της ζήτησης. Αξιοποιώντας τον μηχανισμό προστασίας από τις ιστορικές ελλείψεις (*demand censoring*), εγγυάται ότι η επιχείρηση δεν θα χάσει πωλήσεις λόγω λανθασμένης πρόβλεψης.
- **Μετάβαση σε προληπτική στρατηγική (*Proactive management*):** Μέσω της διαδικτυακής εφαρμογής FM_App, η λειτουργία του τμήματος προμηθειών μετασχηματίζεται. Ο υπεύθυνος αναπλήρωσης (*planner*) απεγκλωβίζεται από τους χρονοβόρους, χειροκίνητους υπολογισμούς

και τις υποκειμενικές εκτιμήσεις. Το σύστημα αυτοματοποιεί την παραγωγή προτάσεων επιτρέποντας στον χρήστη να εστιάσει στη στρατηγική λήψη αποφάσεων, διατηρώντας τον τελικό έλεγχο μέσω της αρχής της ανθρώπινης παρέμβασης (*human-in-the-loop*).

7.4 Περιορισμοί του συστήματος

Κάθε προηγμένο ψηφιακό σύστημα, όσο άρτια και αν έχει σχεδιαστεί, λειτουργεί εντός συγκεκριμένων ορίων τα οποία καθορίζονται από τη φύση των δεδομένων και τις αρχιτεκτονικές επιλογές. Η διαφανής καταγραφή αυτών των περιορισμών αποτελεί δείγμα επιστημονικής ωριμότητας και ορίζει το πλαίσιο για μελλοντική έρευνα:

- **Το πρόβλημα της ψυχρής εκκίνησης (*cold-start*).** Ο σημαντικότερος εγγενής περιορισμός είναι η αδυναμία πρόβλεψης για κωδικούς με ανύπαρκτο ή ελάχιστο ιστορικό ζήτησης στο σύνολο εκπαίδευσης. Κατά μεθοδολογική επιλογή, οι κωδικοί αυτοί απορρίπτονται, καθώς ένα σύστημα που βασίζεται αποκλειστικά στο ιστορικό σήμα δεν μπορεί να προβλέψει ζήτηση για προϊόντα που αγνοεί. Η συνέπεια αυτού του περιορισμού αποτυπώνεται άμεσα στα αποτελέσματα. Η συστηματική υπό-πρόβλεψη της ομάδας *Pareto-A* και η πτώση της επικάλυψης (*overlap*) στην κατάταξη των Top-20 προϊόντων στον ετήσιο ορίζοντα οφείλονται, σε μεγάλο βαθμό, σε κωδικούς που εμφάνισαν εκρηκτική, μη προβλέψιμη αύξηση ζήτησης το 2025 χωρίς προηγούμενο ιστορικό. Πρόκειται για ένα φυσικό όριο κάθε αμιγώς ιστορικού προγνωστικού συστήματος.
- **Απουσία εξωγενών επιχειρησιακών σημάτων.** Η τρέχουσα προσέγγιση είναι αμιγώς ερευνητική. Κατά συνέπεια, το μοντέλο δεν λαμβάνει καμία εξωτερική πληροφορία σχετικά με επιχειρηματικές αποφάσεις, όπως διαφημιστικές καμπάνιες, εμπορικές συμφωνίες ή ακυρώσεις προμηθευτών. Αρκετές από τις παρατηρούμενες ακραίες αποκλίσεις (*spikes*) ανάγονται ακριβώς σε τέτοια γεγονότα, τα οποία σε ένα πλήρως παραγωγικό περιβάλλον θα μπορούσαν να κωδικοποιηθούν ως επιπλέον επεξηγηματικά χαρακτηριστικά.
- **Απλοποιητικές παραδοχές στη σύζευξη με τη διαχείριση αποθεμάτων.** Ο υπολογισμός της ζήτησης κατά τον χρόνο προμήθειας στηρίζεται σε γραμμική κλιμάκωση μεταξύ των διαθέσιμων οριζόντων. Παρότι η παραδοχή αυτή κρίνεται ασφαλής δεδομένης της πυκνότητας του συνόλου των οριζόντων, εισάγει μια μικρή προσέγγιση για περιπτώσεις, κατά τις οποίες ο χρόνος παράδοσης δεν συμπίπτει ακριβώς με κάποιον διαθέσιμο ορίζοντα πρόβλεψης. Επιπλέον, η σημειακή φύση της τελικής πρόβλεψης, αν και εμπλουτισμένη με εμπειρικά διαστήματα, υπολείπεται μιας πλήρως πιθανοτικής μοντελοποίησης της κατανομής της ζήτησης.
- **Στατικότητα και περιοδικότητα της επανεκπαίδευσης.** Το σύστημα λειτουργεί σε καθεστώς περιοδικής, εκτός σύνδεσης (*offline*) επανεκπαίδευσης. Το σύνολο δεδομένων παρέχεται μέσω εξωτερικού αρχείου `csv` και δεν έχει υλοποιηθεί μηχανισμός αυτόματης τροφοδότησης νέων δεδομένων από διασυνδεδεμένο σύστημα ERP.

7.5 Μελλοντικές επεκτάσεις

Οι ακόλουθες προτάσεις χαρτογραφούν μια ρεαλιστική πορεία εξέλιξης, η οποία απαντά απευθείας στους περιορισμούς της Ενότητας 7.5.

7.5.1 Αντιμετώπιση του προβλήματος της ψυχρής εκκίνησης: Ενσωμάτωση του GNN

Η σημαντικότερη μελλοντική επέκταση του συστήματος προκύπτει ως άμεση απάντηση στον κυριότερο περιορισμό του, το πρόβλημα της ψυχρής εκκίνησης. Η προτεινόμενη λύση αφορά την εκμετάλλευση

των συσχετίσεων ανάμεσα στα προϊόντα. Μάλιστα, προτείνεται η ενσωμάτωση ενός Νευρωνικού Δικτύου Γράφων (GNN), το οποίο, εν δυνάμει, είναι ικανότερο από τις απλές διανυσματικές αναπαραστάσεις (embeddings) που χρησιμοποιήθηκαν στο *Sparse True Forecast Pipeline*. Η ιδέα αυτή έχει ήδη θεμελιωθεί και αναλυθεί στην περιγραφή της μεθοδολογίας της Προσέγγισης 1 (βλ. ενότητα 5.4). Μάλιστα, στην ενότητα 5.4.7 αναφέρεται ότι, παρότι εγκαταλείφθηκε η Προσέγγιση 1 στο σύνολό της, προτείνεται η μελλοντική χρήση του GNN ως επέκταση στο τελικό σύστημα, εξαιτίας του πλεονεκτήματός του, όσον αφορά την εκμάθηση τοπολογικών σχέσεων μεταξύ των προϊόντων χωρίς εξάρτηση από το ιστορικό ζήτησης.

Η πρόταση συνίσταται στην ενσωμάτωση ενός **Νευρωνικού Δικτύου Γράφων Πολλαπλών Έργων (MTL-GNN)**, το οποίο θα λειτουργεί ως μια αυτόνομης, εξειδικευμένης υπο-μονάδας, αποκλειστικά αφιερωμένης στους νέους κωδικούς. Σε επιχειρησιακό επίπεδο, η επέκταση αυτή θα υλοποιούνταν ως ένα νέο μενού εισαγωγής νέων προϊόντων εντός της εφαρμογής FM_App. Κατά την εισαγωγή ενός νέου κωδικού, ο οποίος, εξ ορισμού, στερείται ιστορικού και άρα απορρίπτεται από την κύρια ροή πρόβλεψης, ο χρήστης θα παρέχει τα διαθέσιμα περιγραφικά και δομικά χαρακτηριστικά του (κατηγορία, μητρικός κωδικός, τεχνικά γνωρίσματα, σχέσεις συμβατότητας). Με βάση αυτά, το GNN θα εντάσσει τον νέο κόμβο εντός του ήδη εκπαιδευμένου γράφου προϊόντων και, μέσω του μηχανισμού διέλευσης μηνυμάτων (message passing), θα παρήγαγε μια αρχική εκτίμηση ζήτησης αξιοποιώντας τη «συλλογική γνώση» των τοπολογικά γειτονικών, ώριμων κωδικών.

Το κρίσιμο σημείο της πρότασης είναι ότι, στο συγκεκριμένο και περιορισμένο αυτό υπο-πρόβλημα, το GNN αναμένεται να υπερτερεί σαφώς των απλούστερων διανυσματικών αναπαραστάσεων (embeddings) που χρησιμοποιεί το τελικό σύστημα. Ο λόγος είναι ότι τα στατικά embeddings κωδικοποιούν μεν τη σημασιολογική συγγένεια, αλλά αδυνατούν να μοντελοποιήσουν ρητά τις σχεσιακές, δομικές εξαρτήσεις του δικτύου προϊόντων, οι οποίες είναι ακριβώς η πληροφορία που απομένει διαθέσιμη, όταν το ιστορικό ζήτησης απουσιάζει. Έτσι, η αδυναμία που οδήγησε στην εγκατάλειψη του GNN στον βραχύ ημερήσιο ορίζοντα μετατρέπεται, στο πλαίσιο της ψυχρής εκκίνησης, στο μοναδικό του συγκριτικό πλεονέκτημα. Με τον τρόπο αυτό, η συνολική αρχιτεκτονική εξελίσσεται σε ένα υβριδικό σύστημα δύο ταχυτήτων, ένα ελαφρύ, αποδοτικό σύνολο δενδρικών μοντέλων για τη συντριπτική πλειονότητα των κωδικών με ιστορικό, και μια εξειδικευμένη μηχανή GNN ως «πύλη εισόδου» των νέων προϊόντων.

7.5.2 Μεθοδολογικές επεκτάσεις

Πέρα από την αντιμετώπιση της ψυχρής εκκίνησης, η μεθοδολογία πρόβλεψης επιδέχεται περαιτέρω εμβάθυνσης σε δύο κατευθύνσεις.

- **Μετάβαση από την έμμεση στην άμεση εκτίμηση των ποσοστημορίων.** Η τρέχουσα υλοποίηση των αποθεμάτων ασφαλείας (βλ. ενότητα 5.6.14) στηρίζεται σε εμπειρικά διαστήματα πρόβλεψης (βλ. ενότητα 5.6.12) που παράγονται έμμεσα από τα κατάλοιπα (*residuals*) του αναδρομικού ελέγχου (*backtesting*). Μια φυσική εξέλιξη αποτελεί η μετάβαση στην πλήρως **πιθανοτική πρόβλεψη (*probabilistic forecasting*)**, κατάτην οποία το μοντέλο θα εκπαιδεύεται απευθείας σε συνάρτηση απώλειας ποσοστημορίου (Pinball Loss), ώστε το επιθυμητό επίπεδο εξυπηρέτησης να αντιστοιχεί σε ένα ποσοστημόριο της προβλεπόμενης κατανομής χωρίς ενδιάμεσες παραδοχές.
- **Ενσωμάτωση εξωγενών επιχειρησιακών σημάτων.** Η ενσωμάτωση των προαναφερθέντων εξωγενών σημάτων, όπως καμπάνιες, εμπορικές συμφωνίες, ημερολόγιο προωθήσεων, ως επιπλέον χαρακτηριστικών θα μετρίαζε σημαντικά τις ακραίες αποκλίσεις που σήμερα

παραμένουν ανεξήγητες από το ιστορικό σήμα και οδηγούν σε σημαντικές αποκλίσεις κατά την πρόβλεψη ζήτησης.

7.5.3 Επεκτάσεις της εφαρμογής και της υποδομής

Σε επίπεδο λειτουργικότητας, η αρθρωτή αρχιτεκτονική της FM_App αφήνει σημαντικά περιθώρια εξέλιξης. Η ενσωμάτωση μηχανισμού αυθεντικοποίησης και ελέγχου πρόσβασης βάσει ρόλων (*role-based access control*) θα επιτρέπει την ταυτόχρονη, ασφαλή χρήση από πολλαπλούς χρήστες και οργανωτικές μονάδες με απομονωμένα μεταξύ τους δεδομένα. Η άμεση διασύνδεση με συστήματα ERP και WMS μέσω προγραμματιστικών διεπαφών (APIs), αντί της ανταλλαγής αρχείων, θα αυτοματοποιήσει πλήρως τη ροή των δεδομένων. Επίσης, θα είναι δυνατό να λαμβάνει χώρα αυτοματοποιημένη εκτέλεση ολόκληρου του συστήματος πρόβλεψης βάσει δοθέντων κριτηρίων, όπως περιοδικότητα ή μεταβολή στη σύνθεση του συνόλου δεδομένων.

7.6 Επίλογος

Η παρούσα διπλωματική εργασία ξεκίνησε από ένα σύνθετο και ρεαλιστικό επιχειρησιακό πρόβλημα. Αυτό είναι η έξυπνη διαχείριση των παραγγελιών αναπλήρωσης αποθεμάτων στη δευτερογενή αγορά αυτοκινήτου, υπό συνθήκες διαλείπουσας και λογοκριμένης ζήτησης. Μέσα από μια διαφανή, επαναληπτική ερευνητική διαδρομή, τεκμηριώθηκε ότι οι σύγχρονες τεχνικές Μηχανικής Μάθησης μπορούν, όχι μόνο να επιτύχουν αξιόπιστες προβλέψεις σε αυτό το δύσκολο πεδίο, αλλά και να ενσωματωθούν με επιτυχία σε ένα λειτουργικό, φιλικό προς τον χρήστη εργαλείο που παράγει μετρήσιμη επιχειρησιακή αξία.

Πέρα από την επιβεβαίωση της αρχικής ερευνητικής υπόθεσης, η εργασία αναδεικνύει και ένα ευρύτερο συμπέρασμα. Η γέφυρα ανάμεσα στην ακαδημαϊκή έρευνα και την εφαρμόσιμη τεχνολογία δεν κατασκευάζεται μόνο με την επιδίωξη της μέγιστης στατιστικής ακρίβειας, αλλά εξίσου με τη βαθιά κατανόηση του επιχειρησιακού πλαισίου, την ειλικρινή αναγνώριση των περιορισμών και τη συστηματική, σταδιακή ωρίμανση του συστήματος. Οι μελλοντικές επεκτάσεις που διατυπώθηκαν δεν αποτελούν αναγκαία βήματα για τη λειτουργία του συστήματος στην τρέχουσα μορφή του, αλλά χαρτογραφούν την πορεία προς την αναβάθμισή του του από ένα ολοκληρωμένο πρωτότυπο σε ένα πλήρως παραγωγικό σύστημα σε επιχειρησιακή κλίμακα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] A. A. Syntetos and J. E. Boylan, "The accuracy of intermittent demand estimates," *International Journal of Forecasting*, vol. 21, no. 2, pp. 303–314, 2005.

[2] R. H. Teunter, A. A. Syntetos, and M. Z. Babai, "Intermittent demand: Linking forecasting to inventory obsolescence," *European Journal of Operational Research*, vol. 214, no. 3, pp. 606–615, 2011.

[3] F. Petropoulos et al., "Forecasting: theory and practice," *International Journal of Forecasting*, vol. 38, no. 3, pp. 705–871, 2022.

[4] J. D. Croston, "Forecasting and stock control for intermittent demands," *Operational Research Quarterly*, vol. 23, no. 3, pp. 289–303, 1972.

[5] A. A. Syntetos, J. E. Boylan, and J. D. Croston, "On the categorization of demand patterns," *Journal of the Operational Research Society*, vol. 56, no. 5, pp. 495–503, 2005.

[6] J. R. Trapero, E. H. de Frutos, and D. J. Pedregal, "Demand forecasting under lost sales stock policies," *International Journal of Forecasting*, vol. 40, no. 3, pp. 1055–1068, 2024.

[7] A. Jain, N. Rudi, and T. Wang, "Demand estimation and ordering under censoring: Stock-out timing is (almost) all you need," *Operations Research*, vol. 63, no. 1, pp. 134–150, 2015.

[8] I. Svetunkov and A. Sroginis, "Why do zeroes happen? A model-based approach for demand classification," arXiv preprint arXiv:2504.05894v2, 2025.

[9] M. Z. Babai, M. M. Ali, and K. Nikolopoulos, "Impact of temporal aggregation on stock control performance of intermittent demand estimators," *Omega*, vol. 40, no. 6, pp. 713–721, 2012.

[10] M. Z. Babai, M. M. Ali, J. E. Boylan, and A. A. Syntetos, "Forecasting and inventory performance in a two-stage supply chain with ARIMA(0,1,1) demand: Theory and empirical analysis," *International Journal of Production Economics*, vol. 143, no. 2, pp. 463–471, 2013.

[11] E. A. Silver, D. F. Pyke, and D. J. Thomas, *Inventory and Production Management in Supply Chains*, 4th ed. Boca Raton, FL, USA: CRC Press, 2016.

[12] Y. Liu, Q. Zhang, Z. P. Fan, T. H. You, and L. X. Wang, "Maintenance spare parts demand forecasting for automobile 4S shop considering weather data," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 5, pp. 943–955, 2019.

[13] C. F. Chien, C. C. Ku, and Y. Y. Lu, "A stacking ensemble framework for intermittent demand forecasting in automotive spare parts," *Computers & Industrial Engineering*, vol. 179, p. 109203, 2023.

[14] S. Kolassa, "Evaluating predictive count data distributions in retail sales forecasting," *International Journal of Forecasting*, vol. 32, no. 3, pp. 788–803, 2016.

[15] S. Kolassa, "Why the 'best' point forecast depends on the error or accuracy measure," *International Journal of Forecasting*, vol. 36, no. 1, pp. 208–211, 2020.

[16] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021. [Online]. Available: https://otexts.com/fpp3/. Accessed: May 20, 2026.

[17] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley, 2015.

[18] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root," *Journal of Econometrics*, vol. 54, nos. 1–3, pp. 159–178, 1992.

[19] T. Gneiting and M. Katzfuss, "Probabilistic forecasting," *Annual Review of Statistics and Its Application*, vol. 1, pp. 125–151, 2014.

[20] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679-688, 2006.

[21] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "M5 accuracy competition: Results, findings, and conclusions," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1346–1364, 2022.

[22] S. Nahmias and T. L. Olsen, *Production and Operations Analysis*, 7th ed. Prospect Heights, IL, USA: Waveland Press, 2015. [Online]. Available: https://books.google.gr/books/about/Production_and_Operations_Analysis.html?id=SIsoBgAAQBAJ&redir_esc=y . Accessed: May 21, 2026.

[23] W. J. Kennedy, J. W. Patterson, and L. D. Fredendall, "An overview of recent literature on spare parts inventories," *International Journal of Production Economics*, vol. 76, no. 2, pp. 201–215, 2002.

[24] A. Bacchetti and N. Saccani, "Spare parts classification and demand forecasting for stock control: Investigating the gap between research and practice," *Omega*, vol. 40, no. 6, pp. 722–737, 2012.

[25] Q. Hu, J. E. Boylan, H. Chen, and A. W. Labib, "OR in spare parts management: A review," *European Journal of Operational Research*, vol. 266, no. 2, pp. 395–414, 2018.

[26] T. M. Williams, "Stock control with sporadic and slow-moving demand," *Journal of the Operational Research Society*, vol. 35, no. 10, pp. 939–948, 1984.

[27] J. E. Boylan and A. A. Syntetos, "Spare parts management: A review of forecasting research and extensions," *IMA Journal of Management Mathematics*, vol. 21, no. 3, pp. 227–237, 2010.

[28] H. Scarf, "The optimality of (s,S) policies in the dynamic inventory problem," in *Mathematical Methods in the Social Sciences*, pp. 196–202. Stanford, CA, USA: Stanford University Press, 1960.

[29] K. J. Arrow, T. Harris, and J. Marschak, "Optimal inventory policy," *Econometrica*, vol. 19, no. 3, pp. 250–272, 1951.

[30] M. Z. Babai, A. A. Syntetos, and R. Teunter, "On the empirical performance of (T, s, S) heuristics," *European Journal of Operational Research*, vol. 195, no. 2, pp. 442–453, 2009.

[31] S. A. Conrad, "Sales data and the estimation of demand," *Journal of the Operational Research Society*, pp. 123–127, 1976.

[32] S. Nahmias, "Demand estimation in lost sales inventory systems," *Naval Research Logistics*, vol. 41, no. 6, pp. 739–757, 1994.

[33] H.-S. Lau and A.-H. Lau, "Estimating the demand distributions of single-period items having frequent stockouts," *European Journal of Operational Research*, vol. 92, no. 2, pp. 254–265, 1996.

[34] P. C. Bell, "Adaptive Sales Forecasting with Many Stockouts," *Journal of the Operational Research Society*, pp. 865–873, 1981.

[35] F. Petropoulos and N. Kourentzes, "Forecast combinations for intermittent demand," *Journal of the Operational Research Society*, vol. 66, no. 6, pp. 914–924, 2015.

[36] L. Shenstone and R. J. Hyndman, “Stochastic models underlying Croston’s method for intermittent demand forecasting,” *Journal of Forecasting*, vol. 24, no. 6, pp. 389–402, 2005.

[37] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[38] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York, NY, USA: Springer, 2009.

[39] C. Bergmeir and J. M. Benítez, “On the use of cross-validation for time series predictor evaluation,” *Information Sciences*, 2012.

[40] V. Cerqueira, L. Torgo, and I. Mozetič, “Evaluating time series forecasting models: an empirical study on performance estimation methods,” *Machine Learning*, 2020.

[41] L. J. Tashman, “Out-of-sample tests of forecasting accuracy: an analysis and review,” *International Journal of Forecasting*, 2000.

[42] R. Tibshirani, “Regression Shrinkage and Selection Via the Lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[44] T. Januschowski, J. Gasthaus, Y. Wang, D. Salinas, V. Flunkert, M. Bohlke-Schneider, and L. Callot, “Criteria for classifying forecasting methods,” *International Journal of Forecasting*, 2019.

[45] P. G. Giannopoulos, T. K. Dasaklis, I. Tsantilis, and C. Patsakis, “Machine learning algorithms in intermittent demand forecasting: a review,” *International Journal of Forecasting*, 2025.

[46] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, “DeepAR: Probabilistic forecasting with autoregressive recurrent networks,” *International Journal of Forecasting*, 2020.

[47] T. Januschowski, Y. Wang, K. Torkkola, T. Erkkilä, H. Hasson, and J. Gasthaus, “Forecasting with trees,” *International Journal of Forecasting*, 2022.

[48] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, 1st ed. New York, NY, USA: Chapman and Hall/CRC, 1984.

[49] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, pp. 123–140, 1996.

[50] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.

[51] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” *Machine Learning*, vol. 63, pp. 3–42, 2006.

[52] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[53] L. Mason, J. Baxter, P. Bartlett, and M. Frean, “Boosting algorithms as gradient descent,” in *Advances in Neural Information Processing Systems 12*, 1999, pp. 512–518.

[54] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[55] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems 30*, 2017.

[56] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” in *Advances in Neural Information Processing Systems 31*, 2018, pp. 6639–6649.

[57] Scikit-learn developers, “Ensembles: Gradient boosting, random forests, bagging, voting, and stacking,” *scikit-learn*. [Online]. Available: https://scikit-learn.org/stable/modules/ensemble.html. Accessed: May 30, 2026.

[58] D. H. Wolpert, “Stacked generalization,” *Neural Networks*, vol. 5, no. 2, pp. 241–259, 1992.

[59] M. J. van der Laan, E. C. Polley, and A. E. Hubbard, “Super Learner,” *Statistical Applications in Genetics and Molecular Biology*, vol. 6, no. 1, Art. 25, 2007.

[60] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[61] H. Hewamalage, C. Bergmeir, and K. Bandara, “Recurrent neural networks for time series forecasting: Current status and future directions,” *International Journal of Forecasting*, vol. 37, no. 1, pp. 388–427, 2021.

[62] J. L. Elman, “Finding structure in time,” *Cognitive Science*, vol. 14, no. 2, pp. 179–211, 1990.

[63] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[64] K. Cho, B. van Merriënboer, C. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.

[65] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499, 2016.

[66] S. Bai, J. Z. Kolter, and V. Koltun, “An empirical evaluation of generic convolutional and recurrent networks for sequence modeling,” arXiv preprint arXiv:1803.01271, 2018.

[67] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, “Temporal fusion transformers for interpretable multi-horizon time series forecasting,” *International Journal of Forecasting*, 2021.

[68] A. C. Cameron and P. K. Trivedi, *Regression Analysis of Count Data*, 2nd ed. Cambridge, U.K.: Cambridge University Press, 2014.

[69] D. Lambert, “Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing,” *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992.

[70] J. Mullahy, “Specification and testing of some modified count data models,” *Journal of Econometrics*, vol. 33, no. 3, pp. 341–365, 1986.

[71] J. G. Cragg, “Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods,” *Econometrica*, vol. 39, no. 5, pp. 829–844, Sep. 1971.

[72] A. C. Türkmen, T. Januschowski, Y. Wang, and A. T. Cemgil, “Forecasting intermittent and sparse time series: A unified probabilistic framework via deep renewal processes,” *PLOS ONE*, vol. 16, no. 11, p. e0259764, Nov. 2021.

[73] M. C. K. Tweedie, “An index which distinguishes between some important exponential families,” in *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, J. G. K. Ghosh and J. Roy, Eds. Calcutta, India: Indian Statistical Institute, 1984, pp. 579–604.

[74] B. Jørgensen, “Exponential dispersion models,” *J. R. Stat. Soc. B (Methodol.)*, vol. 49, no. 2, pp. 127–145, 1987.

[75] P. K. Dunn and G. K. Smyth, “Series evaluation of Tweedie exponential dispersion model densities,” *Stat. Comput.*, vol. 15, no. 4, pp. 267–280, Oct. 2005.

[76] G. K. Smyth and B. Jørgensen, “Fitting Tweedie’s compound Poisson model to insurance claims data: dispersion modelling,” *ASTIN Bull. J. IAA*, vol. 32, no. 1, pp. 143–157, 2002.

[77] N. Duan, “Smearing estimate: a nonparametric retransformation method,” *J. Am. Stat. Assoc.*, vol. 78, no. 383, pp. 605–610, 1983.

[78] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York, NY, USA: Springer, 2005.

[79] J. de Leeuw, K. Hornik, and P. Mair, “Isotone optimization in R: Pool-adjacent-violators algorithm (PAVA) and active set methods,” *J. Stat. Softw.*, vol. 32, no. 5, pp. 1–24, Oct. 2009.

[80] F. M. T. A. Busing, “Monotone regression: A simple and fast O(n) PAVA implementation,” *J. Stat. Softw.*, vol. 102, no. 1, pp. 1–25, May 2022.

[81] H. Jaeger, “The ‘echo state’ approach to analysing and training recurrent neural networks,” German National Research Center for Information Technology, GMD Tech. Rep. 148, Bonn, Germany, 2001.

[82] N. Kourentzes, “Intermittent demand forecasts with neural networks,” *Int. J. Prod. Econ.*, vol. 143, no. 1, pp. 198–206, 2013.

[83] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30*, 2017.

[84] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, “Geometric deep learning: going beyond Euclidean data,” *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, Jul. 2017.

[85] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61-80, Jan. 2009.

[86] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[87] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep Graph Infomax," in *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 2019.

[88] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41-75, 1997.

[89] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, Long Beach, CA, USA, 2017, pp. 1024-1034.

[90] C. Hoffmann, R. Lasch, and U. Meinig, "Forecasting of intermittent demand with artificial neural networks," *International Journal of Production Research*, vol. 60, no. 13, pp. 4252-4272, 2022.

[91] S. Wang, Y. Kang, and F. Petropoulos, "Combining probabilistic forecasts of intermittent demand," *European Journal of Operational Research*, vol. 315, no. 3, pp. 1038-1048, Jun. 2024.

[92] N. Kozodoi, E. Zinovyeva, S. Valentin, J. Pereira, and R. Agundez, "Probabilistic Demand Forecasting with Graph Neural Networks," *arXiv preprint arXiv:2401.13096*, Jan. 2024.

[93] Z. Fatemi, M. Huynh, E. Zheleva, Z. Syed, and X. Di, "Mitigating cold-start forecasting using cold causal demand forecasting model," *arXiv preprint arXiv:2306.09261*, Jun. 2023.

[94] R. H. Teunter, A. A. Syntetos, and M. Z. Babai, "Determining order-up-to levels under periodic review for compound binomial (intermittent) demand," *European Journal of Operational Research*, vol. 203, no. 3, pp. 619-624, Jun. 2010.

[95] A. A. Syntetos, M. Z. Babai, J. E. Boylan, S. Kolassa, and K. Nikolopoulos, "Supply chain forecasting: Theory, practice, their gap and the future," *European Journal of Operational Research*, vol. 252, no. 1, pp. 1-26, Jul. 2016.

[96] Ç. Pinçe, L. Turrini, and J. Meissner, "Intermittent demand forecasting for spare parts: A critical review," *Omega*, vol. 105, Art. no. 102513, Dec. 2021.

[97] M. Z. Babai, J. E. Boylan, and B. Rostami-Tabar, "Demand forecasting in supply chains: a review of aggregation and hierarchical approaches," *International Journal of Production Research*, vol. 60, no. 1, pp. 324-348, Dec. 2021.

[98] B. S. Nathan, P. M. Aravinth, B. V. S. Reddy, *et al.*, "Primacy of feature engineering over architectural complexity for intermittent demand forecasting," *Scientific Reports*, vol. 16, Art. no. 4792, 2026.

[99] R. Fildes, S. Ma, and S. Kolassa, "Retail forecasting: Research and practice," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1283-1318, 2022.

[100] S. Van der Auweraer, R. N. Boute, and A. A. Syntetos, "Forecasting spare part demand with installed base information: A review," *International Journal of Forecasting*, vol. 35, no. 1, pp. 181-196, Jan. 2019.

[101] N. Kourentzes, "On intermittent demand model optimisation and selection," *International Journal of Production Economics*, vol. 156, pp. 180-190, Oct. 2014.

[102] R. Teunter and B. Sani, "On the bias of Croston's forecasting method," *European Journal of Operational Research*, vol. 194, no. 1, pp. 177-183, Apr. 2009.

[103] A. A. Syntetos, M. Z. Babai, and E. S. Gardner Jr., "Forecasting intermittent inventory demands: simple parametric methods vs. bootstrapping," *Journal of Business Research*, vol. 68, no. 8, pp. 1746-1752, Aug. 2015.

[104] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 Competition: 100,000 time series and 61 forecasting methods," *International Journal of Forecasting*, vol. 36, no. 1, pp. 54-74, Jan. 2020.

[105] M. A. Hoffmann, R. Lasch, and J. Meinig, "Forecasting irregular demand using single hidden layer neural networks," *Logistics Research*, vol. 15, no. 1, Art. no. 6, pp. 1-13, 2022.

[106] R. S. Gutierrez, A. O. Solis, and S. Mukhopadhyay, "Lumpy demand forecasting using neural networks," *International Journal of Production Economics*, vol. 111, no. 2, pp. 409-420, Feb. 2008.

[107] M. Z. Babai, A. Tsadiras, and C. Papadopoulos, "On the empirical performance of some new neural network methods for forecasting intermittent demand," *IMA Journal of Management Mathematics*, vol. 31, no. 3, pp. 281-305, Jul. 2020.

[108] S. Ma and R. Fildes, "Retail sales forecasting with meta-learning," *European Journal of Operational Research*, vol. 288, no. 1, pp. 111-128, Jan. 2021.

[109] S. Praveena and S. Prasanna Devi, "A hybrid demand forecasting for intermittent demand patterns using machine learning techniques," in *Proc. 2022 1st International Conference on Computational Science and Technology (ICCST)*, Chennai, India, 2022, pp. 557-561.

[110] K. K. Chandriah and R. V. Naraganahalli, "RNN / LSTM with modified Adam optimizer in deep learning approach for automobile spare parts demand forecasting," *Multimedia Tools and Applications*, vol. 80, pp. 26145-26159, 2021.

[111] F. Muşat and S. Căbuz, "Switch-Hurdle: A MoE Encoder with AR Hurdle Decoder for Intermittent Demand Forecasting," *arXiv preprint arXiv:2602.22685*, Feb. 2026.

[112] M. Tan, M. A. Merrill, V. Gupta, T. Althoff, and T. Hartvigsen, "Are language models actually useful for time series forecasting?," in *Proc. Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

[113] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2019, pp. 2623-2631.

[114] A. D. Lainder and R. D. Wolfinger, "Forecasting with gradient boosted trees: Augmentation, tuning, and cross-validation strategies: Winning solution to the M5 uncertainty competition," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1426-1433, 2022.

[115] Y. Kang, W. Cao, F. Petropoulos, and F. Li, "Forecast with forecasts: Diversity matters," *European Journal of Operational Research*, vol. 301, no. 1, pp. 180-190, Jul. 2022.

[116] L. Li, Y. Kang, F. Petropoulos, and F. Li, "Feature-based intermittent demand forecast combinations: accuracy and inventory implications," *International Journal of Production Research*, vol. 61, no. 22, pp. 7557-7572, Nov. 2023.

[117] S. T. Manjunath, "Demand forecast optimization using machine learning algorithms," M.S. thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2025.

[118] P. Dodin, J. Xiao, Y. Adulyasak, N. Etebari Alamdari, L. Gauthier, P. Grangier, P. Lemaitre, and W. L. Hamilton, "Bombardier aftermarket demand forecast with machine learning," *INFORMS Journal on Applied Analytics*, vol. 53, no. 6, pp. 425-445, Nov. 2023.

[119] A. Gandhi, Aakanksha, S. Kaveri, and V. Chaoji, "Spatio-temporal multi-graph networks for demand forecasting in online marketplaces," in *Proc. Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track (ECML PKDD 2021)*, vol. 12978, Y. Dong, N. Kourtellis, B. Hammer, and J. A. Lozano, Eds. Cham: Springer, 2021.

[120] W. Fu and C.-F. Chien, "UNISON data-driven intermittent demand forecast framework to empower supply chain resilience and an empirical study in electronics distribution," *Computers & Industrial Engineering*, vol. 135, pp. 240-256, Sep. 2019.

[121] D. Kiefer, F. Grimm, M. Bauer, and C. Van Dinther, "Demand forecasting intermittent and lumpy time series: Comparing statistical, machine learning and deep learning methods," in *Proc. 27th Americas Conference on Information Systems (AMCIS)*, Montreal, QC, Canada, 2021.

[122] P. G. Giannopoulos, T. K. Dasaklis, and N. Rachaniotis, "Development and evaluation of a novel framework to enhance k-NN algorithm's accuracy in data sparsity contexts," *Scientific Reports*, vol. 14, Art. no. 25036, 2024.

[123] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. Multiple Classifier Systems (MCS)*, vol. 1857, Berlin, Heidelberg: Springer, 2000, pp. 1-15.

[124] J. M. Rožanec, B. Fortuna, and D. Mladenić, "Reframing demand forecasting: A two-fold approach for lumpy and intermittent demand," *Sustainability*, vol. 14, no. 15, Art. no. 9295, Aug. 2022.

[125] L. A. C. G. Andrade and C. B. Cunha, "Disaggregated retail forecasting: A gradient boosting approach," *Applied Soft Computing*, vol. 143, art. no. 110283, 2023.

[126] J. Gasthaus, K. Benidis, Y. Wang, S. S. Rangapuram, D. Salinas, V. Flunkert, and T. Januschowski, "Probabilistic forecasting with spline quantile function RNNs," in *Proc. 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, vol. 89, 2019, pp. 1901-1910.

[127] S. Smyl, "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting," *International Journal of Forecasting*, vol. 36, no. 1, pp. 75-85, Jan. 2020.

[128] E. Spiliotis, S. Makridakis, A. A. Semenoglou, *et al.*, "Comparison of statistical and machine learning methods for daily SKU demand forecasting," *Operational Research*, vol. 22, pp. 3037-3061, 2022.

[129] E. Spiliotis, S. Makridakis, A. Kaltsounis, and V. Assimakopoulos, "Product sales probabilistic forecasting: An empirical evaluation using the M5 competition data," *International Journal of Production Economics*, vol. 240, Art. no. 108237, Oct. 2021.

[130] A. A. Ghobbar and C. H. Friend, "Evaluation of forecasting methods for intermittent parts demand in the field of aviation: A predictive model," *Computers & Operations Research*, vol. 30, no. 14, pp. 2097-2114, Dec. 2003.

[131] A. A. Syntetos, K. Nikolopoulos, and J. E. Boylan, "Judging the judges through accuracy-implication metrics: The case of inventory forecasting," *International Journal of Forecasting*, vol. 26, no. 1, pp. 134-143, Jan.-Mar. 2010.

[132] I. Tsantilis, P. G. Giannopoulos, T. K. Dasaklis, and C. Patsakis, "Forecasting intermittent demand using large language models: Evidence from a real world industrial dataset," in *Proc. IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, Glasgow, United Kingdom, 2025, pp. 1-8.

[133] Y. Liu, G. Qin, X. Huang, J. Wang, and M. Long, "AutoTimes: Autoregressive time series forecasters via large language models," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, 2024.

[134] C. Sun, H. Li, Y. Li, and S. Hong, "TEST: Text prototype aligned embedding to activate LLM's ability for time series," in *Proc. International Conference on Learning Representations (ICLR)*, 2024.

[135] Y. Bian, X. Ju, J. Li, Z. Xu, D. Cheng, and Q. Xu, "Multi-patch prediction: Adapting language models for time series representation learning," in *Proc. 41st International Conference on Machine Learning (ICML)*, Vienna, Austria, vol. 235, 2024.

[136] M. Germán-Morales, A. J. Rivera-Rivas, M. J. del Jesus Díaz, and C. J. Carmona, "Transfer learning with foundational models for time series forecasting using low-rank adaptations," *Information Fusion*, vol. 123, Art. no. 103247, Nov. 2025.

[137] M. Lukoševičius and H. Jaeger, "Reservoir computing approaches to recurrent neural network training," *Computer Science Review*, vol. 3, no. 3, pp. 127-149, Aug. 2009.

[138] J. Tsantilis, P. G. Giannopoulos, T. K. Dasaklis, and C. Patsakis, "An analytics-driven approach to sporadic demand forecasting using reservoir computing," *SSRN Preprint*, no. 5209158, 2025.

[139] I. Tsantilis, T. K. Dasaklis, C. Douligeris, and C. Patsakis, "Searching deterministic chaotic properties in system-wide vulnerability datasets," *Informatics*, vol. 8, no. 4, Art. no. 86, Dec. 2021.

[140] M. Z. Babai, A. Syntetos, and R. Teunter, "Intermittent demand forecasting: An empirical study on accuracy and the risk of obsolescence," *International Journal of Production Economics*, vol. 157, pp. 212-219, Nov. 2014.

[141] P. H. Zipkin, *Foundations of Inventory Management*. New York, NY, USA: McGraw-Hill Irwin, 2000.

[142] S. Axsäter, *Inventory Control*, 3rd ed. Cham, Switzerland: Springer, 2015.

[143] J. M. Juran, F. M. Gryna, and R. S. Bingham, *Quality Control Handbook*, vol. 3. New York, NY, USA: McGraw-Hill, 1979, p. 2.

[144] B. E. Flores and D. C. Whybark, "Implementing multiple criteria ABC analysis," *Journal of Operations Management*, vol. 7, no. 1-2, pp. 79-85, Oct. 1987.

[145] C. Anderson, "The long tail," *Wired*, vol. 12, no. 10, pp. 170-177, Oct. 2004.

[146] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual: Python Documentation Manual Part 2*. Scotts Valley, CA, USA: CreateSpace, 2009.

[147] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, pp. 357-362, Sep. 2020.

[148] W. McKinney, "Data structures for statistical computing in Python," in *Proc. 9th Python in Science Conference (SciPy)*, Austin, TX, USA, 2010, pp. 56-61.

[149] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[150] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *Proc. ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[151] W. Wang, F. Wei, L. Dong, H. Bao, X. Huang, and M. Zhou, "MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained Transformers," in *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020.

[152] S. K. Lam, A. Pitrou, and S. Seibert, "Numba: A LLVM-based Python JIT compiler," in *Proc. 2nd Workshop on the LLVM Compiler Infrastructure in HPC*, Austin, TX, USA, 2015, pp. 1-6.

[153] E. A. Silver, D. F. Pyke, and R. Peterson, *Inventory Management and Production Planning and Scheduling*, 3rd ed. New York, NY, USA: Wiley, 1998, p. 30.

[154] C. W. Chase, *Demand-Driven Forecasting: A Structured Approach to Forecasting*. Hoboken, NJ, USA: John Wiley & Sons, 2013.

[155] A. A. Syntetos, K. Nikolopoulos, J. E. Boylan, R. Fildes, and P. Goodwin, "The effects of integrating management judgement into intermittent demand forecasts," *International Journal of Production Economics*, vol. 118, no. 1, pp. 72-81, Mar. 2009.

[156] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *The Semantic Web* (*Lecture Notes in Computer Science*, vol. 10843), A. Gangemi *et al.*, Eds. Cham, Switzerland: Springer, 2018.

[157] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[158] B. Yang, W.-T. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," *arXiv preprint arXiv:1412.6575*, 2014.

[159] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Sardinia, Italy, 2010, pp. 249-256.

[160] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 26, 2013.

[161] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 29, 2016.

[162] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," *arXiv preprint arXiv:2402.03216*, 2024.

[163] Amazon Web Services, *Amazon Bedrock User Guide: Amazon Titan Text Embeddings v2*. Seattle, WA, USA, 2024. [Online]. Available: https://docs.aws.amazon.com/bedrock/latest/userguide/titan-models.html. [Accessed: May 24, 2026].

[164] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A*, vol. 374, no. 2065, 2016.

[165] M. Strathern, "'Improving ratings': audit in the British University system," *European Review*, vol. 5, no. 3, pp. 305-321, Jul. 1997.

[166] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[167] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[168] R. G. Brown, *Statistical Forecasting for Inventory Control*. New York, NY, USA: McGraw-Hill, 1959.

[169] E. S. Gardner Jr., "Exponential smoothing: The state of the art," *Journal of Forecasting*, vol. 4, no. 1, pp. 1-28, 1985.

[170] Y. Zhang, L. Ma, S. Pal, Y. Zhang, and M. Coates, "Multi-resolution time-series transformer for long-term forecasting," in *Proc. 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2024, pp. 4222-4230.

[171] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422-446, 2002.

[172] G. Shani and A. Gunawardana, "Evaluating Recommendation Systems," in *Recommender Systems Handbook*, 1st ed., F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Boston, MA, USA: Springer, 2011, pp. 257–297, doi: 10.1007/978-0-387-85820-3_8.

ΠΑΡΑΡΤΗΜΑ Α : ΑΡΧΕΙΟ ΡΥΘΜΙΣΕΩΝ ΚΥΡΙΑΣ ΡΟΗΣ ΕΡΓΑΣΙΩΝ

```
cfg = PipelineConfig(
    data_path = Path('../data/ML_DATASET_20260107.csv'),
    forecast_origin = pd.Timestamp('2025-01-01'),
    output_dir = Path('../results'),
    cache_dir=Path('../data'),
    supplier_metadata_path = Path('../data/suppliers_unique.csv'),
    horizons = (15, 30, 45, 60, 75, 90, 180, 365),
    nrows = None,
    seed = 42,

    # model selection and training
    models = ("Hurdle_RF", "Hurdle_ET", "Hurdle_HGB", "Hurdle_CAT"),
    snapshot_semi_annual = True,
    preferred_model_by_horizon = None,
    selected_forecast_scale_by_horizon = None,
    ensemble_exclude_models = (),
    # model caching to avoid re-training on every run
    enable_model_cache = True
    model_cache_dir = None,  # Defaults to output_dir/models
    active_sku_lookback_days = None,

    # feature engennering
    use_critical_rolling_features = True,
    # Seasonality + value enrichment for snapshot features.
    use_seasonality_features = True
    use_holiday_flags = True
    use_value_features = True
    use_intermittent_features = True
      #  Extended  intermittent  metrics  alongside  standard  WAPE/Score  (sMAE,
HitRate_Binary, Cond_WAPE).
    enable_intermittent_metrics = True,

    # demand censoring
    use_oos_features = True,
    stock_feature_horizons = (15, 30, 45, 60, 75, 90),
    enable_demand_censoring = True,
    censoring_horizons = (15, 30, 45, 60, 75),
    stock_aux_paths = [
        Path('../data/ml_dataset_2024.csv'),
        Path('../data/ml_dataset_2025.csv'),
    ],
```

```
    # embeddings
    use_transformer_embeddings = True,
                                transformer_embeddings_cache_path               =
Path('../data/transformer_embeddings.pkl'),
    embedding_prefix = 'emb_',
      transformer_model_name  =  "sentence-transformers/paraphrase-multilingual-
MiniLM-L12-v2",
    transformer_embedding_batch_size = 128,
    transformer_max_length = 128,
    transformer_force_recompute = False,
    transformer_embeddings_cache_path = None,

    # calibration
    enable_calibration = True,
    calibration_horizons = (15, 30, 45, 60, 75, 90, 180, 365),
    enable_train_scaling = True,
    train_scale_clip = (0.6, 1.4),
    enable_stratified_scaling = True,
    stratified_scaling_top_n_candidates = (50, 75, 100),
    stratified_scaling_min_revenue_share = 0.05,
    stratified_top_clip = (0.85, 1.60),
    stratified_top_blend_alpha = 0.70,
    enforce_monotone_horizon_predictions = True,
    recency_half_life_by_horizon = {15: 30.0, 30: 60.0, 45: 60.0, 60: 75.0, 75:
90.0, 90: 90.0, 180: 180.0, 365: 365.0},
    calibration_half_life_candidates_by_horizon = {
        15: (None, 30.0, 45.0, 60.0),
        30: (None, 60.0, 75.0, 90.0),
        45: (None, 60.0, 90.0, 105.0),
        60: (None, 75.0, 90.0, 120.0),
        75: (None, 90.0, 105.0, 120.0),
        90: (None, 90.0, 105.0, 120.0),
        180: (None, 120.0, 180.0, 270.0),
        365: (None, 270.0, 365.0, 540.0),
    },
    calibration_scale_clip_by_horizon = {
        15:  (0.85, 1.08),
        30:  (0.85, 1.15),
        45:  (0.90, 1.25),
        60:  (0.90, 1.27),
        75:  (0.93, 1.27),
        90:  (0.93, 1.30),
        180: (0.95, 1.30),
        365: (0.95, 1.40),
    },
```

```
    enable_weighted_calibration = True,
    weighted_calibration_top_n = 1500,
    weighted_calibration_top_weight = 8.0,
    weighted_calibration_by_horizon = {
        15:  {"top_n": 2000, "top_weight": 5.0},
        30:  {"top_n": 1500, "top_weight": 4.5},
        45:  {"top_n": 1200, "top_weight": 3.0},
        60:  {"top_n": 1000, "top_weight": 3.0},
        75:  {"top_n": 1000, "top_weight": 3.0},
        90:  {"top_n": 800,  "top_weight": 3.0},
        180: {"top_n": 600,  "top_weight": 2.5},
        365: {"top_n": 400,  "top_weight": 2.0},
    },
    category_calibration = False,
    # Threshold tuning should use a small holdout slice from the training window,
    # not the same rows used for the classifier fit.
    enable_classifier_threshold_tuning = True,
    classifier_threshold_validation_fraction = 0.05,

    # backtest
    load_backtest = True,
    backtest_max_cutoffs = 3,
    enforce_backtest_cache_fingerprint = True,
    backtest_model_params_override = {
       "Hurdle_CAT": {"clf": {"iterations": 600},  "reg": {"iterations": 1200}},
         "Hurdle_RF":  {"clf": {"n_estimators": 200}, "reg": {"n_estimators":
200}},
         "Hurdle_ET":  {"clf": {"n_estimators": 200}, "reg": {"n_estimators":
200}},
    },
    backtest_model_params_override = None,
    use_backtest_model_selection = True,
    backtest_selection_ratio_weight = 0.40,
    backtest_selection_smoothness_weight = 0.20,

    # Prediction interval (residual-quantile) estimation.
    enable_prediction_intervals = True,
    interval_quantiles = (0.1, 0.9),

    # performance flags
    enable_plots = True,
    plot_show = True,
    plot_save = True,
```

```
    # business decisions
    # Safety stock export
    enable_safety_stock_export = True
    safety_stock_service_levels = (0.90,)
    default_supplier_lead_time_days = 30.0
    supplier_lead_time_column_candidates = (
        "Lead_Time_Days",
        "LeadTimeDays",
        "Supplier_Lead_Time_Days",
        "Lead_Time",
        "LeadTime",
    ),
    default_supplier_lead_time_days = 30.0,
    # Horizons to include in Top-20 quantity reporting (for high-mobility SKUs)
    horizons_top_qty = (15, 30, 45, 60),

    # optuna hyperparameters tuning
    enable_optuna = False,
    optuna_model = ["Hurdle_RF", "Hurdle_ET", "Hurdle_HGB", "Hurdle_CAT"],
    optuna_horizon = [15, 30, 90],
    optuna_trials = 30,
    optuna_max_cutoffs = 3,
    optuna_timeout = None

    # Echo State Network forecast features
    use_esn_features = False,
    esn_horizons = (60, 75, 90, 180, 365),
    enable_esn_tuning = False,
    esn_tuning_sample_skus = 300,
    esn_params_by_horizon = {
            60:  {"n_reservoir":  128,  "lookback_days":  365,  "alpha":  5e-3,
"spectral_radius": 1.0, "sparsity": 0.10},
            75:  {"n_reservoir":  128,  "lookback_days":  365,  "alpha":  5e-3,
"spectral_radius": 1.0, "sparsity": 0.10},
            90:  {"n_reservoir":  128,  "lookback_days":  365,  "alpha":  5e-3,
"spectral_radius": 1.0, "sparsity": 0.10},
           180:  {"n_reservoir":  256,  "lookback_days":  730,  "alpha":  3e-3,
"spectral_radius": 1.1, "sparsity": 0.08},
           365:  {"n_reservoir":  256,  "lookback_days":  1095,  "alpha":  3e-3,
"spectral_radius": 1.1, "sparsity": 0.08},
    },
    esn_normalize_features = True, # Standardize ESN features
    esn_n_jobs = -1,
```

```
    # PCA dimensionality reduction for rolling features (reduces noise, improves
bias)
    use_pca_features = False,
     pca_n_components = 50, #Number of PCA components to extract from rolling
features
    pca_drop_source_features = True

)
```