ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ & ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ

Διπλωματική Εργασία

Πρόβλεψη συμπεριφοράς πελάτη σούπερ μάρκετ λιανικής με ανάπτυξη classification σε γλώσσα python χρησιμοποιώντας dataset συμπεριφορικών και δημογραφικών δεδομένων

Επιβλέπουσα: **Ειρήνη Τριάρχη**

Μέλη Επιτροπής:
Γεώργιος Κόλιας
Παρασκευή Παππά

Ονοματεπώνυμο: **Παππάς Χρηστάκης**
ΑΜ: **357**

Δεκέμβριος 2025,
Πρέβεζα

Prediction of Retail Supermarket Customer Behavior through the Development of Classification Models in Python Using Behavioral and Demographic Datasets

Εγκρίθηκε από τριμελή εξεταστική επιτροπή
Πρέβεζα, 17/12/2025

ΕΠΙΤΡΟΠΗ ΑΞΙΟΛΟΓΗΣΗΣ

1. Επιβλέπουσα Καθηγήτρια:

   Ειρήνη Τριάρχη

   Επίκουρη Καθηγήτρια

2. Μέλος επιτροπής

   Γεώργιος Κόλιας

   Αναπληρωτής Καθηγητής

3. Μέλος επιτροπής

   Παρασκευή Παππά

   Λέκτορας

Δήλωση μη λογοκλοπής

Δηλώνω υπεύθυνα και γνωρίζοντας τις κυρώσεις του Ν. 2121/1993 περί Πνευματικής Ιδιοκτησίας, ότι η παρούσα μεταπτυχιακή εργασία είναι εξ ολοκλήρου αποτέλεσμα δικής μου ερευνητικής εργασίας, δεν αποτελεί προϊόν αντιγραφής ούτε προέρχεται από ανάθεση σε τρίτους. Όλες οι πηγές που χρησιμοποιήθηκαν (κάθε είδους, μορφής και προέλευσης) για τη συγγραφή της περιλαμβάνονται στη βιβλιογραφία.

Παππάς, Χρηστάκης

ΕΥΧΑΡΙΣΤΙΕΣ

Στην Ιωάννα, την Ευαγγελία και την Μαρίνα..

ΠΕΡΙΕΧΟΜΕΝΑ

ΛΙΣΤΑ ΕΙΚΟΝΩΝ

ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία εξετάζει την πρόβλεψη της συμπεριφοράς απόκρισης πελατών στο πλαίσιο του λιανικού εμπορίου σούπερ μάρκετ, αξιοποιώντας τεχνικές εποπτευόμενης μηχανικής μάθησης και δεδομένα δημογραφικού και συμπεριφορικού χαρακτήρα. Στόχος της μελέτης είναι η ανάπτυξη και συγκριτική αξιολόγηση πολλαπλών αλγορίθμων ταξινόμησης, με σκοπό την ακριβή εκτίμηση της πιθανότητας αποδοχής μιας προωθητικής ενέργειας και τη διερεύνηση της επιχειρησιακής τους χρησιμότητας.

Η ανάλυση βασίζεται σε δημόσια διαθέσιμο σύνολο δεδομένων, το οποίο υποβάλλεται σε εκτενή προεπεξεργασία, συμπεριλαμβανομένου του καθαρισμού, της κωδικοποίησης κατηγορικών μεταβλητών, της κλιμάκωσης χαρακτηριστικών και της αντιμετώπισης της ανισορροπίας κλάσεων. Στο πειραματικό πρωτόκολλο εντάσσονται διάφοροι ταξινομητές, όπως Λογιστική Παλινδρόμηση, Δέντρα Απόφασης, Random Forests, Extra Trees, K-Nearest Neighbors, Support Vector Machines, Naive Bayes, AdaBoost, Gradient Boosting και Νευρωνικά Δίκτυα. Η αξιολόγηση πραγματοποιείται σε κοινό σύνολο δοκιμής, με τη χρήση καθιερωμένων μετρικών απόδοσης, όπως Accuracy, Precision, Recall, F1-score και ROC–AUC, επιτρέποντας άμεση και συνεπή σύγκριση.

Τα αποτελέσματα καταδεικνύουν ότι τα ensemble μοντέλα και οι μέθοδοι μέγιστου περιθωρίου υπερέχουν σε όρους συνολικής διακριτικής ικανότητας, ενώ απλούστερα μοντέλα διατηρούν πλεονεκτήματα ερμηνευσιμότητας και σταθερότητας. Η εργασία ολοκληρώνεται με συζήτηση των περιορισμών, σύνδεση με τη σχετική βιβλιογραφία και ανάλυση των επιχειρησιακών συνεπειών, αναδεικνύοντας πώς τα προγνωστικά σκορ μπορούν να υποστηρίξουν πολιτικές στοχευμένου μάρκετινγκ, υπό ρεαλιστικούς περιορισμούς κόστους και δεοντολογίας.

Λέξεις κλειδιά: Μηχανική μάθηση, ταξινόμηση, συμπεριφορά πελατών, λιανικό εμπόριο, προγνωστική ανάλυση, Python, marketing analytics

ABSTRACT

This thesis investigates the prediction of customer response behavior in the context of retail supermarket marketing by employing supervised machine learning techniques on demographic and behavioral data. The primary objective of the study is the development and comparative evaluation of multiple classification algorithms in order to accurately estimate the probability of customer acceptance of promotional campaigns and to assess their operational relevance.

The analysis is conducted on a publicly available dataset, which undergoes extensive preprocessing, including data cleaning, categorical encoding, feature scaling, and treatment of class imbalance. The experimental protocol incorporates a wide range of classifiers, namely Logistic Regression, Decision Trees, Random Forests, Extra Trees, K-Nearest Neighbors, Support Vector Machines, Naive Bayes, AdaBoost, Gradient Boosting and Neural Networks. Model evaluation is performed on a common test set using standard performance metrics such as Accuracy, Precision, Recall, F1-score, and ROC–AUC, enabling consistent and transparent comparison across methods.

The results indicate that ensemble-based models and margin-based classifiers exhibit superior discriminative performance and robustness, while simpler models retain advantages in interpretability and stability. The findings are largely consistent with existing literature on retail response modeling, particularly in settings characterized by mixed-type features and imbalanced target classes.

The thesis concludes with a discussion of methodological limitations, alignment with prior research, and an analysis of the business implications of the results. Emphasis is placed on how predictive scores can be operationalized to support targeted marketing strategies, threshold-based decision rules and cost–benefit–aware campaign design, while adhering to practical constraints related to data governance and ethical considerations.

Key words: Machine learning, classification, customer response prediction, retail analytics, predictive modeling, Python, marketing analytics.

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

Το σύγχρονο λιανικό εμπόριο και ειδικότερα ο κλάδος των σούπερ μάρκετ, λειτουργεί σε ένα περιβάλλον έντονου ανταγωνισμού, παρουσίας σε πληθώρα καναλιών και συνεχούς ροής δεδομένων από φυσικά καταστήματα και ψηφιακά σημεία επαφής. Η ικανότητα μιας επιχείρησης να μετατρέπει αυτά τα δεδομένα σε αξιόπιστες προβλέψεις συμπεριφοράς πελατών αποτελεί κρίσιμο συντελεστή κερδοφορίας. Αυτό συμβαίνει καθώς βελτιώνει την αποτελεσματικότητα των προωθητικών ενεργειών, μειώνει το κόστος επαφής και στηρίζει αποφάσεις σε πραγματικό χρόνο για αποθέματα, τιμολόγηση και προσωποποιημένη επικοινωνία. Στο πλαίσιο αυτό, η παρούσα εργασία εστιάζει στην πρόβλεψη της ανταπόκρισης πελατών σε καμπάνιες σούπερ μάρκετ, αξιοποιώντας μοντέλα ταξινόμησης σε δομημένα δημογραφικά και συμπεριφορικά δεδομένα, με στόχο τη δημιουργία ενός πρακτικά εφαρμόσιμου εργαλείου υποστήριξης αποφάσεων μάρκετινγκ.

Το ερευνητικό κενό που επιχειρεί να καλύψει η μελέτη δεν είναι αμιγώς αλγοριθμικό. Πολλές προσεγγίσεις στέκονται στη σύγκριση μοντέλων χωρίς να εξασφαλίζουν αυστηρή μεθοδολογία και χωρίς να μεταφράζουν συστηματικά τις πιθανότητες πρόβλεψης σε επιχειρησιακούς κανόνες (κατώφλια βάσει κόστους και οφέλους, επιλογή καναλιού, μέγεθος παρτίδας κ.ά.), κάτι που περιορίζει την αξιοπιστία και την εφαρμογή των ευρημάτων στην πράξη. Εδώ, η συμβολή μετατοπίζεται στη γεφύρωση της πρόβλεψης και της απόφασης. Από τον καθορισμό των χαρακτηριστικών και την εκτέλεση ταξινομητών έως την επιλογή τελικού μοντέλου και τον τρόπο με τον οποίο το σκορ μετατρέπεται σε στοχευμένες ενέργειες μάρκετινγκ.

Ως συνέπεια, οι στόχοι της εργασίας είναι αρχικά να διαμορφωθεί ένα σαφές και ικανό ευρείας αναπαραγωγής πλαίσιο προεπεξεργασίας και εκπαίδευσης ταξινομητών για πρόβλεψη ανταπόκρισης. Έπειτα είναι να συγκριθούν αντιπροσωπευτικά μοντέλα (γραμμικά, βασισμένα σε γειτνίαση, δέντρα και σύνολα - ensembles) υπό κοινές μετρικές αξιολόγησης. Ακόμη στοχεύει να προκρίνει τελικό μοντέλο με κριτήρια επιχειρησιακής χρησιμότητας και σταθερότητας, ενώ κρίνεται ωφέλιμο να μεταφραστούν οι προβλεπτικές πιθανότητες σε κανόνες απόφασης, οι οποίοι μεγιστοποιούν το αναμενόμενο όφελος υπό ρεαλιστικούς περιορισμούς κόστους και καναλιών. Τα αντίστοιχα ερευνητικά ερωτήματα αφορούν τη σχετική επίδοση των

μοντέλων σε ανισόρροπο στόχο, την επίδραση κρίσιμων χαρακτηριστικών, όπως Recency και δείκτες δαπάνης, στη διάκριση των τάξεων και τον τρόπο επιλογής κατωφλίου και Top-K. Έτσι τεκμηριώνεται η ισορροπία ανάμεσα σε recall και οικονομική αποδοτικότητα.

Μεθοδολογικά, η εργασία οργανώνεται σε τρία διαδοχικά επίπεδα. Αρχικά στον σχεδιασμό και στα δεδομένα, όπου ορίζονται ο στόχος (Response) και το λεξικό πεδίων. Έπειτα στην προεπεξεργασία και στα χαρακτηριστικά, με διαλογή, κωδικοποίηση κατηγορικών και τυποποίηση αριθμητικών μεταβλητών. Τελευταίο επίπεδο είναι η εκτέλεση ταξινομητών με κοινή ροή και συγκριτική αποτίμηση σε τυποποιημένες μετρικές (Precision, Recall, F1, Accuracy και ROC–AUC). Η αξιολόγηση συγκεντρώνεται στο test set, ενώ η επιλογή τελικού μοντέλου συνδέεται ρητά με κανόνες απόφασης που λαμβάνουν υπόψη το κόστος επαφής και τις ιδιαιτερότητες καναλιών (email, SMS, τηλέφωνο).

Η δομή του κειμένου ακολουθεί τη γνωστή λογική μετάβασης από τη μεθοδολογία στα αποτελέσματα, από εκεί στη συζήτηση και αξιολόγηση και καταλήγει στις προτάσεις. Μετά το Θεωρητικό Πλαίσιο και τη Βιβλιογραφική Ανασκόπηση, το Κεφάλαιο 4 τεκμηριώνει τον σχεδιασμό, την προεπεξεργασία και το εκτελεστικό πλαίσιο των ταξινομητών. Το Κεφάλαιο 5 παρουσιάζει την εξερεύνηση και οπτικοποίηση, την εκτέλεση των μοντέλων και τον συγκριτικό πίνακα μετρικών, καταλήγοντας στην επιλογή τελικού μοντέλου. Το Κεφάλαιο 6 μεταφράζει τα ευρήματα και πραγματεύεται την ερμηνεία τους, κατάδειξη περιορισμών και πως αυτοί επηρεάζουν την σταθερότητα, τη συσχέτιση με βιβλιογραφία και επιχειρησιακές κυρίως συνέπειες (κατώφλι βάσει κόστους - οφέλους, επιλογή καναλιού, batch size). Καταλήγοντας, το Κεφάλαιο 7 συνοψίζει συμπεράσματα και χαρτογραφεί μελλοντικές κατευθύνσεις.

ΚΕΦΑΛΑΙΟ 2. ΘΕΩΡΗΤΙΚΟ ΠΛΑΙΣΙΟ

2.1 Λιανικό εμπόριο και συμπεριφορά των πελατών

Το λιανικό εμπόριο στη σύγχρονη εποχή χαρακτηρίζεται από έντονο ανταγωνισμό και παρουσία σε πληθώρα καναλιών, τα οποία παρέχουν συνεχή ροή δεδομένων. Οι αλληλεπιδράσεις των πελατών πλημμυρίζουν τις εισροές των αλγορίθμων επεξεργασίας μέσω φυσικών καταστημάτων, ψηφιακών καναλιών και υβριδικών μορφών εξυπηρέτησης, όπως *click & collect* κ.ά.. Η διαθεσιμότητα μεγάλου όγκου και ποικιλίας δεδομένων επιτρέπει την ανάπτυξη αναλυτικών εργαλείων που στηρίζουν τη λήψη αποφάσεων, ακόμη και σε πραγματικό χρόνο. Η κατανόηση της δυναμικής σε αυτό το πλαίσιο είναι απαραίτητη προϋπόθεση για στοχευμένες παρεμβάσεις που προσθέτουν τέτοια αξία στην επιχείρηση, ικανή να μετασχηματίσει τις πρακτικές της. Έτσι, η ανάλυση της συμπεριφοράς των πελατών καθίσταται κεντρικό εργαλείο για την κατανόηση προτιμήσεων, την κατηγοριοποίηση των πελατών και καταληκτικά στην εξατομίκευση προσφορών.

Στο εμπόριο λιανικής οι αποφάσεις σχετικά με την αγορά ενός προϊόντος είναι συχνές και δεν περιέχουν μεγάλο ρίσκο, ενώ διαμορφώνονται κυρίως από την τιμή, το προωθητικό πλαίσιο και την προσβασιμότητα σε αυτό, στοιχεία που παραπέμπουν έντονα στο μίγμα μάρκετινγκ (Esfandiari κ.ά., 2025). Συνεπώς, τη συμπεριφορά των αγοραστών συνθέτουν δημογραφικοί παράγοντες, καταναλωτικές συνήθειες και ερεθίσματα από το κατάστημα και την ψηφιακή τους περιήγηση (Yokoyama, Azuma & Kim, 2022). Ως εκ τούτου, η πρόβλεψη ανταπόκρισης σε ενέργειες μάρκετινγκ ανέρχεται πια σε επιχειρησιακό ζητούμενο, με σημαντικό ανταγωνιστικό χαρακτήρα.

Στην προσπάθεια κατανόησης του καταναλωτή, η μετάβαση σε πρακτικές που έχουν στο επίκεντρο τα δεδομένα επιτρέπει την ποσοτικοποίηση της πιθανότητας ενέργειας, όπως η ανταπόκριση σε προσφορά που πραγματεύεται και η παρούσα εργασία, αλλά και την πλήρη διαμόρφωση επιχειρηματικών αποφάσεων, από τον σχεδιασμό του χώρου πώλησης μέχρι την διακοπή εμπορίας προϊόντων. Τεχνικές όπως η ομαδοποίηση (*clustering*), οι κανόνες συσχέτισης, η ανάλυση χρονοσειρών και γενικότερα τα επί μέρους στοιχεία των μοντέλων ταξινόμηση, καθώς και τα μοντέλα καθαυτά, προσφέρουν εκτιμήσεις και πρακτικές απαντήσεις σε ζωτικά επιχειρησιακά ζητήματα (Silverio κ.ά., 2025). Αυτά δύνανται να είναι ζητήματα που άπτονται τιμολόγησης, προώθησης, διαχείρισης αποθεμάτων, ή ακόμη σε επίπεδο στρατηγικής να υπαγορεύουν πολιτικές στόχευσης,

διαχείρισης εκπτώσεων, κ.ά. (Silverio κ.ά., 2025 ; Hagberg κ.ά., 2016). Ωστόσο, η αξιοπιστία αυτών των πληροφοριών για τη λήψη αποφάσεων με στόχο να αυξηθεί ο λόγος μετατροπής επαφής σε πώληση, είναι άρρηκτα συνδεδεμένη με την ορθή προεπεξεργασία των δεδομένων, τη χρήση κατάλληλων μετρικών, αλλά και την αποφυγή μεροληπτικών συμπεριφορών (*bias*).

Το πλαίσιο της λιανικής πώλησης λοιπόν, προσεγγίζεται ως γέφυρα μεταξύ του διεπιστημονικού υπόβαθρου της θεωρίας του μάρκετινγκ και της τεχνολογίας δεδομένων, ή ειδικότερα της καταναλωτικής συμπεριφοράς και των υπολογιστικών εργαλείων πρόβλεψης, προκειμένου τα αποτελέσματα ανάλυσης να μετατραπούν σε επιχειρησιακά κέρδη. Η έμφαση δίνεται σε ενδείξεις που μπορούν να αξιοποιηθούν επιχειρησιακά, όπως η σχέση δαπανών και ανταπόκρισης ανά κατηγορία προϊόντος, προφίλ καταναλωτή, η επίδραση καναλιών αγορών κ.ά. (Ogeawuchi κ.ά., 2022). Εν προκειμένω, ναι μεν τα δεδομένα είναι αυτά που βρίσκονται στο επίκεντρο, αυτά δε, αφορούν τον καταναλωτή και αποσκοπούν στην εργαλειοποίησή τους για να επιτύχουν μελλοντικές πωλήσεις.

Ο πελάτης είναι αυτός που κρίνει και η κρίση του ξεκινάει με τα χαρακτηριστικά του πυρήνα του καταστήματος, όπως οι τιμές, το προσωπικό, η ποικιλία προϊόντος, η ευκολία πρόσβασης και προεκτάσεις αυτών των παραμέτρων (Stylianou & Pantelidou, 2025). Οι συνδυασμοί αυτών των γνωρισμάτων συγκροτούν την εικόνα του καταστήματος και επηρεάζουν άμεσα την ικανοποίηση και την πρόθεση επιστροφής του πελάτη (*reselling)*. Εμπειρικές μελέτες σε σούπερ μάρκετ αναδεικνύουν σταθερά τον ρόλο της ποικιλίας, της διάταξης και της τιμολόγησης στην ικανοποίηση του πελάτη (Ajiga κ.ά., 2024 ; Ma, Zhang & Zheng, 2025).

Έπειτα, κομβικό ρόλο παίζει ο τρόπος που επιτυγχάνεται η αγορά. Η παρουσία σε πληθώρα καναλιών προσαρμόζει τις καταναλωτικές συνήθειες, καθώς πολλοί πελάτες συνδυάζουν επίσκεψη στο κατάστημα με παραγγελία από εφαρμογή. Σε πρόσφατη έρευνα των Sharma κ.ά. (2025), ανεδείχθη ότι το μερίδιο των καταναλωτών που χρησιμοποιούν online επιλογές ξεπερνά το 50%, ενώ οι προτιμήσεις διαφέρουν ανά ηλικία και νοικοκυριό. Ως αποτέλεσμα, το αποτύπωμα της υβριδικής συμπεριφοράς φτάνει να επηρεάζει, αν όχι να ορίζει, ακόμη και την οργάνωση καταστημάτων και των αποθεμάτων τους.

Ακόμη, ο υψηλός ρυθμός τεχνολογικών αλλαγών και ο έντονος ανταγωνισμός στο λιανικό εμπόριο, αναγκάζει τις επιχειρήσεις να προσαρμοστούν ταχύτατα σε ένα σύστημα καταναλωτισμού, στο οποίο η εμπειρία του πελάτη έχει ισάξια, ή και μεγαλύτερη, σημασία από το προϊόν καθαυτό. Η ενσωμάτωση των ψηφιακών τεχνολογιών στις

καθημερινές αγορές έχει δημιουργήσει ένα νέο μοντέλο καταναλωτή, περισσότερο ενημερωμένο, απαιτητικό και επιρρεπή στην προσωποποιημένη εξυπηρέτηση, την οποία πλέον εξοικειώθηκε και θεωρεί δεδομένη (Yokoyama, Azuma & Kim, 2022). Η άνοδος του ηλεκτρονικού εμπορίου, έχει ωθήσει αμιγώς λιανικής εμπορίας επιχειρήσεις, να επανεξετάσουν το επιχειρηματικό τους προφίλ και να προβούν σε στρατηγικές ολοκλήρωσης, όπως το *omnichannel*, όπου το ψηφιακό και φυσικό κανάλι αλληλοσυμπληρώνονται (Hagberg, Sundström & Egels-Zandén, 2016).

Ειδικά τα σούπερ μάρκετ, ως η παραδοσιακή μορφή λιανικού εμπορίου, αποτελούν αυτοτελή εργαστήρια ανάλυσης συμπεριφοράς, συγκεντρώνοντας τεράστιο όγκο δεδομένων από αγορές ρουτίνας, εποχικές προτιμήσεις, αντιδράσεις σε εκπτώσεις και συμπεριφορές και γενικότερα οποιοδήποτε στοιχείο μπορεί να μετρηθεί από την εμπειρία του πελάτη (Ajiga κ.ά., 2024). Το αποτέλεσμα είναι η επιτυχία της επιχείρησης να μην εξαρτάται αποκλειστικά από το εύρος των προϊόντων και των τιμών τους, ως είθισται, αλλά από την ικανότητα της να αξιοποιεί τα δεδομένα που συλλέχθηκαν. Αυτά χρησιμοποιούνται για να προβλεφθούν ανάγκες, να κατανοηθούν κίνητρα και να ερμηνευθούν συσχετίσεις, δημιουργώντας αξία μέσα από την εμπειρία και ακολούθως αυτή να εξατομικευθεί το περισσότερο δυνατό, εδραιώνοντας την γνώση των πελατών της ως ανταγωνιστικό της πλεονέκτημα.

Αν μη τι άλλο, η συμπεριφορά του πελάτη δεν είναι τυχαία και διαμορφώνεται από ένα πλέγμα παραγόντων βασισμένο σε κοινωνικά, ψυχολογικά, οικονομικά και πολιτισμικά ερεθίσματα, επιλέγοντας προϊόντα όχι μόνο βάσει της χρηστικής τους αξίας, αλλά των συναισθημάτων, των στάσεων και των εμπειριών που τα έχει συνδέσει (Min, 2006). Η θεωρία της αγοραστικής συμπεριφοράς λογίζει ως καθοριστικές παραμέτρους τη σημασία της αντίληψης αξίας, της εμπιστοσύνης στο εμπορικό σήμα και της ικανοποίησης από την εμπειρία αγορών, αναγάγοντας έτσι τον πελάτη ως ενεργό συμμετέχων στη διαδικασία αγοράς και όχι ως παθητικό αποδέκτη προσφορών (Esfandiari κ.ά., 2025). Η απόφαση πλέον είναι δυναμική και διαμορφώνεται από την ευρύτερη πληροφόρησή του, όπως κριτικές άλλων καταναλωτών, σύγκριση επιλογών, ψηφιακή παρουσία κ.ά., αυξάνοντας έτσι τις προσδοκίες, αλλά μειώνοντας την ανοχή στα λάθη. Αυτό έχει ως αποτέλεσμα οι έμποροι λιανικής πώλησης, να παρακολουθούν συνεχώς τις καταναλωτικές συμπεριφορικές τάσεις, αναγνωρίζοντας έγκαιρα τις μεταβολές, ώστε να προσαρμόσουν την τιμολογιακή τους πολιτική και τις αποφάσεις προωθητικών ενεργειών.

Ως φίλτρο στη λήψη αποφάσεων λειτουργούν οι δημογραφικοί παράγοντες. Ηλικία, μέγεθος νοικοκυριού και εισόδημα, μεταξύ άλλων, συνδέονται με τη συχνότητα αγορών,

την εκμετάλλευση προσφορών και την επιλογή καναλιού (Yokoyama, Azuma & Kim, 2022). Πιο συγκεκριμένα, η ανάλυση συμπεριφοράς υποστηρίζεται από αναπαραστάσεις που συνοψίζουν την αξία πελάτη, όπως το *RFM*, δηλαδή πόσο πρόσφατα έγινε η αγορά (Recency), με ποια συχνότητα (Frequency) και με τι συνολική αξία (Monetary Value), καθώς και επεκτάσεις της (de Mooij & Hofstede, 2002). Μοντέλα τύπου *LRFMP* ή *RFM-V* αποδίδουν πιο εξελιγμένα μοτίβα, προσθέτοντας την παράμετρο της διάρκειας (Length), της περιοδικότητας (Periodicity) ή της ποικιλίας (Variety), ώστε να επιτυγχάνουν στοχευμένη τμηματοποίηση σε είδη λιανικού εμπορίου (de Mooij & Hofstede, 2002).

Η χρήση τους στην ανάλυση συμπεριφοράς βασίζεται στα δεδομένα συναλλαγής, τροφοδοτώντας ωστόσο την αλγοριθμική τους επεξεργασία, υπό το πρίσμα των κανόνων του μάρκετινγκ. Η πρόσφατη αγορά συνδέεται με πιθανότητα επιστροφής, ενώ η συχνότητα με την πίστη στο κατάστημα και η αξία με το δυνητικό κέρδος (Min, 2006). Επίσης, αν το σύστημα λιανικής έχει πολλά κανάλια προστίθενται τα χαρακτηριστικά τους (φυσικό, εφαρμογή κτλ.) και οι ρυθμοί επίσκεψης, ώστε να καταγράφεται η πραγματική συμβολή της κάθε αλληλεπίδρασης (Ma, Zhang & Zheng, 2025).

Από προωθητική σκοπιά, η πρόβλεψη ανταπόκρισης σε προωθητικές εκστρατείες με μηχανική μάθηση, βοηθά να επιλεγεί το κοινό στόχος και η ένταση προσφοράς. Ερευνητικά έργα που μελετήθηκαν κατά την βιβλιογραφική ανασκόπηση, τεκμηριώνουν βελτίωση στόχευσης και ανάδειξη κρίσιμων παραγόντων μέσω δέντρων απόφασης και συναφών τεχνικών πρόβλεψης, τα οποία χρησιμοποιεί και το μοντέλο που αναπτύχθηκε για τις ανάγκες της εργασίας (Prasetiyowati, Maulidevi & Surendro, 2021 ; Biau, 2012).

Εξ' ορισμού της, η απόφαση αγοράς λιανικής είναι συνήθως γρήγορη και επαναλαμβανόμενη. Κατευθύνεται κυρίως από την προσλαμβανόμενη αξία, τη συνήθεια και χαρακτηριστικά όπως η τιμή, η διαθεσιμότητα και τυχών προωθητικές ενέργειες (Yokoyama, Azuma & Kim, 2022). Οι προωθητικές αυτές ενέργειες ωστόσο, μετακινούν βραχυπρόθεσμα τη ζήτηση και πολλές φορές επαναλαμβάνονται τακτικά, με αποτέλεσμα να αναμορφώνουν τις προσδοκίες του αγοραστή για την κανονική τιμή (Allaway κ.ά., 2011).

Η ελαστικότητα της τιμής διαφέρει μεταξύ κατηγοριών πελατών. Τα νοικοκυριά με περιορισμένο χρόνο δίνουν βάρος στην ευκολία, ενώ τα νοικοκυριά με περιορισμένο εισόδημα ανταποκρίνονται εντονότερα σε εκπτώσεις (de Mooij & Hofstede, 2002). Διακρίνεται επίσης υποκατάσταση μεταξύ κατηγοριών (π.χ. νωπά με κατεψυγμένα) και συμπληρωματικότητα (π.χ. κρασί με τυρί) (de Mooij & Hofstede, 2002). Η κατανόηση αυτών των σχέσεων είναι κρίσιμη για την χωροθετική λήψη αποφάσεων, όπως κατανομή

χώρου ραφιού και για την τιμολόγηση, ή συνδυαστική προώθηση, κατά κατηγορία προϊόντος.

Ό,τι αφορά τη δραστηριοποίηση του μάρκετινγκ, το ζητούμενο είναι η πιθανότητα ανταπόκρισης. Στα δεδομένα καμπανιών, η θετική ανταπόκριση είναι συνήθως σε λίγες περιπτώσεις, άρα προκύπτει ανισορροπία κλάσεων (Ajiga κ.ά., 2024). Αυτό απαιτεί προσοχή στην επιλογή μετρικών και στη ρύθμιση ορίων απόφασης, γιατί το απλό ποσοστό ορθής ταξινόμησης θολώνει την εικόνα. Η εύστοχη ερμηνεία των πιθανοτήτων και η βαθμονόμησή τους ανάγονται σε ζωτικής σημασίας, όταν οι αποφάσεις αφορούν επιτυχία στόχευση και προϋπολογισμό. Σε αυτό ωστόσο, δεν βοηθάει ότι η καταναλωτική συμπεριφορά δεν είναι στατική. Το γεγονός ότι επηρεάζεται από εποχικότητα, οικονομικές συνθήκες και μάθηση του ίδιου του πελάτη από προηγούμενες εμπειρίες, απαιτεί ένα σχήμα ανάλυσης που ενημερώνεται τακτικά και ελέγχει συστηματικά για μεταβολές (Mani & Shenoy, 2025). Αυτό επιτρέπει να επικαιροποιούνται τα δεδομένα του μοντέλου, ώστε να διατηρείται η ακρίβεια πρόβλεψης και να προσαρμόζονται άμεσα οι τακτικές μάρκετινγκ, που στοχεύουν στην υψηλότερη απόδοση επένδυσης.

Όπως είναι λογικό, οι πελάτες δεν ανταποκρίνονται με τον ίδιο τρόπο στις ίδιες ενέργειες. Διαφέρουν σε παραμέτρους όπως ευαισθησία τιμής, προτίμηση καναλιού, προθυμία δοκιμής νέων προϊόντων και χρόνο που διαθέτουν για αγορές (Esfandiari κ.ά., 2025). Η ίδια προώθηση μπορεί να ενεργοποιήσει έναν πελάτη ευκαιρίας και να αφήσει αδιάφορο έναν πελάτη άνεσης. Η ετερογένεια αυτή θα πρέπει να λαμβάνεται υπόψη κατά τη στόχευση και τη μέτρηση αποτελεσμάτων.

Ο κύκλος ζωής ενός πελάτη περιλαμβάνει φάσεις ένταξης, ωρίμανσης, κάμψης, με διαφορετική συχνότητα αγορών και διαφορετικό καλάθι (Min, 2006). Τα ορόσημα ζωής (γέννηση παιδιού, μετακόμιση κτλ.) μεταβάλλουν το μίγμα καναλιών και κατηγοριών, ενώ τα μοντέλα που ενσωματώνουν χρονικές μεταβολές και εποχικότητα, αποδίδουν πιο σταθερές προβλέψεις σε περιβάλλον λιανικής πώλησης (Stylianou & Pantelidou, 2025). Η πιστότητα ως έννοια δεν είναι ενιαία και οριζόντιας προσέγγισης. Άλλοι πελάτες είναι πιστοί στο κατάστημα, άλλοι στην κατηγορία και άλλοι στο εμπορικό σήμα. Οι ανταμοιβές πιστότητας και οι εξατομικευμένες προσφορές λειτουργούν όταν ταυτίζονται με το κίνητρο του αγοραστή. Η υπερβολική χρήση εκπτώσεων π.χ., μπορεί να συνηθίσει τον πελάτη σε χαμηλότερες τιμές και να μειώσει το περιθώριο κέρδους του προϊόντος (Allaway κ.ά., 2011).

Η ερμηνεία της συμπεριφοράς αυτής όμως, επηρεάζεται άμεσα από την ποιότητα των δεδομένων. Ασυνέπειες σε κωδικούς, απώλειες συναλλαγών ή ασαφείς ορισμοί καναλιών

οδηγούν κατά κανόνα σε λάθος συμπεράσματα. Η καθαρή οριοθέτηση μεταβλητών, η παρακολούθηση και ερμηνεία αλλαγών στον κατάλογο προϊόντων και η αντιστοίχιση πελάτη με κατηγορία νοικοκυριού, είναι απαραίτητα βήματα πριν από κάθε συμπέρασμα (Ma, Zhang & Zheng, 2025). Υπό το ίδιο πρίσμα, τα ζητήματα ιδιωτικότητας και δικαιοσύνης δεν είναι τυπικές λεπτομέρειες. Η χρήση δεδομένων πρέπει να σέβεται τη συγκατάθεση και να αποφεύγει πρακτικές που τιμωρούν ομάδες πελατών. Σύμφωνα με τους Yokoyama, Azuma και Kim, (2022), ο έλεγχος για μεροληψία και η διαφάνεια στον τρόπο λήψης αποφάσεων, αντανακλούν στην αγορά ενισχύοντας την εμπιστοσύνη και την αποδοχή των δράσεων μάρκετινγκ.

2.2 Συμπεριφορά σε σούπερ μάρκετ

Η εικόνα ενός καταστήματος συνοψίζει το πώς ο πελάτης αντιλαμβάνεται την αξία που λαμβάνει. Κρίσιμα στοιχεία που συμβάλλουν σε αυτό είναι η τοποθεσία, η προσβασιμότητα και ο χρόνος εξυπηρέτησης (ουρές, ταμείο, ράφια), χαρακτηριστικά που επηρεάζουν άμεσα τη συχνότητα επίσκεψης και το μέγεθος καλαθιού στα σούπερ μάρκετ (Silverio κ.ά., 2025). Η ποικιλία που παρέχεται, τόσο στο εύρος όσο και στο βάθος επιλογών, αλλά και η διαθεσιμότητα καθορίζουν την πιθανότητα εύρεσης του κατάλληλου προϊόντος σε μία επίσκεψη. Η σαφήνεια διάταξης και σήμανσης μειώνει το κόστος αναζήτησης και ενισχύει την αίσθηση ευκολίας του πελάτη, ενώ η καθαριότητα, ο φωτισμός και η θερμοκρασία λειτουργούν ως ενδείξεις ποιότητας, ιδίως στα ευπαθή προϊόντα, επηρεάζοντας άμεσα την αίσθηση φρεσκάδας και ασφάλειας (Silverio κ.ά., 2025).

Η φρεσκάδα και η επάρκεια αποθέματος είναι άμεσα συνδεδεμένες με την εμπιστοσύνη και επανάληψη αγοράς (Esfandiari κ.ά., 2025). Η τιμή και ο τρόπος παρουσίασης (π.χ. σύγκριση σε αναγωγή μονάδας) διαμορφώνουν αντίληψη δικαιοσύνης και εξοικονόμησης (de Mooij & Hofstede, 2002). Διαχρονικά, οι προωθήσεις ερμηνεύονται σε ψυχολογική βάση μέσω της κατάλληλης πλαισίωσης, αποτελώντας εργαλεία στήριξης της άμεσης επιλογής (Silverio κ.ά., 2025). Η εξυπηρέτηση από την άλλη, η οποία κρίνεται από χαρακτηριστικά όπως διαθεσιμότητα προσωπικού, ευγένεια, διάθεση επίλυσης προβλημάτων, ευθυμία, επαγγελματισμό κ.ά., επηρεάζει την ικανοποίηση και την πρόθεση επιστροφής (de Mooij & Hofstede, 2002). Σε περιβάλλον πολλών καναλιών, το φυσικό κατάστημα παραμένει το μοναδικό σημείο ολοκληρωμένης εμπειρίας, γι’ αυτό η συνέπεια μεταξύ ραφιού, φυλλαδίου και εφαρμογής είναι κομβική (Padmanabhan κ.ά.,

2025). Προσοχή χρειάζεται η αναντιστοιχία τιμής ή διαθεσιμότητας μεταξύ των καναλιών, καθώς διαβρώνει γρήγορα και εύκολα την πίστη του καταναλωτή (Padmanabhan κ.ά., 2025).

Οι προαναφερθείσες διαστάσεις μετρώνται συνήθως με κλίμακες πολλών χαρακτηριστικών και συγκροτούν λανθάνοντα μεγέθη, όπως η άνεση και η αξιοπιστία, τα οποία αποτελούν ποιοτικές μετρικές και δεν ποσοτικοποιούνται εύκολα. Τα μεγέθη αυτά προβλέπουν ικανοποίηση, πρόθεση επαναγοράς, προθυμία σύστασης κ.ά. Στα σούπερ μάρκετ και όχι μόνο, η σύνδεση τους με δείκτες καλαθιού και συχνότητας, τροφοδοτεί την πρακτική αξιοποίηση σε λειτουργίες αλγοριθμικής επεξεργασίας και εφαρμογές μάρκετινγκ. (Mani & Shenoy, 2025).

Σε ό,τι αφορά την τιμή, αυτή λειτουργεί ως άμεση ένδειξη αξίας και ως μονάδα σύγκρισης με την αναμενόμενη από τον πελάτη τιμή, λειτουργώντας ως έννοια παράλληλη με την προσλαμβανόμενη αξία (Min, 2006). Η απόκλιση από την αναμενόμενη τιμή επηρεάζει την πιθανότητα αγοράς και κατ' επέκταση το μέγεθος καλαθιού (Theodoridis & Chatzipanagiotou, 2009). Για ουσιαστική αποτίμηση, εκτιμώνται ίδιες και διασταυρούμενες με άλλα προϊόντα ελαστικότητες τιμής βάσει δεδομένων πωλήσεων, με ελέγχους για εποχικότητα και προσφορά. Αυτό συμβαίνει διότι η ευαισθησία στην τιμή διαφέρει ανά κατηγορία και νοικοκυριό, με εντονότερη διαφοροποίηση σε βασικά και συχνά αγοραζόμενα είδη.

Οι προωθήσεις, ούσες άμεσα συνυφασμένες με την τιμή, κινούν βραχυπρόθεσμα τη ζήτηση, αλλά έχουν και μακροπρόθεσμες επιδράσεις (*carryover effect)*. Σε περίπτωση θετικού *carryover effect* διατηρείται το νέο πελατολόγιο. Σε περίπτωση αρνητικού όμως, παρατηρούνται φαινόμενα αποθήκευσης (*stockpiling*), χρονικής μετακύλισης της αγοράς και στις περιπτώσεις με επαναλαμβανόμενα σχήματα έκπτωσης, ατονία της αγοράς λόγω αναμονής της έκπτωσης (*promotion fatigue*) και παγίωση του αναμενόμενου κόστους στη χαμηλότερη τιμή (Stylianou & Pantelidou, 2025). Ωστόσο, η αξιολόγηση γίνεται μέσα από αύξηση καθαρών πωλήσεων και οριακού κέρδους και όχι μόνο από τον όγκο πωλήσεων. Ως εκ τούτου, έχει στρατηγική σημασία αν το περιθώριο κέρδους προήλθε μόνο από προσωρινό όγκο πωλήσεων με χαμηλότερο συντελεστή κερδοφορίας (*markup*), ή αν δημιουργήθηκε σταθερότερη βάση πωλήσεων μετά την προωθητική ενέργεια. Αυτό εξετάζεται με το πέρας της προσφοράς, διακρίνοντας αν κρατήθηκε μέρος της ζήτησης, ή αν αντικαταστάθηκαν με πωλήσεις από το μέλλον.

Τέτοιες μέθοδοι αξιολόγησης προσφοράς μπορεί να είναι η εξέταση σε ομάδες ελέγχου, όπου συγκρίνονται αλλαγές σε παρόμοιους πελάτες, ή σε καταστήματα που δεν πήραν

την προσφορά. Εκτελούνται ως πειραματικά σχέδια, τα οποία δίνουν τυχαία την προσφορά στη μία ομάδα, με επιλεγμένα ισοδύναμα με την άλλη χαρακτηριστικά, και μόλις επιτευχθεί στατιστική ισχύς αλλάζει η ομάδα στόχος. Έπειτα, μετρώντας διαφορές των διαφορών (*DiD - Difference in Difference*) πριν και μετά μεταξύ ομάδας δράσης και ελέγχου, απομονώνονται οι κοινές τάσεις τις αγοράς. Κατά την διαδικασία του πειράματος θα πρέπει να αποφεύγεται η έκθεση του μηνύματος προσφοράς της μίας ομάδας στην άλλη, γι' αυτό και σε online προώθηση ή φυλλαδίου θα πρέπει να αλλάζει, ή να περιορίζεται χωρικά (Mani & Shenoy, 2025).

Οι κατηγορίες των προϊόντων, όπως αναφέρθηκε στο προηγούμενο υποκεφάλαιο, συνδέονται μεταξύ τους είτε με υποκατάσταση, είτε με διάχυση μεταξύ κατηγοριών λόγω συμπληρωματικών προϊόντων (*halo*). Η σωστή μέτρηση αυτών των αντιβαλλόμενων ελαστικοτήτων περιορίζει τον «κανιβαλισμό» (δηλαδή μία τιμή ενός προϊόντος να επηρεάσει αρνητικά τη ζήτηση άλλου προϊόντος υπό την ίδια εταιρική ομπρέλα), αφού ελέγχεται και βελτιώνεται η κατανομή προϋπολογισμού σε καμπάνιες ανά κατηγορία (Stylianou & Pantelidou, 2025). Ο έλεγχος αυτός βέβαια, συνοδεύεται από αναλύσεις διαχωρισμένες σε τμήματα βάσει ευαισθησίας τιμής και άνεσης για την εκάστοτε υποομάδα αγοράς.

Η τελική αξιολόγηση ωστόσο γίνεται πολυπαραγοντικά βάσει δεικτών (*KPIs – Key Performance Indicators*) και παραμέτρων, όπως καθαρή αύξηση μετρικών λόγω της προωθητικής ενέργειας (*uplift*), οριακού περιθωρίου κέρδους, ποσοστό εξαργύρωσης κουπονιών (*redeem rate*), *ROI* (Return on Investment) και *DiD* της εκστρατείας, στρωμάτωση (*matching)*, διαρροή προς άλλες κατηγορίες κ.ά. (Ajiga κ.ά., 2024). Έτσι, η απόφαση δεν βασίζεται μόνο σε μία ένδειξη, αλλά συνδυάζεται με περιορισμούς λόγω αποθέματος και λειτουργικού κόστους, ώστε η πολιτική προσφορών να είναι βιώσιμη και σταθερή στον χρόνο, εξυπηρετώντας την στρατηγική της επιχείρησης.

Στην αγορά των πελατών στα σούπερ μάρκετ τώρα, οι επιλογές καναλιού δεν είναι τυχαίες και ακολουθούν σταθερά μοτίβα. Ο ίδιος πελάτης άλλοτε αγοράζει στο κατάστημα, άλλοτε μέσω εφαρμογής και άλλοτε με παραλαβή από το κατάστημα. Αυτό εξαρτάται από τον διαθέσιμο χρόνο, την ανάγκη για αμεσότητα στην παράδοση και την αναμενόμενη αξία καλαθιού του (Padmanabhan κ.ά., 2025). Η καταγραφή του μεριδίου καναλιού ανά πελάτη (π.χ. 50% κατάστημα, 40% εφαρμογή, 10% click & collect), αποτυπώνει τις προτιμήσεις εξυπηρέτησης και ορίζει την στόχευση προωθήσεων. Σταθερά μοτίβα καναλιού επιδεικνύουν τις διαφορετικές προτιμήσεις εξυπηρέτησης και αντανακλούν την ανεκτικότητα σε χρόνο αναμονής.

Η αποστολή αγοράς (*shopping mission*) περιγράφει τον λόγο της επίσκεψης και την ταξινομεί βάσει του σκοπού και του μεγέθους καλαθιού. Διακρίνεται κυρίως σε (Yokoyama, Azuma & Kim, 2022):

- γρήγορη συμπλήρωση - λίγα προϊόντα, άμεσα απαραίτητα, μικρός χρόνος παραμονής
- μεγάλη προμήθεια - πλήρες καλάθι, μακροπρόθεσμη κάλυψη αναγκών, προγραμματισμένες αγορές
- επείγουσα ανάγκη - περιορισμένος κατάλογος επιλογών, αντικατάσταση είδους που μόλις τελείωσε, προτεραιότητα στην ταχύτητα

Κάθε αποστολή έχει τυπικά σημάδια που την ταξινομούν. Δεδομένα όπως η ημέρα και ώρα, το μέσο καλάθι, η σύνθεση κατηγοριών και το προτιμώμενο κανάλι, είναι μερικά από αυτά (Yokoyama, Azuma & Kim, 2022). Η σωστή αναγνώριση αποστολής βοηθά στη λήψη αποφάσεων για πρόταση προϊόντων και στη διαχείριση διανομής αυτών. Από τη σκοπιά της διαχείρισης διανομής όταν κυριαρχεί η ταχύτητα ως κριτήριο, στόχος είναι ο μικρότερος χρόνος ανακύκλωσης, ενώ όταν προβλέπεται μεγάλο καλάθι η βέλτιστη αξιοποίηση των πόρων (Silverio κ.ά., 2025). Γενικότερα, βάσει και του *RFM,* η υψηλή ένδειξη σε πρόσφατη δράση είναι συνδεδεμένη με απόκριση σε κουπόνια, η υψηλή αξία αγορών με σταθερότητα επισκεψιμότητας και η μεγάλη συχνότητα με την σημαντικότητα του πελάτη για το κατάστημα (Yokoyama, Azuma & Kim, 2022).

Υπό το δημογραφικό πρίσμα, το εισόδημα καθορίζει τη στρατηγική τιμής ως προς την ποιότητα και την προτίμηση σε μάρκες. Τα μεσαία και χαμηλά στρώματα εμφανίζουν υψηλότερη ελαστικότητα ζήτησης και ταχύτερη μετατόπιση μεταξύ καταστημάτων, ενώ τα υψηλότερα εισοδήματα δίνουν έμφαση στην ευκολία, την ποιότητα και τις *premium* υποκατηγορίες (Allaway κ.ά., 2011). Η εκπαίδευση και το επάγγελμα επίσης, συνδέονται με γνώση προϊόντος, αναγνωρισιμότητα σημάτων και ανάγνωση διατροφικών ετικετών (Stylianou & Pantelidou, 2025). Ο επαγγελματικός φόρτος από την άλλη, επιδρά στη συχνότητα αγορών και στην επιλογή καναλιού, η οποία επηρεάζεται εξίσου και από τη χωρική διάσταση (αστική κατοικία, προσβασιμότητα κτλ.) (Hagberg, Sundström & Egels-Zandén, 2016).

Η εκτίμηση της αξίας από τον κάθε πελάτη επηρεάζεται από τις στάσεις, τις αντιλήψεις και τα κίνητρα που τον χαρακτηρίζουν, ενώ η προδιάθεση για να μετατραπεί τελικά η επαφή σε αγορά διαμορφώνεται σύμφωνα με την αντιλαμβανόμενη ποιότητα και την αναγνωρισιμότητα της μάρκας (Min, 2006). Όσο μεγαλύτερη η εμπλοκή του με την

κατηγορία, τόσο μεγαλύτερο το βάθος αναζήτησης και σύγκρισης. Στις κατηγορίες ευαισθησίας (βιολογικά, υγιεινά κ.ά.) οι κοινωνικές ομάδες αναφοράς καθορίζουν τις προτιμήσεις, καθώς η κοινωνική επιρροή λειτουργεί μέσω κανόνων, μίμησης και κοινωνικής απόδειξης, όπως και ένταξης (Esfandiari κ.ά., 2025). Ως εκ τούτου, οι κριτικές και οι βαθμολογίες ενισχύουν την προσλαμβανόμενη αξιοπιστία και μειώνουν την αβεβαιότητα. Αντίκτυπο στην ψυχολογική διάσταση παρουσιάζει επίσης η ατμόσφαιρα του καταστήματος, η μουσική, το άρωμα, η διακόσμηση και άλλες υποσυνείδητες παράμετροι, ικανές να επηρεάσουν τον ρυθμό και το μέγεθος κατανάλωσης (Silverio κ.ά., 2025). Εν αντιθέσει, η κόπωση στην απόφαση και ο μεγάλος φόρτος επιλογών κάνουν τις προεπιλεγμένες λύσεις να ευδοκιμούν (Silverio κ.ά., 2025).

Γενικότερα, τα προφίλ πελατών σχηματίζονται από συνήθειες και περιορισμούς. Ως συνήθεια θεωρείται μία σταθερή μέρα ή ώρα επίσκεψης σε μακροπρόθεσμη βάση, ή η προτίμηση σε μία συγκεκριμένη μάρκα (Stylianou & Pantelidou, 2025). Περιορισμός μπορεί να είναι ο μικρός διαθέσιμος χρόνος αγορών, προσανατολίζοντας τον πελάτη σε παραγγελία με λίγα κλικ και σαφή χρόνο παράδοσης, ή το χαμηλό εισόδημα, το οποίο ταυτίζεται με ευαισθησία και ανταπόκριση σε προσωποποιημένες εκπτώσεις (Stylianou & Pantelidou, 2025). Το μέγεθος νοικοκυριού και η παρουσία παιδιών είναι δημογραφικοί παράγοντες που μεταβάλλουν τη συχνότητα αγοράς, το μέσο καλάθι και την αξιοποίηση προσφορών, ενώ τα μονοπρόσωπα νοικοκυριά συνηθίζουν να επιλέγουν μικρές συσκευασίες και παρουσιάζουν μεγαλύτερη ποικιλία σε δοκιμές (Sharma κ.ά., 2025). Επίσης, ηλικιακά στάδια και κύκλος ζωής σχετίζονται άμεσα με διαφορετικές ανάγκες, όπως κατανάλωση έτοιμων γευμάτων στους νεότερους, συγκεκριμένων ειδών υγιεινής στους ηλικιωμένους και βασικών προϊόντων στα μεγαλύτερα νοικοκυριά (Sharma κ.ά., 2025).

Για την ανάλυση, είναι κομβικό να ορίζονται μετρήσιμα μεγέθη επανάληψης και μετατόπισης ανά κανάλι και αποστολή αγοράς για κάθε πελάτη Κάποια εξ' αυτών είναι μερίδιο καναλιού, μέσο καλάθι ανά αποστολή, ρυθμός επίσκεψης ανά ημέρα ή ζώνη ώρας, δείκτες ανταπόκρισης σε προωθήσεις, *RFM*, δείκτες αξίας κύκλου ζωής των χρηστών (*CLV – Customer Lifetime Value*), ποσοστό απώλειας πελατών (*churn rate*) κ.ά.. Οι δείκτες αυτοί επιτρέπουν ουσιαστική τμηματοποίηση των πελατών με πραγματικά συμπεριφορικά μοτίβα, χωρίς να επιβάλλονται προκατασκευασμένα και αυθαίρετα στερεότυπα. Επιτυγχάνουν επίσης ιεράρχηση προτεραιοτήτων επαφών και βελτιστοποίηση του κόστους απόκτησής τους, αφού πρώτα έχουν ενισχύσει την ακρίβεια πρόβλεψης και τη σχετικότητα των προτάσεων.

Συνεπαγωγικά, οι δημογραφικοί παράγοντες λειτουργούν αθροιστικά και αλληλεπιδρούν ευκρινώς με την πολιτική τιμών, τη διανομή και την τοποθέτηση προϊόντων. Για γρήγορη συμπλήρωση ταιριάζουν σύντομες προσφορές σε βασικά είδη και εύκολες λίστες, ενώ για μεγάλη προμήθεια ταιριάζουν πακέτα προσφορών και οικονομίες κλίμακας. Στις παραδόσεις η αξιοπιστία χρόνου είναι καθοριστική για την ικανοποίηση και διατήρηση πελατών που επιλέγουν την άνεση (Mani & Shenoy, 2025). Συνεπώς, κατόπιν σύζευξης συμπεριφορικών και δημογραφικών δεδομένων, η στόχευση μάρκετινγκ προσαρμόζεται στο προφίλ, όπως και στην αποστολή αγοράς, με σαφείς κανόνες εξυπηρέτησης και αποδοτική χρήση πόρων. Εντός αυτού του πλαισίου, τα δεδομένα συναλλαγών ορίζουν τις διαφοροποιήσεις μεταξύ κατηγοριών και με τις εκτιμήσεις πρόβλεψης που αναλύονται στο επόμενο κεφάλαιο, μετατρέπονται σε πρακτικές αποφάσεις.

2.3 Προβλεπτική μοντελοποίηση στο λιανικό εμπόριο

2.3.1 Εισαγωγή στην προβλεπτική ανάλυση στο λιανικό εμπόριο

Εάν έπρεπε να δοθεί ένας ορισμός για την προβλεπτική μοντελοποίηση στο λιανικό εμπόριο, αυτός θα την περιέγραφε ως η χρήση ιστορικών δεδομένων για εκτίμηση μελλοντικών συμπεριφορών και αποτελεσμάτων, όπως ανταπόκριση σε προσφορά, επιλογή καναλιού ή πρόθεση επαναγοράς. Συνδυάζει τη στατιστική μοντελοποίηση με τη μηχανική μάθηση ώστε να παραχθούν πιθανότητες, ταξινομήσεις και κανόνες που συνδέονται άμεσα με αποφάσεις μάρκετινγκ και λειτουργίας. Στο πλαίσιο των σούπερ μάρκετ, εστιάζει σε πιθανότητες ενεργειών που σχετίζονται με το καλάθι, τη συχνότητα, την ευαισθησία τιμής και την πιστότητα (Ajiga κ.ά., 2024). Η ουσία βρίσκεται στη μετατροπή δεδομένων ρουτίνας σε πρόγνωση, με πρακτικό όφελος στη λήψη αποφάσεων σχετικά με τη στόχευση, την τιμολόγηση, τα αποθέματα κ.ά.

Κρίσιμο προαπαιτούμενο είναι η ενοποίηση πηγών. Αυτό σημαίνει δεδομένα όπως συναλλαγές POS, στοιχεία πιστότητας, ψηφιακά ίχνη και δεδομένα αποθεμάτων να χαρτογραφούνται σε κοινό σχήμα ανά πελάτη, χρόνο και κατηγορία. Η προτυποποίηση επιτρέπει συνεπή ορισμό δεικτών και επαναληψιμότητα πειραμάτων (Mani & Shenoy, 2025). Τυχών ελλείψεις ή ασυμβατότητες στο σχήμα περιορίζουν την ακρίβεια κάθε επόμενου βήματος.

Η διαδικασία οργανώνεται ως μονοπάτι δεδομένων (*data pipeline*). Προηγούνται καθαρισμός, χειρισμός ελλιπών τιμών, εξομάλυνση ακραίων τιμών και σωστή

κωδικοποίηση κατηγορικών δεδομένων (*encoding*). Έπειτα ακολουθούν κλιμάκωση (*scaling*), κατασκευή χαρακτηριστικών (*feature* engineering) και αντιμετώπιση ανισορροπίας κλάσεων (*class imbalance*), εφόσον κριθεί απαραίτητο, εντός των πτυχών της επικύρωσης του μοντέλου (*model validation*). Η τήρηση της σειράς είναι αδιαπραγμάτευτη, καθώς αποτρέπει διαρροή πληροφορίας (*data* leakage) και αυξάνει τη γενίκευση.

Η *feature engineering* μεταφράζει ακατέργαστες στήλες σε ενδείξεις με προγνωστική αξία. Πιο συγκεκριμένα στήλες όπως *RFM*, εποχικότητα ανά εβδομάδα ή μήνα, ελαστικότητα τιμής με βάση ιστορικές μεταβολές, ένταση και τύπος προωθήσεων, αλληλεπιδράσεις δημογραφικών δεδομένων με κατηγορίες και χρονικές υστερήσεις, ώστε να αποδοθεί αδράνεια ή συνήθεια, είναι μερικά μόνο παραδείγματα. Στο πλαίσιο της εργασίας, τέτοια χαρακτηριστικά τροφοδοτούν τους ταξινομητές (*classifiers*) που εξετάζονται για ανταπόκριση σε ενέργεια και κινδύνους απόρριψης.

Ο καθορισμός στόχου και χρονικού διαστήματος πρόβλεψης (*prediction window)* προηγείται της μοντελοποίησης. Ορίζεται σαφώς τι σημαίνει θετικό γεγονός (π.χ. ανταπόκριση σε προσφορά εντός του *prediction window* 30 ημερών), ποιος είναι ο χρονικός ορίζοντας εκπαίδευσης (*training horizon*) και ποιο το διάστημα αξιολόγησης (*evaluation period)*. Η απόφαση αυτή συνάδει με τη χρήση της πρόβλεψης, καθώς ο σκοπός υπαγορεύει τον ορίζοντα και η διαφοροποίησή του τροποποιεί τις παρεμβάσεις της επιχείρησης βάσει της πρόβλεψης (*action arrows)*, οι οποίες δύνανται να είναι σε οποιοδήποτε κανάλι επιλεγεί (Mani & Shenoy, 2025).

Η εκπαίδευση και το *validation* στηρίζονται σε διασταυρούμενη αξιολόγηση (*CV - cross validation*), δηλαδή διαχωρισμό των δεδομένων σε *k* υποσύνολα (*folds*) για να εκπαιδευτεί το μοντέλο με *k-1*, κρατώντας αυτό το ένα για δοκιμή. Αυτό επαναλαμβάνεται τόσες φορές όσα και τα *folds*, με το κάθε fold να έχει αποτελέσει *k-1* φορές δεδομένα εκπαίδευσης και μία φορά δεδομένα δοκιμής, με αποτέλεσμα να υπολογιστεί η μέση απόδοση του μοντέλου για όλα τα *folds* από όλες τις επαναλήψεις. Για δεδομένα χωρίς ισχυρή χρονική εξάρτηση (χρονικότητα) χρησιμοποιούνται k-fold, ενώ για έντονη ακολουθία χρόνου (δηλαδή κομβική σημασία στη χρονική σειρά των δεδομένων) εφαρμόζονται χρονικοί διαχωρισμοί (*time-based splits*), όπου τα *folds* δεν χωρίζονται τυχαία, αλλά σε χρονικά διαστήματα (Ghosh κ.ά., 2024). Στην παρούσα εργασία τα δεδομένα δεν παρουσιάζουν ισχυρή χρονικότητα που να επιβάλλουν *time-based splits*.

Για κάθε μοντέλο *classifier* θα πρέπει να ρυθμίζονται οι υπερπαράμετροί του. Αυτές είναι ανεξάρτητες από τα δεδομένα και αποτελούν τιμές που ρυθμίζουν τον τρόπο μάθησης

ενός *classifier*. Ο *classifier* επιλέγει κάποιες εγγραφές (*train set*) και εκπαιδεύει το μοντέλο πάνω σε αυτές (70-80% δείγματος) και αφήνει το υπόλοιπο μέρος του δείγματος για τον τελικό έλεγχο. Στην *CV*, η ρύθμιση τους γίνεται εντός του μοντέλου, ενώ μπορεί να παρουσιαστεί και ως *Nested CV*, όταν απαιτείται περισσότερο αυστηρή εκτίμηση, όχι όμως στην περίπτωση της παρούσης. Το τελικό μοντέλο ελέγχεται στο ανεξάρτητο δείγμα που δεν έχει φανεί σε κανένα στάδιο εκπαίδευσης και έχει διατηρηθεί απομονωμένο καθ' όλη τη μοντελοποίηση εξ' αρχής (*test set*).

Η επιλογή των μετρικών αξιολόγησης του μοντέλου ευθυγραμμίζεται με το επιχειρησιακό ερώτημα. Γενικά, τα προβλήματα ανταπόκρισης ανάγονται σε ταξινόμησης, όπως αυτό που εξετάζεται εδώ. Προτιμώνται λόγω χαμηλού θετικού ρυθμού (μικρό *positive rate*) και κόστους επαφής, μετρικές όπως ακρίβεια θετικών ανταποκρίσεων (*precision*), ευαισθησία (*recall*), ισορροπία μεταξύ *precision* και *recall* (*F1*). Συχνά επιστρατεύεται το *confusion matrix*, το οποίο είναι ένας δισδιάστατος πίνακας που συνοψίζει την επιτυχία του *classifier* στο ερώτημα ανταπόκρισης με τιμές επιτυχημένων και αποτυχημένων θετικών και αρνητικών προβλέψεων TP, FP, TN, FN και από το πλήθος αυτών των μεταβλητών προκύπτουν οι τιμές του *precision* = TP / (TP+FP) και του *recall* = TP / (TP+FN) (Szymański & Kajdanowicz, 2019).

Εξετάζεται επίσης ο δείκτης *ROC-AUC* (*Area under Receiver Operating Characteristic Curve*) αναλογία ρυθμού σωστής θετικής πρόβλεψης (*TPR – True Positive Rate*) και λανθασμένης θετικής πρόβλεψης (*FPR – False Positive Rate*), ιδιαίτερα χρήσιμος σε ισορροπημένες κλάσεις και ο δείκτης *PRC–AUC* (*Area under Precision–Recall Curve*) που χρησιμοποιείται για την αξιολόγηση της θετικής απόδοσης κυρίως σε κλάσεις με ανισορροπία (Johnson & Khoshgoftaar, 2019). Η *ROC–AUC* δίνει σφαιρική εικόνα ανεξάρτητη από κατώφλι (*threshold*), αλλά η απόφαση παίρνεται στο συγκεκριμένο κατώφλι που ισορροπεί έσοδα και κόστος. Αναφορικά, στη ζήτηση, ως πρόβλημα παλινδρόμησης, προέχει η πρόγνωση μετρικών σφάλματος, επειδή συνδέεται άμεσα με απώλεια πωλήσεων ή υπεραπόθεμα, καθώς όσο μεγαλύτερο το μέγεθος σφάλματος, τόσο μεγαλύτερο το κόστος (Peng, Lee & Ingersoll, 2002).

Η μετάβαση από το θεωρητικό μοντέλο σε ενέργεια απαιτεί μηχανισμό απόφασης. Συνεπώς, ορίζεται κατώφλι με βάση πίνακα κόστους και οφέλους, όπου αντιστοιχίζουν τα έσοδα ανά επιτυχή επαφή με το κόστος επαφής, επιλέγεται το πλήθος στόχευσης ως μέγεθος παρτίδας και καθορίζονται κανάλια επικοινωνίας ανά στόχο. Η αποτελεσματικότητα τεκμηριώνεται με πειραματισμό ή ισοδύναμα σχήματα (όπως διαφημίσεις βιτρίνας ως γεωγραφικός πειραματισμός, εμπειρικών πειραματισμών τύπου

multi-armed bandits κ.ά.) ώστε το κέρδος να αποδίδεται στην πρόβλεψη και όχι σε εξωγενείς μεταβολές) (Ma, Zhang & Zheng, 2025).

Ωστόσο, δεν αρκεί απλά η δημιουργία ενός μοντέλου, αλλά και να φτάσει στη λειτουργική του ωριμότητα, καθώς είναι αυτή που κλείνει τον κύκλο. Παρακολούθηση απόδοσης και βασικών μετρικών (*precision, recall, F1 ROC-AUC*), έλεγχος μετατόπισης δεδομένων (*data drift*), κανόνες επανεκπαίδευσης για προκαθορισμένες τιμές (*triggers*) και αρχειοθέτηση εκδόσεων, ώστε να υπάρχει η δυνατότητα *rollback*, εξασφαλίζουν τη σταθερότητα της ποιότητας (Goldmann, Machado & Osterrieder, 2025). Η ερμηνευσιμότητα μέσω σημαντικότητας χαρακτηριστικών (*feature importance*) και τοπικών εξηγήσεων σε επίπεδο πρόβλεψης ενισχύει την διαφάνεια, τον έλεγχο μεροληψίας και γενικότερα την εμπιστοσύνη στο μοντέλο, ιδίως όταν οι προβλέψεις στηρίζουν εκπτώσεις ή αποκλεισμούς.

Πιο συγκεκριμένα, η σχέση μεταξύ του ερωτήματος, των μετρικών και της απόφασης, πρέπει να είναι ρητή. Σε ανταπόκριση καμπάνιας, ζητούμενο είναι να εντοπίζεται μικρή θετική τάξη με κόστος επαφής, γι' αυτό προτιμώνται *precision, recall, F1* και *PRC–AUC* (Padmanabhan κ.ά., 2025). Σε πρόβλεψη ζήτησης, προέχει η ελαχιστοποίηση σφάλματος και η απόφαση μεταφράζεται σε παραγγελίες, αναπλήρωση και προγραμματισμό προσφορών. Σε πρόβλεψη επόμενης πρότασης (*next best*), το πρόβλημα είναι κατάταξης. Αποτυπώνονται μετρικές πληροφοριακής ανάκτησης και η απόφαση είναι ποιες προτάσεις θα εμφανίζονται πρώτες στην εφαρμογή ή στο ταμείο.

Το ερώτημα είναι άμεσα συνυφασμένο με τον ορισμό του στόχου, ο οποίος απαιτεί σαφή χρονικό ορίζοντα. Για παράδειγμα, εάν το θετικό γεγονός μετριόταν ως ανταπόκριση σε κουπόνι κρασιού εντός 30 ημερών, το παράθυρο παρατήρησης προηγείται και τροφοδοτεί χαρακτηριστικά, όπως αγορές τελευταίων 90 ημερών, ενώ το διάστημα πρόβλεψης είναι μεταγενέστερο και δεν επικαλύπτεται. Η διάκριση αυτή αποτρέπει διαρροή πληροφορίας. Ομοίως, αν αλλάξει ο ορίζοντας, προσαρμόζεται και ο επιχειρησιακός στόχος (ταχύτερη επαφή, μικρότερη παρτίδα).

Το κατώφλι που θα επιλεγεί, θα πρέπει να αποτιμάται με τη σχέση κόστους ως προς το όφελος. Κάθε σωστή πρόβλεψη έχει αναμενόμενο κέρδος (περιθώριο − κόστος επαφής), ενώ κάθε λάθος έχει κόστος (σπατάλη έκπτωσης ή κορεσμός πελάτη) (Ma, Zhang & Zheng, 2025). Το βέλτιστο κατώφλι είναι εκεί όπου η διαφορά μεταξύ οφέλους και κόστους μεγιστοποιείται. Η βελτιστοποίηση μπορεί να γίνει ανά τμήμα πελατών, ώστε να λαμβάνεται υπόψη διαφορετική ευαισθησία τιμής ή διαφορετικό κόστος καναλιού (Ma, Zhang & Zheng, 2025).

Όσον αφορά την αξιολόγηση, η απόδοση θα πρέπει να τεκμηριώνεται πειραματικά. Τα τεστ μεταξύ ομάδων πελατών ή καταστημάτων απομονώνουν το καθαρό αποτέλεσμα από εποχικότητα και θόρυβο (Ajiga κ.ά., 2024). Εναλλακτικά, χρησιμοποιούνται σχεδιασμοί με ισοδύναμες ομάδες ελέγχου. Η ανάλυση δεν περιορίζεται σε όγκο πωλήσεων, αλλά περιλαμβάνει οριακό περιθώριο, *uplift* και μετατόπιση σε συμπληρωματικές ή ανταγωνιστικές κατηγορίες.

Στον δρόμο για την ολοκλήρωση, η λειτουργική ενσωμάτωση χρίζει εποπτείας και ελέγχου αλλαγών. Ορίζονται κανόνες για επικαιροποίηση δεδομένων, παρακολούθηση μετατόπισης και επανεκπαίδευση, ενώ κάθε έκδοση μοντέλου συνοδεύεται από τεκμηρίωση σκοπού, χαρακτηριστικών και μετρικών (Mani & Shenoy, 2025). Η ροή εγκρίνεται από κοινού από το τμήμα μάρκετινγκ και λειτουργιών, ώστε να διασφαλίζεται η συνέπεια μεταξύ πρόβλεψης και εκτέλεσης.

Ακόμη, η βαθμονόμηση πιθανοτήτων (*calibration*) πρέπει να πάρει την προσοχή που της αρμόζει, καθώς αποτελεί γέφυρα προς τις επιχειρησιακές αποφάσεις. Ένα καλό ταξινομητικό μοντέλο δεν εγγυάται αξιόπιστες πιθανότητες (Ali, Shamsuddin & Ralescu, 2013). Η βαθμονόμηση (isotonic regression ή Platt scaling) εναρμονίζει την προβλεπόμενη πιθανότητα με τη συχνότητα πραγματοποίησης. Όταν το κατώφλι ορίζεται με οικονομικούς όρους, η ορθή βαθμονόμηση μειώνει λάθη στόχευσης και βελτιώνει το πραγματικό περιθώριο. Το θέμα της βαθμονόμησης αναπτύσσεται εκτενέστερα στο κεφάλαιο 2.3.4. Επίσης, να σημειωθεί πως ευαίσθητα θέμα ηθικής φύσεως, ιδιωτικότητας και δικαιοσύνης ενσωματώνονται από το στάδιο του σχεδιασμού. Η ελαχιστοποίηση δεδομένων στα απαραίτητα, η ανωνυμοποίηση και οι έλεγχοι μεροληψίας αποτελούν μέτρα ασφαλείας και προστατεύουν τον πελάτη, μειώνοντας παράλληλα τους νομικούς κινδύνους (Stylianou & Pantelidou, 2025).

2.3.2 Δεδομένα και τεχνικές για προβλεπτική ανάλυση

Η μέτρηση της ποιότητας των προβλέψεων καθορίζεται από το εύρος και τη συνέπεια των πηγών δεδομένων. Βασικές πηγές είναι οι συναλλαγές POS, τα προγράμματα πιστότητας (πόντοι επιβράβευσης κτλ.), τα ψηφιακά ίχνη και η ανακύκλωση αποθεμάτων. Η ενοποίηση τους γίνεται σε ενιαίο σχήμα ανά πελάτη, συναρτήσει του χρόνου και της κατηγορίας, με πρωτεύοντα κλειδιά (*PK – Primary Key*) και σαφείς ορισμούς μονάδων

(π.χ. ημέρα ως βασικό χρονικό βήμα) (Hagberg, Sundström & Egels-Zandén, 2016).

Όπου κρίνεται απαραίτητο, η κατασκευή χαρακτηριστικών αντιστοιχίζει και μετατρέπει ακατέργαστες στήλες σε προγνωστικούς δείκτες. Συνήθεις δείκτες είναι οι *RFM* και *LRFMP*, δείκτες εποχικότητας (ημέρα εβδομάδας, μήνας), ρυθμοί αλλαγής και μεταβλητότητα. Από πλευράς τιμής και προώθησης, χρήσιμες είναι η ένταση και ο τύπος έκπτωσης, καθώς και απλοί δείκτες ελαστικότητας μέσω ιστορικών μεταβολών. Οι συσχετίσεις μεταξύ στηλών δημιουργούν νέα εξαρτώμενα *features* (π.χ. εισόδημα * δείκτης κατηγορίας Χ) και οι χρονικές υστερήσεις (*lags*), αναδιατάσσουν ιστορικά δεδομένα για να φανερώσουν αδράνεια και συνήθεια μεταξύ χρονικών φαινομένων.

Σημαντικοί επίσης είναι και οι δείκτες καναλιών και οι αποστολές αγοράς. Μερίδιο καναλιού ανά πελάτη, μέσο καλάθι και σύνθεση ανά αποστολή, καθώς και ρυθμός επίσκεψης ανά ζώνη ώρας, παρέχουν καθαρή εικόνα συμπεριφορικών τάσεων (Silverio κ.ά., 2025). Οι ανισορροπίες κλάσεων αντιμετωπίζονται μέσα από τεχνικές που περιλαμβάνουν επαναπροσδιορισμό τον βαρών (*reweighting*), βάρη κλάσεων (*class weights*) ή τεχνικές συνθετικών δειγμάτων (*SMOTE*), όπως χρησιμοποιούνται και στο μοντέλο που εφαρμόστηκε στην παρούσα. Ο σκοπός δεν είναι η τεχνητή ισορροπία ανά πάσα στιγμή, καταλήγοντας σε ανεπιθύμητα αποτελέσματα υπερπροσαρμογής (*overfitting*), αλλά η σταθερή απόδοση στην άγνωστη διανομή παραγωγής. Για αυτό επιστρατεύονται τεχνικές ποσοτικά προσανατολισμένου επαναδειγματισμού (*oversampling* ή *undersampling*). Η *CV* αξιοποιείται και σε αυτή την περίπτωση, όχι όμως απαραίτητα επαναλαμβάνοντας όλα το *folds*, αλλά σταματώντας μόλις παρατηρηθεί επαρκή προσαρμογή (*early stopping*). Καταληκτικά, η επιλογή λύσης εξαρτάται από το κόστος σφάλματος και τη σταθερότητα των μετρικών.

Τα προβλήματα που χρίζουν λύσης από την άλλη, διακρίνονται σε ταξινόμησης, παλινδρόμησης και σειρών χρόνου (*time series*). Η ταξινόμηση καλύπτει ανταπόκριση σε καμπάνιες και αποχώρηση, η παλινδρόμηση προασπίζει ζήτηση και μέσο καλάθι, ενώ οι σειρές χρόνου αποτυπώνουν ροές πελατών και ανάγκες αναπλήρωσης (Yokoyama, Azuma & Kim, 2022). Η επιλογή λύσης τώρα, ορίζεται από το επιχειρησιακό ερώτημα και τον χρονικό ορίζοντα δράσης.

Οι αλγόριθμοι που χρησιμοποιεί η εργασία επαρκούν για τις κύριες χρήσεις στη λιανική. Αυτοί, όπως θα τους δούμε αναλυτικότερα στο επόμενο κεφάλαιο, είναι επιγραμματικά οι εξής:

- Λογιστική παλινδρόμηση (LR - Logistic Regression): Υπολογίζει ένα γραμμικό συνδυασμό των χαρακτηριστικών και τον περνά από σιγμοειδή συνάρτηση για να

δώσει πιθανότητες. Γραμμική, ερμηνεύσιμη, καλή ως ενδεικτική και βασική εκτίμηση.

- Δέντρα αποφάσεων (DT – Decision Trees): Χωρίζει επαναληπτικά τον χώρο των χαρακτηριστικών με κανόνες (*splits*) που μειώνουν την αταξία, φτιάχνοντας μονοπάτια αποφάσεων. Εύχρηστα, ερμηνεύσιμα, με περιορισμούς ωστόσο σε μεγάλη διακύμανση των δεδομένων.
- Random Forest & Extra Trees: Εκπαιδεύει πολλά τυχαιοποιημένα δέντρα (*bootstrap* και τυχαία *thresholds*) και επιλέγεται η πρόβλεψη πλειοψηφικά για να μειώσει τη διασπορά. Ανθεκτικά *ensembles* (σύστημα επί μέρους μοντέλων), καλή απόδοση χωρίς περίτεχνο *tuning* (βελτιστοποίηση παραμέτρων), κατάλληλα για μικτή κλίμακα χαρακτηριστικών.
- KNN (K-Nearest Neighbors): Βρίσκει τα *k* κοντινότερα δείγματα στον χώρο και αποφασίζει με πλειοψηφικό μέσο όρο των γειτόνων. Απλό, χρήσιμο σε καλά κλιμακωμένα και ομοιόμορφης απόστασης δεδομένα, απαιτεί προσοχή σε σπανιότητα (αραιά δεδομένα).
- SVM (Support Vector Machine): Βρίσκει το υπερεπίπεδο με μέγιστο περιθώριο ανάμεσα στις κλάσεις και προβάλλει τα δεδομένα σε χώρο όπου διαχωρίζονται γραμμικά. Αποδοτικό σε μεσαίου μεγέθους σύνολα (μερικές χιλιάδες) με κατάλληλο *kernel* (μετασχηματιστής προβολής των δεδομένων στο υπερεπίπεδο), χρειάζεται *scaling* (κανονικοποίηση στο υπερεπίπεδο).
- Naive Bayes: Υπολογίζει τις υπό συνθήκη πιθανότητες και εφαρμόζει τον κανόνα του Bayes, θεωρώντας ανεξαρτησία ανάμεσα στα χαρακτηριστικά. Γρήγορο, βασικό σε αραιά ή κατηγορικά σενάρια, χρήσιμο ως αναφορά σύγκρισης.
- AdaBoost & Gradient Boosting: Χτίζουν διαδοχικά ασθενείς μαθητές. Ο AdaBoost αναβαθμίζει δύσκολα δείγματα με βάρη, ενώ ο Gradient Boosting προσαρμόζει κάθε νέο δέντρο στο υπόλοιπο (αρνητικό *gradient* - αρνητική κλίση) της απώλειας. E*nsemble* μοντέλα, υψηλή ακρίβεια με προσοχή σε *overfitting*, χρήσιμα όταν υπάρχουν μη γραμμικότητες μεταξύ μεταβλητών.
- Νευρωνικό δίκτυο (NN – Neural Network): Συνθέτει πολλαπλά στρώματα γραμμικών μετασχηματισμών και μη γραμμικών νευρωνικών ενεργοποιήσεων, μαθαίνοντας τα βάρη με ανάδραση σφάλματος (*backpropagation*) για να προσεγγίσει πολύπλοκες σχέσεις. Ευέλικτο σε μεγάλα δεδομένα, απαιτεί προσεκτική προεπεξεργασία και ρύθμιση.

Το κάθε μοντέλο έχει τις δικές του πτυχές, ωστόσο λίγο ή πολύ όλα διέπονται από τις ίδιες αρχές. Μέσω του *validation* διασφαλίζεται η επιτυχής γενίκευση. Για ανεξάρτητες παρατηρήσεις προτείνεται k-fold προσέγγιση, ενώ σε έντονη χρονικότητα *time-based splits*. Κρίσιμο είναι το σωστό σημείο εφαρμογής κλιμάκωσης και επαναδειγματοληψίας, μετά τον χωρισμό σε folds για αποφυγή διαρροής. Σκοπός της βαθμονόμησης που ακολουθεί είναι η αντιστοίχιση πιθανότητας και συχνότητας, ώστε οι αποφάσεις που βασίζονται στο κατώφλι και στο κόστος να είναι αξιόπιστες. Αντίστοιχα, η επιλογή κατωφλιού γίνεται με πίνακα κόστους-οφέλους και συνδέεται με πειραματική τεκμηρίωση.

Επίσης, ορισμένες εξειδικευμένες τεχνικές αυξάνουν την επιχειρησιακή σημασία των μοντέλων. Τέτοιες είναι το *uplift modeling*, το οποίο εκτιμά διαφορικό όφελος επαφής και βελτιστοποιεί τη στόχευση και η ανάλυση επιβίωσης, η οποία αποδίδει χρονικό κίνδυνο αποχώρησης και βοηθά στον χρονικό προγραμματισμό παρέμβασης (Szymański & Kajdanowicz, 2019). Οι τεχνικές αυτές χρησιμοποιούνται συμπληρωματικά, όπου το ερώτημα το απαιτεί.

2.4 Μοντέλα ταξινόμησης με βάση την Python

2.4.1 Λογιστική παλινδρόμηση - Logistic Regression (LR)

Η LR εκτιμά την πιθανότητα ένταξης στην θετική κλάση μέσω λογιστικής συνάρτησης, υποθέτοντας γραμμικότητα στον λογάριθμο των πιθανοτήτων επιτυχίας ή αποτυχίας (log-odds = $\log(\frac{p}{1-p})$). Το όριο απόφασης τίθεται σε κατώφλι πιθανότητας και όχι σε ωμή βαθμολογία, ενώ μπορεί να είναι ευαίσθητο σε κάποια παράμετρο, όπως π.χ. το κόστος, ώστε να μεγιστοποιεί αναμενόμενο κέρδος. Η ερμηνεία γίνεται με συντελεστές σε log-odds. Δηλαδή, θετικός συντελεστής αυξάνει την πιθανότητα ανταπόκρισης, αρνητικός τη μειώνει. Η κλιμάκωση χαρακτηριστικών βελτιώνει τη σταθερότητα και σύγκλιση των εκτιμήσεων και αποτρέπει κυριαρχία μεταβλητών με μεγάλες μονάδες μέτρησης (Peng, Lee & Ingersoll, 2002). Για κατηγορικά χαρακτηριστικά εφαρμόζεται τεχνική μετατροπής κατηγορικής μεταβλητής σε αριθμούς *one-hot encoding*, φτιάχνοντας μία δυαδική στήλη με τιμές 0 ή 1 για κάθε πιθανή τιμή.

Η κανονικοποίηση λειτουργεί ως τροχοπέδη πολυπλοκότητας στην ουσία. L2 ποινή περιορίζει τα βάρη προς το μηδέν, L1 μπορεί να μηδενίσει βάρη και να επιτελέσει επιλογή χαρακτηριστικών, ενώ η *elastic-net* ισορροπεί και τα δύο. Η ισχύς της ποινής ελέγχεται

από την παράμετρο C και επιλέγεται με *CV*, μικρό C = ισχυρότερη ποινή. Σε ανισορροπία κλάσεων, η στάθμιση κλάσεων εξισορροπεί τη συνεισφορά θετικών – αρνητικών, αλλά ενδέχεται να μην χρειαστεί λόγω *reweighting*. Προσοχή απαιτεί η συγγραμμικότητα, καθώς διογκώνει αβεβαιότητες στους συντελεστές, η χρήση ποινής ωστόσο βοηθάει. Μη γραμμικότητες στις σχέσεις (λόγω καμπυλότητας ή εξάρτηση σε άλλο feature) απαιτούν ρητούς όρους, αλλιώς η LR τις χάνει.

Στο πλαίσιο της παρούσας εργασίας, η LR λειτουργεί ως γραμμική βάση αναφοράς για το δυαδικό «Response» (η στήλη πρόβλεψης ανταπόκρισης του μοντέλου). Προσφέρει καθαρή χαρτογράφηση μεταξύ δημογραφικών και συμπεριφορικών δεικτών και πιθανότητας ανταπόκρισης. Η βαθμονόμηση πιθανοτήτων είναι συνήθως καλή, γεγονός που διευκολύνει κατώφλια βάσει κόστους - οφέλους. Η αξιολόγηση δίνει βάρος σε PRC-AUC, F1 και ROC-AUC συμπληρωματικά, όταν το θετικό ποσοστό είναι μικρό.

Στην Python, ο εκτιμητής LR προσφέρει επιλογές για ποινές (L1, L2 και elastic-net), λύτες βελτιστοποίησης (π.χ. liblinear για L1, saga για elastic-net) και στάθμιση κλάσεων (class_weight = ‘balanced’). Η σωστή πρακτική είναι για το *pipeline* να γίνει πρώτα το *encoding*, έπειτα η LR και αν χρειάζεται μετά βαθμονόμηση, όλα εντός του *fold* για αποφυγή διαρροής. Το κατώφλι επιλέγεται στο *validation* με καμπύλες κέρδους ή αναμενόμενο ROI.

2.4.2 Δέντρο Αποφάσεων – Decision Tree (DT)

Τα Decision Trees (DT) χωρίζουν επαναληπτικά τον χώρο χαρακτηριστικών με κανόνες τύπου *If - else* που μειώνουν την αταξία (Gini ή Entropy). Κάθε διαχωρισμός δημιουργεί μονοπάτια ερμηνείας σε επίπεδο παρατήρησης, χρήσιμα όταν το ζητούμενο είναι η αιτιολόγηση προσφορών. Δεν απαιτούν κλιμάκωση και διακρίνουν φυσικά αλληλεπιδράσεις και μη γραμμικότητες, αλλά χρειάζονται μιμητική συμπλήρωση (*imputation*) για κενές τιμές. Απαιτεί προσοχή σε κατηγορικά με πολλές διαφορετικές τιμές, άρα και υψηλού κινδύνου μεροληψίας, γιατί τα δέντρα τείνουν να τα ευνοούν παρουσιάζονται φαινόμενα *overfitting* (Prasetiyowati, Maulidevi & Surendro, 2021).

Τα φαινόμενα αυτά παρουσιάζονται και όταν δεν τεθούν περιορισμοί στο μοντέλο του DT. Εκτός από max_depth, min_samples_split ή leaf, max_leaf_nodes, είναι χρήσιμο το minimal cost-complexity pruning (ccp_alpha) για *post-pruning* με *validation*. Αυτές αποτελούν εντολές που «κλαδεύουν» στην ουσία το δέντρο απόφασης, είτε από πριν περιορίζοντάς το όσο φτιάχνεται (*pre-pruning*), είτε έπειτα φτιάχνοντάς το μεγαλύτερο

και «κλαδεύεται» βάσει κριτήριου κόστους, πολυπλοκότητας κτλ. (*post-pruning*). Μικρότερο βάθος βελτιώνει γενίκευση αλλά μειώνει λεπτομέρεια, ενώ η ισορροπία μεταξύ *bias* και *variance* επιτυγχάνεται με συντηρητικό βάθος (φύλλα) και *pruning*. Οι εκτιμήσεις πιθανότητας από το DT τείνουν να είναι *overconfident* στα φύλλα (Prasetiyowati, Maulidevi & Surendro, 2021). Αν χρησιμοποιούνται για κατώφλια επιθυμητών αποτελεσμάτων τύπου ROI, συνίσταται βαθμονόμηση εντός της CV.

Σε δεδομένα σούπερ μάρκετ, τα DT αποτυπώνουν ευανάγνωστα σημεία καμπής (π.χ. όριο εισοδήματος - οικογενειακή σύνθεση ή κατώφλια έντασης έκπτωσης). Χρίζει προσοχής σε ανισορροπία κλάσεων (απαιτείται και εδώ class_weight='balanced' ή reweighting) και σε μεροληψία σημαντικοτήτων που βασίζονται σε υψηλή εντροπία. Για πιο αξιόπιστη ερμηνεία προτιμάται μέτρηση της πτώσης απόδοσης όταν εμπλέκονται τυχαίες τιμές ενός χαρακτηριστικού (*permutation importance*) ή τοπικές εξηγήσεις. Το DT είναι ένα μόνο δέντρο αποφάσεων και ως μεμονωμένος classifier σπάνια παραμένει σταθερό, γι' αυτό λειτουργεί ως βάση αναφοράς σε *ensembles* (RF και Gradient Boosting) που γενικεύουν καλύτερα, όταν ζητάμε ακρίβεια (Prasetiyowati, Maulidevi & Surendro, 2021).

Στην Python, το DecisionTreeClassifier επιτρέπει επιλογή κριτηρίου, όπως Gini, Entropy ή log_loss, ελέγχους περιορισμού και random_state για αναπαραγωγιμότητα. Ο έλεγχος max_features επηρεάζει την ποικιλία των διασπάσεων. Όταν εφαρμόζεται, θα πρέπει να τεκμηριώνεται ρητά το κριτήριο, το βάθος και το ccp_alpha, καθώς και να συνδέονται ευκρινώς με τις επιχειρησιακές μετρικές (precision κτλ.) και ομοίως και εδώ η επιλογή κατωφλίου να γίνεται μετά από τη βαθμονόμηση.

2.4.3 Τυχαίο Δάσος - Random Forest (RF)

Το RF είναι από τα περισσότερο εφαρμοσμένα μοντέλα ταξινόμησης. Συνδυάζει πολλά δέντρα εκπαιδευμένα σε δειγματοληπτικά υποσύνολα παρατηρήσεων (*bootstrap*) και χαρακτηριστικών (max_features). Η τυχαιοποίηση μειώνει τη συσχέτιση μεταξύ δέντρων και άρα τη διασπορά, κρατώντας χαμηλά τη μεροληψία. Αποτυπώνει μη γραμμικότητες και αλληλεπιδράσεις (π.χ. τιμή με προώθηση και με σύνθεση νοικοκυριού) χωρίς ρητούς όρους. Στη βασική βιβλιοθήκη μηχανικής μάθησης της Python scikit-learn χρειάζεται *imputation* για κενά, καθώς ούτε αυτό χειρίζεται μη συμπληρωμένες τιμές (NaN).

Παρακάτω αναφέρονται οι πιο κρίσιμες υπερπαράμετροι του RF για πρόβλημα

ανταπόκρισης:

n_estimators: αυξάνει τη σταθερότητα. Χρήσιμη η OOB αξιολόγηση (Out of Bag παρατηρήσεις που δεν επιλέχθηκαν στο bootstrap δείγμα κάθε δέντρου, oob_score=True), για να καταστεί σαφές πότε σταθεροποιείται το σφάλμα χωρίς έξτρα *validation*.

max_depth, min_samples_split/min_samples_leaf, max_leaf_nodes: ελέγχουν την προσαρμογή. Συντηρητικό βάθος και μεγαλύτερα φύλλα μειώνουν *overfitting* και βελτιώνουν συχνά την βαθμονόμηση πιθανοτήτων.

criterion (gini ή entropy): μικρές αλλά υπαρκτές διαφορές, η gini είναι λίγο πιο επιθετική και βρίσκει πόσο συχνά θα γίνει το λάθος, ενώ η entropy την αταξία της κατανομής, ως πιο ευαίσθητη σε αμφιλεγόμενες κλάσεις. Εμπειρικά εξετάζεται πρώτα με gini και αν υπάρχουν πολλές ισοβαθμίες, εξετάζεται η entropy και διατηρείται όποια έχει καλύτερη CV απόδοση (Szymański & Kajdanowicz, 2019).

class_weight: balanced (σταθεροί συντελεστές βάρους για όλες τις διασπάσεις) ή balanced_subsample (τα βάρη υπολογίζονται μέσα σε κάθε *bootstrap* δείγμα, άρα κάθε δέντρο έχει λίγο διαφορετικά βάρη) για ανισορροπία, βοηθά ουσιαστικά σε σενάρια υψηλού recall. Δίνει μεγαλύτερο βάρος στη σπάνια κλάση, ώστε το δέντρο να ενδιαφέρεται να μη τη χάσει.

max_features: Κάθε κόμβος εξετάζει ένα τυχαίο υποσύνολο χαρακτηριστικών μεγέθους max_features. Ελέγχει την τυχαιότητα ρυθμίζοντας τη στοχαστικότητα ανά διάσπαση (τυπικό default: sqrt). Μικρότερες τιμές κάνουν τα δέντρα να μοιάζουν λιγότερο, άρα μικρότερη συσχέτιση και χαμηλότερη διασπορά στον μέσο όρο του δάσους.

max_samples: ορίζει πόσα δείγματα (από το σύνολο εκπαίδευσης) θα τραβήξει κάθε δέντρο από το *bootstrap* του (εφόσον bootstrap=True). Μικρότερη τιμή ισούται με πιο γρήγορη εκπαίδευση και λιγότερη συσχέτιση μεταξύ των δένδρων. Η γενίκευση βελτιώνεται με την αύξηση των δέντρων. Ωστόσο, κάθε δένδρο βλέπει λιγότερα δεδομένα και αυξάνεται λίγο η μεροληψία, αντισταθμίζεται όμως με περισσότερα n_estimators.

random_state, n_jobs: αναπαραγωγιμότητα λόγω σταθεροποίησης των τυχαίων διαδικασιών του αλγορίθμου. Ορίζονται οι πυρήνες της CPU που θα χρησιμοποιήσει ο αλγόριθμος για παράλληλη εκτέλεση. Εάν επιθυμούνται όλοι, τότε n_jobs = -1.

Οι πιθανότητες του RF προκύπτουν ως μέσος όρος συχνοτήτων στα φύλλα των δέντρων και μπορεί να είναι *over* ή *under confident*. Οι *impurity-based* σημαντικότητες (ενδεικτικοί μείωσης αταξίας όταν ένα χαρακτηριστικό διασπά το δέντρο, άμεσα σχετιζόμενοι με gini και entropy) δίνουν γρήγορη πρώτη εικόνα, αλλά είναι μεροληπτικοί προς χαρακτηριστικά με πολλές μοναδικές τιμές. Για αξιόπιστη αποτίμηση

χρησιμοποιείται permutation_importance (μετρά πτώση επίδοσης) και TreeSHAP (αξιολογεί συνεισφορές δένδρων), όπου χρειάζεται, για συνολικές ή τοπικές εξηγήσεις. Καλή πρακτική είναι ο έλεγχος σταθερότητας σημαντικοτήτων σε επαναληπτικά *bootstraps*.

Για το validation και τις μετρικές, το *tuning* με *Stratified CV*, όπου χωρίζονται τα δεδομένα σε *folds* με την ίδια αναλογία θετικών – αρνητικών (ή OOB ως βοηθητικός δείκτης), και η εκτέλεση όλων των βημάτων μέσα στο *fold* (scaling, encoding, reweighting, calibrator) διασφαλίζουν την αποφυγή διαρροής. Η αξιολόγηση γίνεται με PRC–AUC, F1, ROC–AUC συμπληρωματικά λόγω χαμηλού θετικού ρυθμού. Στο τελικό τμήμα τεστ δείγματος, οφείλεται να συγκρίνονται και οικονομικές μετρικές, όπως κέρδος, ROI κτλ.

Στην παρούσα εργασία, ο RF ρυθμίστηκε με GridSearchCV πάνω σε πλέγμα που περιλάμβανε criterion, min_samples_split και max_depth, με σταθερό n_estimators. Η τελική επιλογή μοντέλου επικυρώθηκε με διασταύρωση και αποτιμήθηκε στο ανεξάρτητο δείγμα μέσω των μετρικών ταξινόμησης precision, recall και ROC-AUC.

2.4.4 Extra Trees - Extremely Randomized Trees (ET)

Ο ET μοιράζεται τη λογική των δασών, αλλά αυξάνει την τυχαιοποίηση στο σημείο διάσπασης. Αντί να αναζητά το βέλτιστο κατώφλι σε ένα χαρακτηριστικό, δοκιμάζει τυχαία πολλά κατώφλια και επιλέγει το καλύτερο μεταξύ αυτών. Το αποτέλεσμα είναι ακόμη μικρότερη συσχέτιση δέντρων και συχνά χαμηλότερη διασπορά με παρόμοιο ή ταχύτερο χρόνο εκπαίδευσης.

Οι υπερπαράμετροι που φάνηκαν χρήσιμες και στο μοντέλο της παρούσης είναι παρόμοιες με του RF, τόσο στην λειτουργία όσο και στην χρηστικότητα τους. Αυτές είναι οι n_estimators, κριτήριο gini ή entropy, max_depth ή min_samples_leaf και max_features. Πλεονέκτημα του ET είναι η ταχύτητα και η ανθεκτικότητα γενίκευσης όταν υπάρχει θόρυβος, ειδικά σε σύνολα με πολλά, εν μέρει συσχετισμένα χαρακτηριστικά. Σε προβλήματα ανταπόκρισης όπου προέχει η κατάταξη και η σταθερότητα σε OOB σενάρια, συχνά αποδίδει πολύ κοντά, ή καλύτερα από RF, με λιγότερο tuning. Και τα δύο μοντέλα προσφέρουν ισχυρή ακρίβεια χωρίς ανάγκη κλιμάκωσης και χειρίζονται φυσικά τη μηχανική αλληλεπιδράσεων από μόνα τους, χωρίς φτιαχτά *features*, όπως στα γραμμικά μοντέλα.

Στο πλαίσιο του παρόντος μοντέλου, ο ET ρυθμίστηκε με GridSearchCV πάνω σε

n_estimators και criterion. Η τελική διαμόρφωση αξιολογήθηκε με τις ίδιες μετρικές ταξινόμησης και ελέγχθηκε με πίνακα σύγχυσης και καμπύλες ROC και PRC. Για χρήση κατωφλιού βάσει κόστους, ισχύει η ίδια οδηγία με τον RF, δηλαδή ότι απαιτείται βαθμονόμηση όταν οι πιθανότητες θα τροφοδοτήσουν αποφάσεις στόχευσης.

2.4.5 K-Nearest Neighbors (KNN)

Ο KNN, ακόμη ένας συχνά εφαρμοζόμενος αλγόριθμος ταξινόμησης, λειτουργεί με βάση τη γειτνίαση. Κάθε παρατήρηση λαμβάνει την κλάση που υπερισχύει στους k κοντινότερους γείτονες. Η απόσταση λειτουργεί ως μέτρο ομοιότητας, άρα η κλιμάκωση των χαρακτηριστικών (π.χ. *standardization* ή *robust scaling*) είναι απαραίτητη για να μην κυριαρχούν μεταβλητές με μεγάλες μονάδες. Η επιλογή μετρικής επηρεάζει ουσιαστικά τη συμπεριφορά. Η Minkowski με p=2 (Ευκλείδεια μέτρηση απόστασης) είναι η προεπιλογή, αλλά η p=1 (Manhattan) είναι πιο ανθεκτική σε τιμές εκτοξευτές. Το μοντέλο είναι μη παραμετρικό και εκπαιδεύεται σε τοπικούς κανόνες, χρήσιμο όταν οι σχέσεις είναι ανομοιογενείς στον χώρο των χαρακτηριστικών.

Η επιλογή του k ελέγχει τη λειτουργία εξομάλυνσης. Μικρό k μειώνει *bias* αλλά αυξάνει διασπορά και ευαισθησία σε θόρυβο. Μεγάλο k αυξάνει ροπή προς την πλειονότητα και μπορεί να λειάνει υπερβολικά τα όρια. Το μειονέκτημα της διαστασιμότητας (πλήθος χαρακτηριστικών στον χώρο) είναι ότι αποδυναμώνεται η ισχύς των αποστάσεων όσο μεγαλώνει διάσταση. Ως αντιμετώπιση προτείνεται συμπίεση ή επιλογή χαρακτηριστικών, με τη χρήση μεθόδων όπως η PCA και η UMAP για ανίχνευση χαμηλότερης δομής, ή εποπτευόμενη επιλογή με κανονικοποίηση (Halder κ.ά., 2024). Σε ανισορροπία κλάσεων, βαρύτητες ανά απόσταση (weights = ‘distance’) και ρύθμιση κατωφλιού στην πιθανότητα ψήφου περιορίζουν τα ψευδώς θετικά. Εναλλακτικά εφαρμόζεται επαναδειγματοληψία (*SMOTE* και *undersampling*) επίσης μέσα στο fold. Η πιθανότητα κλάσης ερμηνεύεται ως ποσοστό ψήφων στους *k* γείτονες, αλλά είναι συχνά φτωχά βαθμονομημένη, καθώς οι λίγοι γείτονες καθιστούν τα τοπικά ποσοστά θορυβώδη, ενώ μεγάλοι *k* τα λειαίνουν υπερβολικά (Halder κ.ά., 2024).

Από πλευράς υλοποίησης, κρίσιμες ρυθμίσεις είναι n_neighbors (πλήθος γειτόνων που κοιτάει), metric (μέτρο απόστασης), weights (πως ψηφίζουν οι γείτονες) και ο αλγόριθμος αναζήτησης γειτόνων (algorithm = ‘auto’ με kd_tree για μεσαίες διαστάσεις με συνεχείς μεταβλητές και ball_tree για πιο ιδιάζουσες κατανομές). Σε μέτριες διαστάσεις τα δέντρα επιταχύνουν την αναζήτηση. Σε πολύ υψηλή διάσταση η απόδοση συγκλίνει σε

εξαντλητική σύγκριση (*brute force*). Το KNN δεν έχει εκπαίδευση με βάρη, άρα το κόστος είναι μετατοπισμένο στον χρόνο πρόβλεψης. Για πολύ μεγάλα σύνολα συναλλαγών μπορούν να εξεταστούν προσεγγιστικές μέθοδοι γειτόνων.

Στο παρόν μοντέλο βρέθηκε βέλτιστο n_neighbors=1, ένδειξη έντονα τοπικής δομής μετά το scaling. Αυτό προσφέρει υψηλή ευαισθησία σε μικρές συγκεντρώσεις θετικών, αλλά απαιτεί αυστηρό έλεγχο σταθερότητας με διασταύρωση και πιθανή εξομάλυνση (δοκιμή με k>1 ώστε η πρόβλεψη να είναι μέσος όρος ψήφων μεγαλύτερης γειτονιάς ή με weights = ‘distance').

2.4.6 Support Vector Machine (SVM)

ΤΟ SVM αναζητά υπερεπίπεδο που μεγιστοποιεί το περιθώριο μεταξύ κλάσεων. Με *kernels* προβάλλει τα δεδομένα σε χώρο όπου επιτυγχάνεται γραμμικός διαχωρισμός. Η μέθοδος απαιτεί αυστηρή κλιμάκωση χαρακτηριστικών (π.χ. StandardScaler σε *pipeline*) ώστε οι αποστάσεις να είναι συγκρίσιμες και ο όρος κανονικοποίησης να λειτουργεί σταθερά.

Οι βασικές υπερπαράμετροι (C, kernel, gamma, degree, class_weight) καθορίζουν το σχήμα του ορίου. Το C ρυθμίζει την εναλλαγή μεταξύ σφαλμάτων εκπαίδευσης και πλάτους περιθωρίου (μικρό C = φαρδύ περιθώριο και μεγαλύτερο bias). Ο *kernel* ορίζει τον μετασχηματισμό (γραμμικό ή RBF ή πολυωνυμικό) και η gamma ελέγχει το εύρος επιρροής (πόσο μακριά δηλαδή επηρεάζει ένα σημείο εκπαίδευσης το όριο απόφασης) για RBF ή Poly (με το gamma = ‘scale’ συνήθως ασφαλέστερο από auto). Στον πολυωνυμικό μετασχηματισμό προστίθενται degree και coef0, οι οποίοι ελέγχουν καμπυλότητα και την επίδραση του σταθερού όρου coef0 (*offset*). Η ανισορροπία αντιμετωπίζεται με class_weight = ‘balanced’ ή/και ρύθμιση κατωφλιού στην decision_function.

Οι πιθανότητες του SVM δεν είναι εγγενώς αξιόπιστες. Όταν απαιτούνται επιχειρησιακά κατώφλια τύπου ROI, εφαρμόζεται βαθμονόμηση με CalibratedClassifierCV εντός διασταύρωσης, ακολουθούμενη από επιλογή κατωφλιού βάσει PRC και καμπύλες κέρδους. Σε μεγάλους όγκους και διαστάσεις προτιμώνται γραμμικές παραλλαγές (LinearSVC ή SGDClassifier με hinge loss) για υπολογιστική αποδοτικότητα (Passos & Mishra, 2022). Για μικρότερα και μεσαία σύνολα ο RBF είναι συχνά ισχυρή προεπιλογή, ενώ πολυωνυμικός *kernel* (π.χ. degree = 5) υποδηλώνει ανάγκη για υψηλής τάξης αλληλεπιδράσεις, αλλά απαιτεί αυστηρό έλεγχο *overfitting* με διασταύρωση και παρακολούθηση του αριθμού στηριγμάτων (support vectors) που επηρεάζει τη γενίκευση

(Passos & Mishra, 2022). Από πλευράς ερμηνείας, αξιοποιούνται αποστάσεις από το υπερεπίπεδο και τοπικές εξηγήσεις (π.χ. SHAP στη decision_function).

Στη μελέτη που πραγματεύεται η παρούσα εργασία, το grid κατέληξε σε πολυωνυμικό πυρήνα με degree = 5, C = 5 και gamma = 'auto', επιλογή που αυξάνει την εκφραστικότητα, την ικανότητα δηλαδή του μοντέλου να προσεγγίζει πολύπλοκα μοτίβα. Απαιτεί όμως προσεκτικό *tuning* για αποφυγή *overfitting* και κατώφλι προσαρμοσμένο αντίστοιχα.

2.4.7 Naive Bayes (NB)

Ο NB, ως ακόμη ένας πολυχρησιμοποιημένος classifier, εκτιμά την πιθανότητα κλάσης μέσω του κανόνα του Bayes κάνοντας την αφελή υπόθεση, υπό όρους ανεξαρτησίας των χαρακτηριστικών μέσα σε κάθε κλάση. Έτσι, το συνολικό αποτέλεσμα γράφεται ως γινόμενο επιμέρους πιθανοτήτων και στην πράξη υπολογίζεται σε λογαριθμική κλίμακα, ώστε οι συνεισφορές να αθροίζονται σταθερά αριθμητικά. Η υπόθεση ανεξαρτησίας σπάνια ισχύει πλήρως, αλλά συχνά δίνει σωστή κατάταξη, ακόμη κι αν οι απόλυτες πιθανότητες είναι κακώς βαθμονομημένες, γι' αυτό ο NB είναι γρήγορος και σταθερό σημείο αναφοράς (Peretz, Koren & Koren, 2024). Όταν υπάρχουν ισχυρά συσχετισμένα γνωρίσματα, ο NB τείνει να λαμβάνει υπόψη δύο φορές την ίδια πληροφορία, ενισχύοντας υπερβολικά έτσι την πιθανότητα. Αυτό μετριάζεται με επιλογή ή αποσύνδεση χαρακτηριστικών (π.χ. με PCA ή εποπτευόμενη επιλογή), ό,τι και να επιλεγεί ωστόσο πρέπει να γίνεται μέσα στη CV για αποφυγή *data leakage*.

Η παραλλαγή επιλέγεται από τη φύση των γνωρισμάτων. Η GaussianNB μοντελοποιεί συνεχείς μεταβλητές με κανονικές κατανομές ανά κλάση. Η ρύθμιση var_smoothing σταθεροποιεί εκτιμήσεις όταν οι διακυμάνσεις είναι πολύ μικρές ή τα δείγματα λίγα. Η MultinomialNB ταιριάζει σε καταμετρήσεις μη αρνητικές (ειδικά συχνότητες ή μετρήσεις) και χρησιμοποιεί εξομάλυνση Laplace με παράμετρο alpha για να αποφύγει μηδενικές πιθανότητες σε σπάνιες τιμές. Η παραλλαγή ComplementNB είναι συχνά πιο ανθεκτική σε ανισορροπία, διότι αντί να εκτιμά τα βάρη ενός όρου για μια κλάση απευθείας από τα δείγματα της, τα εκτιμά από τα υπόλοιπα δείγματα, ως συμπλήρωμά της. Η BernoulliNB δουλεύει με δυαδικά γνωρίσματα (παρουσία και απουσία). Σε σούπερ μάρκετ με δημογραφικά χαρακτηριστικά και συνεχείς δείκτες (εισόδημα, συχνότητες επισκέψεων, ένταση πρόσφατων αγορών κτλ.), το GaussianNB είναι συνήθως η ορθή εκκίνηση. Αν ορισμένα γνωρίσματα είναι δεδομένα καταμέτρησης (π.χ. πλήθος αγορών

ανά κατηγορία) το MultinomialNB ενδέχεται να αποδώσει καλύτερα μετά από κατάλληλο μετασχηματισμό (π.χ. *scaling* μη αρνητικών).

Για επιχειρησιακές αποφάσεις η ανισορροπία δεν αντιμετωπίζεται με class weights (δεν υπάρχουν βάρη στον NB), αλλά με κατώφλι απόφασης, επαναδειγματοληψία και τις μετρικές αξιοπιστίας, αντί για απλή ακρίβεια. Ο NB είναι επίσης χρήσιμος ως τεστ διαρροής. Αν αποδώσει υπερβολικά καλά χωρίς σωστό *pipeline* (να έχει προηγηθεί δηλαδή imputation ή κλιμάκωση ή επιλογή χαρακτηριστικών κτλ.) μέσα στη διασταύρωση, τίθεται υποψία διαρροής Peretz, Koren & Koren, 2024). Από πλευράς ερμηνείας, οι λογαριθμικές συνεισφορές ανά γνώρισμα δίνουν καθαρό αποτύπωμα του τι ωθεί την απόφαση, ενώ στο τελικό στάδιο το κατώφλι επιλέγεται με καμπύλες PRC και κέρδους.

2.4.8 AdaBoost

Το AdaBoost είναι σταδιακή πρόσθεση ασθενών μαθητών, ή αλλιώς βάσεων, (συνήθως ρηχά δέντρα) που εκπαιδεύονται διαδοχικά με αναβάθμιση βαρών σε δείγματα που σφάλανε νωρίτερα. Στη δυαδική εκδοχή του ελαχιστοποιεί τη συνάρτηση απώλειας (*exponential loss*), βάζοντας εκθετικά μεγάλο βάρος στα δείγματα με λάθος πρόσημο σε κάθε επανάληψη. Το τελικό σκορ προκύπτει ως ζυγισμένο άθροισμα των βάσεων, αφού δεν συνεισφέρει ισότιμα η κάθε μία, με τις πιο ακριβείς να έχουν μεγαλύτερο βάρος.

Στην υλοποίηση SAMME.R της scikit-learn το σκορ συνδέεται λογιστικά και δίνει ομαλότερες πιθανότητες από το κλασικό SAMME. Η ισορροπία απόδοσης με σταθερότητα ελέγχεται κυρίως από n_estimators και learning_rate με παράμετρο shrinkage. Περισσότερα στάδια αυξάνουν την ισχύ του μοντέλου, ενώ μικρό learning_rate δρα ως φρένο πολυπλοκότητας, απαιτώντας ωστόσο περισσότερους εκτιμητές (Szymański & Kajdanowicz, 2019). Η βάση (ρηχά δέντρα βάθους 1-3) καθορίζει την εκφραστικότητα των διορθώσεων και υλοποιείται ως DecisionTreeClassifier(max_depth=1/2/3 ή max_leaf_nodes=2/4/8). Υπερβολικά δυνατές βάσεις θα πρέπει να αποφεύγονται, καθώς υπάρχει κίνδυνος να οδηγήσουν σε *overfitting.*

Κρίσιμες τεχνικές λεπτομέρειες για το AdaBoost είναι πως παρουσιάζει ευαισθησία σε θόρυβο ή λανθασμένες ετικέτες και επειδή τα δείγματα με ακραίες τιμές παίρνουν αυξανόμενα βάρη, χρειάζεται έλεγχος ποιότητας δεδομένων. Επίσης είναι πολύ χρήσιμο να εφαρμοστεί όπου αρμόζει, *early stopping* με CV (σταματά όταν η μετρική του

μοντέλου παύει να βελτιώνεται). Σε ανισορροπία κλάσεων, προτιμάται η πρακτική με στάθμιση στη βάση (π.χ. class_weight στα δέντρα) ή sample weights στο στάδιο εκπαίδευσης (*fit*) και επιλογή κατωφλιού εκ νέου στο τελικό σκορ βάσει επιθυμητού γνώμονα. Σε προβλήματα πολλών κλάσεων, προτιμάται SAMME.R για καλύτερη σταθερότητα πιθανοτήτων. Από πλευράς ερμηνείας, οι feature_importances_ που προκύπτουν από τις βάσεις δίνουν μια πρώτη εικόνα αξιολόγησής, αλλά για πιο αμερόληπτη αποτίμηση προτιμάται η permutation_importance, όπου μετριέται η πτώση της επίδοσης (π.χ. με PRC-AUC) ανακατεύοντας τις τιμές ενός *feature* στο *validation*.

Στην προσέγγιση της παρούσης κατά την υλοποίηση του AdaBoost, η αναζήτηση υπερπαραμέτρων έγινε με πλέγμα στο learning_rate (σταθερό n_estimators=500) και το βέλτιστο αναδύθηκε κοντά στο 1.7 σε CV, ένδειξη έντονων γνωρισμάτων που επιτρέπουν πιο επιθετικό ρυθμό χωρίς μεγάλη απώλεια γενίκευσης, Η τελική επίδοση αποτιμήθηκε σε ανεξάρτητο δείγμα με *confusion matrix* και ταξινομητικά σκορ. Πρακτικά, το AdaBoost ταιριάζει σε σενάρια ανταπόκρισης όπου τα γνωρίσματα είναι καθαρά, αλλά ανομοιόμορφα (π.χ. ισχυρή επίδραση συγκεκριμένων προωθήσεων σε υποομάδες).

2.4.9 Gradient Boosting (GB)

Ο GB μαθαίνει διαδοχικές διορθώσεις στο υπόλοιπο σφάλμα της τρέχουσας πρόβλεψης. Σε κάθε στάδιο ένα ρηχό δέντρο (weak learner) προσαρμόζεται στο αρνητικό gradient της επιλεγμένης συνάρτησης απώλειας και προστίθεται με μικρή διόρθωση (*shrinkage*) και ρυθμό εκμάθησης, άρα και βαθμό επίδρασης, ότι προστάζει η παράμετρος learning rate. Έτσι, το σύνολο αποτυπώνει μη γραμμικότητες και αλληλεπιδράσεις χωρίς ρητούς όρους. Στη δυαδική ταξινόμηση χρησιμοποιείται συνήθως λογιστική απώλεια, αλλά εναλλακτικές (π.χ. deviance με robust επιλογές) είναι χρήσιμες όταν υπάρχουν εκτοξευτές τιμών και θόρυβος.

Κεντρικοί μοχλοί ελέγχου είναι το learning rate, ο αριθμός σταδίων n_estimators και η πολυπλοκότητα του base learner (π.χ. max_depth ή max_leaf_nodes, min_samples_leaf ή min_samples_split). Τυπικά, μικρό learning rate και περισσότερα στάδια μειώνει τη διασπορά και βελτιώνει τη γενίκευση. Με την επιλογή της υπερπαραμέτρου subsample < 1 (stochastic GB) εισάγεται τυχαιοποίηση τύπου *bagging* (τεχνική μείωσης της διασποράς) και λειτουργεί ως κανονικοποίηση μειώνοντας τη συσχέτιση μεταξύ των δέντρων. Επίσης, η υποδειγματοληψία χαρακτηριστικών (π.χ. max_features) περιορίζει τη συσχέτιση σταδίων, κάνοντας τα όλα να βασίζονται σε ελάχιστα και ίδια για όλα τα

δέντρα *features*, με αποτέλεσμα συχνά να σταθεροποιεί την επίδοση (Natekin & Knoll, 2013).

Από πλευράς εκπαίδευσης, ο GB ευνοεί *early stopping* σε *validation*, ώστε να παγώνει όταν παύει να βελτιώνεται η PRC–AUC. Αυτό παρουσιάζεται ιδιαίτερα χρήσιμο όταν το learning rate είναι σχετικά υψηλό. Η μέθοδος απαιτεί imputation για κενές τιμές και σωστή κλιμάκωση. Αν υπάρχουν απόλυτα εξαρτημένα χαρακτηριστικά, κατά κανόνα δεν θα δώσουν το ίδιο *split*, συνεπώς κρατείται το 1 από τα 2, ώστε αφενός να μην δημιουργηθεί πλεονασμός και αφετέρου να μην αυξηθεί ο κίνδυνος αστάθειας. Η ερμηνεία υποστηρίζεται από σημαντικότητες χαρακτηριστικών (impurity και permutation) και PDPs (Partial Dependence Plots) για να φανεί το μέσο οριακό αποτέλεσμα ενός *feature*. Για τοπικές εξηγήσεις μπορούν να χρησιμοποιηθούν και εδώ SHAP τιμές πάνω στα δέντρα.

Στην εφαρμογή του GB στην εργασία, η αναζήτηση έγινε με n_estimators = 300 και πλέγμα στο learning rate. Οι υψηλότερες τιμές έδειξαν καλή απόδοση στη διασταυρούμενη επικύρωση (π.χ. 0.5), ενώ το τελικό μοντέλο διατηρήθηκε με learning rate περίπου 0.3, ένδειξη έντονων γνωρισμάτων που ωφελούνται από ήπια κανονικοποίηση. Πρακτικά, προτείνεται έλεγχος της πολυπλοκότητας δέντρων (π.χ. max_depth ή max_leaf_nodes, επαρκές min_samples_leaf) και ενεργοποίηση subsample ή max_features για πρόσθετη σταθερότητα.

2.4.10 Νευρωνικά Δίκτυα – Neural Networks (NN)

Τα NN είναι ενδεχομένως ο περισσότερο πολύπλοκος αλγόριθμος ταξινόμησης. Είναι ένας τρόπος παροχής πληροφόρησης σχέσεων από δεδομένα, χωρίς να γράφονται ως κανόνες. Αποτελούν επί μέρους στρώματα απλών υπολογισμών. Κάθε στρώμα δέχεται αριθμούς, κάνει μια γραμμική μίξη (σαν σταθμισμένο άθροισμα) και φιλτράρει το αποτέλεσμα από μια απλή μη γραμμική λειτουργία (π.χ. ReLU που κόβει τα αρνητικά στο μηδέν). Στοιβάζοντας τέτοια στρώματα, το δίκτυο δύναται να προσεγγίσει πολύπλοκα μοτίβα, με ερωτήματα που συνδυάζουν τιμές από πολλά *features* για να προβλέψει την τιμή ενός άλλου, της ανταπόκρισης.

Η εκπαίδευσή του NN γίνεται με *backpropagation*. Το δίκτυο κάνει μια πρόβλεψη, συγκρίνει με την πραγματική τιμή μέσω μιας συνάρτησης απώλειας (π.χ. για δυαδική ανταπόκριση), μετρά το σφάλμα και έπειτα προσαρμόζει λίγο τα βάρη του ώστε να

ελαχιστοποιείται το σφάλμα σε κάθε επανάληψη. Αυτό επαναλαμβάνεται πολλές φορές, με μικρά πακέτα δεδομένων (*mini-batches*) και ένας βελτιστοποιητής (Adam) αποφασίζει το μέγεθος των βημάτων διόρθωσης. Αν το μέγεθος είναι υπερβολικά μεγάλο, το μοντέλο δεν σταθεροποιείται, ενώ αν είναι υπερβολικά μικρό αργεί ή ακινητοποιείται. Γι' αυτό συνήθως εφαρμόζεται πρόγραμμα ρυθμού μάθησης που ξεκινά πιο γρήγορα και πέφτει σταδιακά η ταχύτητα (Klabunde κ.ά., 2025).

Λόγω της μεγάλης τους χωρητικότητας σε μοτίβα, τα NN μπορούν εύκολα να εκπαιδευτούν και στον θόρυβο και να απομνημονεύσουν τις ιδιαιτερότητες του δείγματος, χάνοντας όμως γενίκευση. Εδώ συμβάλλουν οι υπερπαράμετροι όπως το *early stopping* και το *dropout* που ακυρώνει τυχαία νευρώνες στη διάρκεια της μάθησης για να μειωθεί η εξάρτηση μεταξύ τους. Το weight decay (L2) τιμωρεί πολύ μεγάλα βάρη, ώστε το μοντέλο να μην γίνεται πολύ σύνθετο. Σχεδόν πάντα χρειάζεται scaling των αριθμητικών χαρακτηριστικών, ώστε όλα να κινούνται σε παρόμοιες κλίμακες. Στα κατηγορικά, αν είναι πολλά και σύνθετα, αντί για τεράστιο one-hot χρησιμοποιούνται *embeddings* (εκπαιδεύονται μικρά διανύσματα που συμπυκνώνουν την πληροφορία της κατηγορίας).

Όπως ήδη αναφέρθηκε, τα προβλήματα σούπερ μάρκετ η ανισορροπία είναι κανόνας, καθώς λίγοι ανταποκρίνονται και πολλοί όχι. Στα NN αυτό αντιμετωπίζεται δίνοντας μεγαλύτερο βάρος στα σπάνια θετικά ή με *focal loss,* που εστιάζει στα δύσκολα παραδείγματα. Για τη συνολική εικόνα, αξιολογούνται σημαντικότητες χαρακτηριστικών με *permutation*. Για τοπικές εξηγήσεις ανά πελάτη χρησιμοποιείται SHAP για αποκωδικοποίηση των γνωρισμάτων που σπρώχνουν προς ή μακριά από τη θετική κλάση. Στο παρόν πλαίσιο, τα NN λειτουργούν συμπληρωματικά προς τα δενδροειδή *ensembles*. Προσφέρουν εκφραστικότητα όταν υπάρχουν υψηλής τάξης σχέσεις, αλλά πριν την χρήση τους απαιτούν πειθαρχημένο tuning και ρητή βαθμονόμηση.

ΚΕΦΑΛΑΙΟ 3. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

3.1 Υπάρχουσα έρευνα στην πρόβλεψη συμπεριφοράς πελατών σούπερ μάρκετ

Γενικά, η έρευνα στη λιανική αξιοποιεί συναλλαγές και στοιχεία πιστότητας για να εξηγήσει επιλογές καταστήματος και ανταπόκριση σε ερεθίσματα. Στη μελέτη των Sharma κ.ά. (2025), για σημεία πώλησης που έχουν στο επίκεντρο την υγεία, η προτίμηση μεγάλων αλυσίδων σχετίζεται με πληρέστερη πληροφόρηση, όπως προέλευση και

πιστοποιήσεις, ενώ η εμπιστοσύνη σε θεσμούς ενισχύει τόσο οργανωμένες αγορές όσο και μεγάλες αλυσίδες, με τις κοινωνικές συστάσεις να ωφελούν τα μικρά καταστήματα. Η αντιλαμβανόμενη πρόσβαση δρα καταλυτικά, ιδίως σε περιοχές με περιορισμένες επιλογές. Τα ευρήματα υποδεικνύουν ότι η ποιότητα πληροφόρησης και η πρόσβαση είναι σταθεροί προγνωστικοί παράγοντες επιλογής καναλιού.

Στο πλαίσιο του *omnichannel*, η θεωρία λόγων υπέρ / κατά, όπου οι πελάτες ζυγίζουν θετικά και αρνητικά πριν χρησιμοποιήσουν ένα κανάλι, δείχνει ότι η αντιλαμβανόμενη ποιότητα προϊόντος και η ευελιξία αγοράς ευνοούν την υιοθέτηση ψηφιακής βιτρίνας των φυσικών αλυσίδων. Αντίθετα, ανησυχίες για καθυστερήσεις παράδοσης και ελκυστικές εναλλακτικές λειτουργούν αποτρεπτικά. Η επίδραση διαφοροποιείται ανά τύπο κατηγορίας (π.χ. ηλεκτρικά - ένδυση), καθώς οι προσδοκίες είναι αντικειμενικές και η ποιότητα είναι ευκολότερο να αξιολογηθεί διαδικτυακά, τα υπέρ τείνουν να υπερισχύον (Padmanabhan κ.ά., 2025). Στην ένδυση, όπου το ρίσκο κακής εφαρμογής, υφής κτλ. είναι υψηλότερο, τα κατά έχουν μεγαλύτερο βάρος. Αυτή η διαφοροποίηση ανά κατηγορία είναι κρίσιμη στο πλαίσιο στοχευμένης πρόβλεψης και σχεδιασμό καναλιού.

Η εικόνα καταστήματος και η ικανοποίηση έχουν τεκμηριωθεί σε ελληνικό δείγμα στην έρευνα των Theodoridis & Chatzipanagiotou (2009), με αναλύσεις πολλών ομάδων. Διαπιστώνεται ότι ιδίως τα χαρακτηριστικά που αφορούν το προϊόν και την ποικιλία του και η πολιτική τιμολόγησης, αποτελούν κοινούς οδηγούς ικανοποίησης σε διαφορετικά προφίλ αγοραστών, ενώ τα υπόλοιπα γνωρίσματα παρουσιάζουν μεταβλητότητα. Το εύρημα στηρίζει μοντέλα που σταθμίζουν τις ίδιες διαστάσεις με διαφορετικές βαρύτητες ανά τμήμα, όπως στην περίπτωση της έρευνας των Allaway κ.ά. (2011). Η εικόνα του καταστήματος προσεγγίζεται ως σύνολο διαστάσεων στρατηγικής που στηρίζουν την καταναλωτική μνήμη και διαμορφώνουν την αφοσίωση. Μέσω ανάλυσης παραγόντων αναδεικνύουν οχτώ οδηγούς (επίπεδο εξυπηρέτησης, ποιότητα - ποικιλία προϊόντος, προγράμματα επιβράβευσης, προσπάθεια διατήρησης πελατών, τιμές, διάταξη, τοποθεσία και εμπλοκή στην κοινότητα) και δύο εκβάσεις βάσει αξίας της μάρκας (συναισθηματική αφοσίωση και φανατισμός) ως διακριτοί μηχανισμοί απόφασης (Allaway κ.ά., 2011).

Για την στάθμιση, οι εκτιμήσεις προκύπτουν με LR των δύο εκβάσεων πάνω στους οδηγούς. Ως προς τη συναισθηματική αφοσίωση, η προσπάθεια διατήρησης πελατών έχει τον υψηλότερο συντελεστή, έπειτα ακολουθεί η ποιότητα - ποικιλία και κατόπιν το επίπεδο εξυπηρέτησης, ενώ τα προγράμματα επιβράβευσης έχουν αρνητικό πρόσημο. Για τον φανατισμό, κυριαρχούν εκ νέου η προσπάθεια και η ποιότητα - ποικιλία, με θετικές αλλά μικρότερες επιδράσεις από την εμπλοκή στην κοινότητα. Έτσι, η εικόνα

καταστήματος δεν είναι μονοδιάστατη, αλλά με ποικιλία γνωρισμάτων να φέρουν διαφορετικά βάρη στις εκβάσεις της πιστότητας.

Σε σχεδιασμό εντός του καταστήματος, όπως μελετούν οι Silverio κ.ά. (2025), προσεγγίσεις διαγραμματικού σχήματος (*planogram*), με γνώμονα τη συμπεριφορά σε ιεραρχική ταξινόμηση, μετρούν την ομοιότητα διατάξεων ραφιού βάσει αναμενόμενης διαδρομής και παράλληλων αγορών. Τα αποτελέσματα δείχνουν ότι ισοδύναμες, από πλευράς συμπεριφοράς, διατάξεις μπορούν να αναγνωριστούν και να αναδιαταχθούν για καλύτερη απόδοση. Συνεχίζοντας στη μελέτη τους αναφέρουν επίσης πως η «*λογική της αποστολής αγοράς στηρίζεται εμπειρικά από σχέσεις κατηγορίας και καλαθιού»* (Silverio κ.ά., 2025). Η ορθή γειτνίαση συμπληρωματικών προϊόντων σε ράφι ενισχύει προσθήκες στο καλάθι και αυξάνει την αξία συναλλαγής. Η γνώση αυτή τροφοδοτεί χαρακτηριστικά μοντέλων για συμπληρωματικές προτάσεις και *cross-selling*. Επί τούτου, η χρήση δεδομένων συμπεριφοράς για σχεδιασμό ραφιού αποτελεί πρακτικό παράδειγμα μετατροπής ευρημάτων σε λειτουργικές παρεμβάσεις.

Σύμφωνα με την έρευνα των Stylianou και Pantelidou (2025), στην ανάλυση καλαθιού και τμηματοποίηση μελέτες μεγάλης κλίμακας συνδυάζουν κανόνες συσχέτισης Apriori, K-Means, συστήματα συστάσεων και μοντέλα χρονοσειρών (ARIMA). Αναδεικνύονται σταθερά μοτίβα συγγενών αγορών και ευκρινή τμήματα με βάση συνήθειες και εποχικότητα στη ζήτηση. Σημαντική είναι η ανάδειξη σημαντικότητας της ενοποίησης συναλλαγών και πιστότητας, διευκολύνοντας τις προβλέψεις που τροφοδοτούν μάρκετινγκ, αποθέματα και πολιτικές τιμών.

Όσον αφορά την ανταπόκριση σε προωθήσεις, τα παραπάνω ευρήματα συνδέονται με την ανάγκη συνέπειας τιμής και προσφορών στα κανάλια. Η μελέτη των Padmanabhan κ.ά. (2025), στα *omnichannel* αναδεικνύουν ότι οι ασυμφωνίες μεταξύ ραφιού, φυλλαδίου και εφαρμογής υπονομεύουν την πρόθεση χρήσης του διαδικτυακού καναλιού της ίδιας αλυσίδας και μειώνουν την αξιοπιστία. Ως εκ τούτου, η συνέπεια πολιτικής βελτιώνει την πιθανότητα ανταπόκρισης και την πρόθεση επαναγοράς. Εξετάζοντας υγιεινά περιβάλλοντα αγοράς, ανακύπτουν μεταβλητές για πρόβλεψη επιλογής σημείου πώλησης, βασιζόμενες σε στοιχεία όπως τύποι πληροφόρησης, εμπιστοσύνη σε πιστοποιήσεις και αντιλαμβανόμενη προσβασιμότητα. Οι μεταβλητές αυτές μπορούν να αποδοθούν ως χαρακτηριστικά πελάτη, χρόνου και κατηγορίας και να βελτιώσουν τους *classifiers* για επιλογή καναλιού ή καταστήματος.

Σε ακόμη περισσότερο ψηφιακά οικοσυστήματα, μελέτες *social* και εμπορίου ζωντανής μετάδοσης (*live commerce*) δείχνουν ότι η ποιότητα πληροφορίας, η ποιότητα υπηρεσίας

και η διάδοση από στόμα σε στόμα ενισχύουν δέσμευση και πρόθεση επαναγοράς (Herzallah, 2025). Παράλληλα, στην εμπειρική τους μελέτη οι Xu κ.ά. (2024), εξετάζοντας τη συμπεριφορά του κοινού σε *live commerce* πλατφόρμες, κατέληξαν πως οι εναλλαγές συναισθημάτων και η διατήρηση κοινού σε ζωντανές ροές συνδέονται έντονα με πρόθεση αγοράς. Τα συμπεράσματα αυτά στηρίζει και η μελέτη των Li, Frutos και Ortega (2025), πηγαίνοντας ένα βήμα παραπέρα καταλήγουν πως αυτές οι ενδείξεις υποδηλώνουν ανάγκη για νέους δείκτες στα μοντέλα ανταπόκρισης σε ψηφιακές καμπάνιες.

Συμπερασματικά, ως προς τις τεχνικές, η χρήση συνδυαστικών μεθόδων (κανόνες συσχέτισης, τμηματοποίηση, πρόγνωση) σε ενιαίο πλαίσιο φαίνεται να αποδίδει όταν υπάρχουν μεγάλα σύνολα συναλλαγών. Η σύνδεση προτύπων μικρόκοσμου, όπως παράλληλες αγορές τύπου *halo*, με μακροοικονομικούς δείκτες κατανάλωσης (πληθωρισμός κτλ.) αναδεικνύεται ως προοπτική πολιτικής, αλλά και ως πηγή *drift* στο μοντέλο. Αυτό συμβαίνει διότι αν αλλάξει κάτι στο μάκρο περιβάλλον, ενδέχεται να επηρεάσει τα μοτίβα στο μίκρο, οπότε απαιτεί τακτική επαναξιολόγηση.

Εντούτοις, η βιβλιογραφική έρευνα στις εφαρμογές πρόβλεψης συμπεριφοράς καταδεικνύει ορισμένες σταθερές. Αφενός, η αξιόπιστη πληροφόρηση προϊόντος και η πρόσβαση αυξάνουν την πιθανότητα επιλογής σημείου ή καναλιού. Αφετέρου, η συνέπεια πολιτικής τιμής - προώθησης σε όλα τα κανάλια επικροτείται μέσω της ανταπόκρισης. Στο πλαίσιο της προώθησης, ο σχεδιασμός ραφιών με βάση πραγματικές συνυπάρχουσες αγορές ενισχύει συμπληρωματικές πωλήσεις (Silverio κ.ά., 2025). Επίσης, τα ετερογενή δεδομένα (POS, πιστότητα, ψηφιακά ίχνη κτλ.) χρειάζονται ενιαίο σχήμα προσέγγισης και τακτική επικαιροποίηση για αξιόπιστη πρόβλεψη.

Οι παραπάνω σταθερές μεταφράζονται σε πρακτικές επιλογές χαρακτηριστικών (πληροφορία προϊόντος, δείκτες πρόσβασης, σταθερότητα τιμής και προσφορών, ζεύγη αγορών κ.ά.) και σε απόδοση προτεραιότητας σε μετρικές που δηλώνουν ανταπόκριση και περιθώριο κέρδους. Η ένταξη τέτοιων μεταβλητών στους *classifiers* αναμένεται να ενισχύσει την ακρίβεια και κυρίως τη χρησιμότητα των προβλέψεων για μάρκετινγκ και επιχειρησιακές λειτουργίες.

3.2 Σύνθεση και επιχειρησιακές επιπτώσεις της προβλεπτικής ανάλυσης

Η πρόβλεψη αποκτά αξία όταν καταλήγει σε συνεπή απόφαση. Το αποτέλεσμα του μοντέλου μετατρέπεται σε πράξη μέσω πίνακα κόστους - οφέλους, επιλογής κατωφλιού

και ορισμού μεγέθους παρτίδας στόχευσης. Η ίδια λογική εφαρμόζεται και στη ζήτηση. Το σφάλμα πρόγνωσης συνδέεται με κόστος έλλειψης ή υπεραποθέματος και καθοδηγεί την παραγγελία.

Η επιλογή κατωφλιού δεν είναι τεχνικό βήμα αλλά οικονομική απόφαση. Για κάθε πιθανότητα ανταπόκρισης υπολογίζεται το αναμενόμενο περιθώριο κέρδους μετά το κόστος επαφής. Το βέλτιστο σημείο είναι εκεί όπου μεγιστοποιείται το καθαρό όφελος. Το κατώφλι μπορεί να διαφοροποιείται ανά τμήμα πελατών, κατηγορία ή κανάλι, όταν τα κόστη και οι ελαστικότητες διαφέρουν. (Πίνακας: «Κατώφλι ανά τμήμα και αναμενόμενο περιθώριο».)

Η βαθμονόμηση πιθανοτήτων είναι κρίσιμη για σωστές αποφάσεις. Μετρά αν τα *scores* ανταποκρίνονται σε συχνότητες (π.χ. αν κάποια εγγραφή ταξινομείται με 0.7, αν όντως το 70% πράγματι ανταποκρίνεται). Ένα ακριβές μοντέλο μπορεί να διακρίνει πολύ καλά το ποιοι να πάρουν την προσφορά πριν από ποιους, αλλά αυτό να συμβαίνει λόγω υπέρ ή υπό αισιόδοξων πιθανοτήτων, οπότε οι αποφάσεις (κατώφλι, προϋπολογισμός ROI κτλ.) να βγαίνουν λάθος. Μέθοδοι όπως *isotonic regression* ή *Platt scaling* συνδέουν αξιόπιστα πρόβλεψη και συχνότητα. Έτσι μειώνονται απευκταία λάθη στόχευσης και βελτιώνεται το πραγματικό κέρδος. Η βαθμονόμηση ελέγχεται περιοδικά, γιατί η συμπεριφορά αλλάζει στον χρονικό ορίζοντα

Η τεκμηρίωση του κέρδους απαιτεί πειραματισμό, δεν αρκεί η πληροφορία ότι ανέβηκαν οι πωλήσεις, καθώς μπορεί να εμπεριέχεται μεροληψία. Τα τεστ 2 ομάδων λήψης και μη της προσφοράς ή ισοδύναμα σχέδια με ομάδες ελέγχου, απομονώνουν το καθαρό αποτέλεσμα από εποχικότητα και τάσεις. Έτσι, πέρα από τον όγκο πωλήσεων, μετρώνται οριακό περιθώριο, *uplift* και ποσοστό εξαργύρωσης. Όπου είναι εφικτό, καταγράφονται και παράπλευρες επιδράσεις σε συμπληρωματικές και ανταγωνιστικές κατηγορίες (*halo* και κανιβαλισμό).

Για να τρέξει ένα μοντέλο σε παραγωγή δεν αρκεί η συγγραφή του κώδικα. Η ενσωμάτωση σε λειτουργία απαιτεί το χτίσιμο ενός σταθερού *pipeline* και τοποθέτηση κανόνων που το κρατούν εύρωστο. Ορίζονται επίσης έλεγχοι ποιότητας δεδομένων, παρακολούθηση μετατόπισης, όρια για επανεκπαίδευση και αρχειοθέτηση εκδόσεων. Η διαδικασία εγκρίνεται από κοινού με εμπορικές και λειτουργικές ομάδες, ώστε η εκτέλεση να υπηρετεί τον επιχειρησιακό σκοπό του μοντέλου.

Η ερμηνευσιμότητα ενισχύει την εμπιστοσύνη και την ορθή χρήση, καθώς υπάρχει ενημέρωση πέραν του τι προτάθηκε και γιατί προτάθηκε. Επιρροή συγκεκριμένων παραγόντων (χαρακτηριστικών), μερικές εξαρτήσεις κατά την τροποποίηση ενός

παράγονται και τοπικές εξηγήσεις για την θέση κατάταξης ενός πελάτη, δείχνουν τον λόγο που προτείνεται μια ενέργεια (Ma, Zhang & Zheng, 2025). Έτσι αποφεύγονται ανεπιθύμητες επιδράσεις, όπως τιμωρία πιστών πελατών με χαμηλές εκπτώσεις ή υπέρμετρη στόχευση σε ευαίσθητες ομάδες. Για να έχει αξία, η ερμηνεία θα πρέπει να συνοδεύεται από σαφείς κανόνες πολιτικής (*fairness checks*), ώστε οι εσωτερικές διεργασίες να μεταφράζονται σε σωστές αποφάσεις (Ajiga κ.ά., 2024).

Η δεοντολογία και η ιδιωτικότητα εντάσσονται από τον σχεδιασμό του συστήματος. Χρησιμοποιούνται μόνο τα απαραίτητα δεδομένα, εφαρμόζεται ανωνυμοποίηση και ελέγχεται αν το μοντέλο κλίνει υπέρ ή κατά κάποιων ομάδων μεροληπτικά. Δεν είναι αρκετή η καλή συνολική ακρίβεια, για αυτό εξετάζεται η ισορροπία απόδοσης μεταξύ ομάδων (π.χ. ίδιο *precision* και *recall* ανάμεσα σε νέους και ηλικιωμένους) και τεκμηριώνονται σαφώς τα κριτήρια στόχευσης (κατώφλια, αποκλεισμοί κτλ.).

Η σύνδεση με το μοντέλο της παρούσας εργασίας είναι άμεση. Οι *classifiers* που υλοποιούνται (LR, DT, RF, Extra Trees, KNN, SVM, Naive Bayes, AdaBoost, Gradient Boosting, NN) τροφοδοτούν πιθανότητες για ανταπόκριση. Αυτές μεταφράζονται σε κατώφλια και παρτίδες στόχων, ικανές να υποβληθούν σε πειραματική μέτρηση του καθαρού οφέλους. Η βαθμονόμηση εφαρμόζεται πριν τον ορισμό κατωφλιού, ώστε το ποσοστό επιλογής να αντιπροσωπεύει πράγματι το επιθυμητό πλήθος. Ο κύκλος του μοντέλου πρόβλεψης, κλείνει με την παρακολούθηση της επίδοσης και τη σύσταση κανόνων επανεκπαίδευσης.

3.2 Κενά στην υπάρχουσα βιβλιογραφία

Η βιβλιογραφία εξετάζει αποσπάσματα των πραγματικών ενεργειών και όχι το σύνολο της εμπειρίας τους καταναλωτή. Σε κάποιες περιπτώσεις αναλύονται μόνο στοιχεία POS ή μόνο πιστότητας, χωρίς συνεπή ενοποίηση δημογραφικών, συμπεριφορικών και συμφραζομένων (π.χ. χρόνος, κατάστημα, κανάλι κτλ.). Λείπει ένα κοινό σχήμα δεδομένων μεταξύ του πελάτη, του χρόνου και της κατηγορίας, που να καθιστά τους δείκτες συγκρίσιμους και τα πειράματα επαναλήψιμα. Δυστυχώς η χρονικότητα συχνά αγνοείται. Αντί χρονικά συνεπών ελέγχων (*rolling* ή *forward chaining*) χρησιμοποιείται τυχαίο k-fold, οδηγώντας σε αισιόδοξες εκτιμήσεις, ενώ σπάνια τεκμηριώνεται σχέδιο για *drift*, *re-training* και *re-calibration*.

Η ανισορροπία κλάσεων αντιμετωπίζεται τεχνικά, όχι οικονομοτεχνικά. Συγκρίνονται SMOTE, *undersampling* και *class weights* με *accuracy* ή ROC–AUC, χωρίς ρητή

σύνδεση με καθαρό κέρδος και σπατάλη επαφών. Υπάρχει επίσης υποτίμηση των πιθανοτήτων. Παραδίδονται *scores* αντί για βαθμονομημένες πιθανότητες, παρότι οι επιχειρησιακές αποφάσεις απαιτούν αξιόπιστη αντιστοίχιση της πιθανότητας με τη συχνότητα. Ακόμη, η πλειονότητα των μελετών παραμένει προβλεπτική, όχι αιτιώδης. Εκτιμάται πιθανότητα ανταπόκρισης χωρίς διάκριση από την αυθόρμητη αγορά. Λείπει συστηματική χρήση *uplift modeling* και ομάδων ελέγχου με καθαρή μέτρηση αυξητικού κέρδους, παρά μόνο σε ελάχιστες περιπτώσεις.

Η εξωτερική εγκυρότητα είναι περιορισμένη. Πολλά ευρήματα βασίζονται σε μία αλυσίδα, λίγα καταστήματα ή στενή χρονική περίοδο. Δεν συναντήθηκε καμία φορά περίπτωση με πολυδιάστατες επαναλήψεις σε άλλες περιοχές, χρονικά παράθυρα, εποχές, κατηγορίες και διαστήματα αβεβαιότητας ανά υποομάδα. Υποτιμάται σε πολλές έρευνες επίσης η διάρκεια των επιδράσεων. Αντ' αυτού, μετράται η άμεση ανταπόκριση, χωρίς *carryover* ή φθορά (*wear-out*) από επαναλαμβανόμενες ενέργειες. Απαιτούνται αναλύσεις με *lags*, μετρικές τύπου *days since last promo* και έλεγχοι που να μετράνε το *promotion fatigue*. Οι συνδυαστικοί παράγοντες ανταγωνισμού δεν ελέγχονται επαρκώς (ανταγωνιστικές τιμές, διαθεσιμότητα, διαφορές καταστήματος κτλ.). Χρειάζονται σταθερά αποτελέσματα και συναφείς μεταβλητές προσφοράς και ζήτησης και εργαλεία για μεροληψία επιλογής.

Η ερμηνευσιμότητα συχνά περιορίζεται σε μια λίστα *feature importances*. Λείπουν τοπικές εξηγήσεις σε επίπεδο απόφασης και μερικές εξαρτήσεις με ελέγχους σταθερότητας ανά υποομάδα. Συνίσταται συνδυασμός τοπικής και διεθνούς σύγκριση με παραδείγματα επιλογής και απόρριψης ενός πελάτη. Ακόμη, η λειτουργική ωρίμανση συχνά αγνοείται. Δεν βρίσκονται μετρήσεις μετατοπίσεις δεδομένων και scores, όρια συναγερμού και κανόνες επανεκπαίδευσης η επαναβαθμονόμησης μετά την παραγωγική ένταξη. Απαιτούνται ορισμένοι δείκτες ελέγχου, συχνότητες ελέγχων και ρητές συνθήκες ενεργοποίησης.

Σημαντική απουσία είναι ότι η συνέπεια ταυτοτήτων δεν τεκμηριώνεται στα *features*. Δεν ορίζονται σταθερά κλειδιά πελάτη και νοικοκυριού, ούτε χειρισμός πολλαπλών ταυτοτήτων για το ίδιο πρόσωπο. Ίδιο πρόβλημα παρουσιάζεται και στο νοικοκυριό, αφού αγνοείται η σύνδεση πελατειακών ταυτοτήτων του ίδιου νοικοκυριού. Δεν εξετάστηκε πουθενά επίσης το *cold-start,* πρόβλεψη για πελάτες με μηδενικά ή ελάχιστα ιστορικά δεδομένα. Χρειάζονται μοντέλα εκκίνησης με ελάχιστα γνωρίσματα και στρατηγικές ομοιότητας για τις πρώτες εβδομάδες για πιο ολοκληρωμένο μοντέλο.

Η ποιότητα χαρακτηριστικών φαίνεται να υποτιμάται. Σπανίζουν πειράματα

προσθαφαίρεσης χαρακτηριστικών, τεχνικών ή αρχιτεκτονικών δομών ανά οικογένεια (π.χ. RFM, seasonality, lags, promos κτλ.) και έλεγχοι σταθερότητας στον χρόνο. Τουλάχιστο θα έπρεπε να γίνεται τεκμηρίωση του κέρδους και του κόστους, όταν αφαιρούνται ή προστίθενται ομάδες και ορισμός ελάχιστης χρήσιμης δέσμης, δηλαδή το μικρότερο σύνολο *features* που δίνει επαρκή απόδοση.

Ακόμη, η δικαιοσύνη και η ιδιωτικότητα σπάνια εντάσσονται ρητά στην αξιολόγηση. Απαιτούνται μετρικές ισότητας ευκαιριών ανά ομάδα, ελαχιστοποίηση δεδομένων, τεκμηριωμένη ανωνυμοποίηση και σαφείς εξουσιοδοτήσεις πρόσβασης, ώστε τα ευρήματα να είναι εφαρμόσιμα σε περιβάλλοντα με κανονιστικές απαιτήσεις. Η μεταφορά στην πράξη παραβλέπει συχνά λειτουργικούς περιορισμούς. Λείπει μοντελοποίηση προϋπολογισμών, χωρητικότητας καναλιών και κανόνων συχνότητας επαφής. Η κατάταξη πρέπει να συνδέεται με βελτιστοποίηση παρτίδας υπό περιορισμούς και επιλογή καναλιού ανά άτομο ή μήνυμα.

Συνοψίζοντας, τα κενά που παραμένουν ουσιαστικά δεν είναι μόνο τεχνικά. Αφορούν την αξιόπιστη μετάφραση πρόβλεψης σε απόφαση αξιοποιώντας σωστά δεδομένα και *pipelines*, χρονικά συνεπή *validation*, βαθμονομημένες πιθανότητες, αιτιώδη τεκμηρίωση αποτελέσματος, οικονομικούς περιορισμούς, ερμηνευσιμότητα, λειτουργική ωρίμανση και κανονιστική συμμόρφωση. Μόνο έτσι η λύση δύναται να είναι υπεύθυνη, επαναλήψιμη και οικονομικά τεκμηριωμένη σε πραγματικό πλαίσιο λιανικού εμπορίου. Αντίθετα, οι περισσότερες μελέτες είναι ακαδημαϊκού προσανατολισμού, απέχοντας παρασάγγας από τις προβλεπτικές υλοποιήσεις υψηλού επιπέδου, τις οποίες και εξασκούν οι μεγάλοι παίχτες του λιανικού εμπορίου στο παγκόσμιο γίγνεσθαι.

ΚΕΦΑΛΑΙΟ 4. ΜΕΘΟΔΟΛΟΓΙΑ

Η υλοποίηση πραγματοποιείται σε Python 3.x με Jupyter Notebook ως κύριο περιβάλλον εργασίας. Το Notebook επιτρέπει διαδοχική εκτέλεση, σχολιασμό δίπλα στον κώδικα και άμεσο έλεγχο ενδιάμεσων αποτελεσμάτων. Έτσι, εξασφαλίζεται ιχνηλασιμότητα του *pipeline* και ακολουθεί τη μεθοδολογική λογική που παρουσιάστηκε στο Κεφ. 2.3, δηλαδή με διακριτά στάδια από τον καθαρισμό έως την αξιολόγηση.

Για τη διαχείριση δεδομένων χρησιμοποιείται ο συνδυασμός βιβλιοθηκών (*modules*) pandas και numpy. Τα αντικείμενα που δημιουργούνται της κλάσης DataFrames διευκολύνουν τυποποιημένους μετασχηματισμούς (επιλογή, φιλτράρισμα, συγχώνευση στηλών), ενώ τα nd-arrays στηρίζουν αριθμητικές πράξεις υψηλής απόδοσης. Η επιλογή αυτού του ζεύγους είναι συμβατή με την ανάγκη ενοποίησης των πηγών σε κοινό σχήμα πελάτη, χρόνου και κατηγορίας, όπως τεκμηριώθηκε στο υποκεφάλαιο 2.3.1.

Η μοντελοποίηση βασίζεται στο εκτενές *module* scikit-learn, που παρέχει ενιαίο API για προεπεξεργασία, εκπαίδευση και αξιολόγηση των *classifiers*. Χρησιμοποιούνται μετασχηματιστές για κωδικοποίηση κατηγορικών και κλιμάκωση χαρακτηριστικών με την κλάση StandardScaler για τη θέσπιση κοινής κλίμακας, σύμφωνα με τις αρχές του 2.3.2. Η επιλογή υπερπαραμέτρων πραγματοποιείται με την GridSearchCV κλάση, ώστε κάθε αλγόριθμος να ελέγχεται σε σαφώς ορισμένο χώρο παραμέτρων, βάσει του καρτεσιανού συνδυασμού που προκύπτει από τις τιμές τους. Η γεννήτρια τυχαίων αριθμών ορίζεται ίδια για όλα τα στάδια, με σταθερούς σπόρους (seeds) τυχαιοποίησης (random_state) για αναπαραγωγιμότητα, σε συναρτήσεις που αφορούν το διαμοιρασμό σε *folds*, τα *bootstrap* δείγματα κτλ, δηλαδή η τιμή του random_state μένει ίδια. Η σχεδίαση αυτή άλλωστε, επιβάλλει το ίδιο πρωτόκολλο σε όλα τα μοντέλα, όπως ζητείται από την άρτια πρακτική σύγκρισης στο 2.3.3. Όσο για την απεικόνιση, αξιοποιούνται τα *modules* οπτικοποίησης δεδομένων matplotlib και seaborn για συνοπτικές κατανομές και σχέσεις, όπου χρειάζεται. Τα διαγνωστικά ταξινόμησης (π.χ. confusion matrix) αποδίδονται μέσω της Yellowbrick βιβλιοθήκης πάνω στον ήδη εκπαιδευμένο εκτιμητή, ώστε η οπτικοποίηση να μην παρεμβαίνει στη διαδικασία εκμάθησης.

Η αντιμετώπιση ανισορροπίας κλάσεων γίνεται με τις δυνατότητες του scikit-learn (π.χ.

class_weight = «balanced» όπου υποστηρίζεται) και όπου απαιτείται πειραματικά, με τεχνικές επαναδειγματοληψίας της βιβλιοθήκης imbalanced-learn (π.χ. με κλάσεις RandomOverSampler ή SMOTE) ενταγμένες εντός της CV, όπως προτείνεται ρητά στο 2.3.2 για αποφυγή διαρροής πληροφορίας. Η θέση των βημάτων αυτών στο *pipeline* τηρεί την αρχή που θέλει το fit αποκλειστικά σε training set (*fold*), transform σε training set και στη συνέχεια στο validation set, ώστε οι εκτιμήσεις να είναι αμερόληπτες, αφού το μοντέλο δεν μαθαίνει κάτι από τα δεδομένα που δεν έχει δει ακόμα.

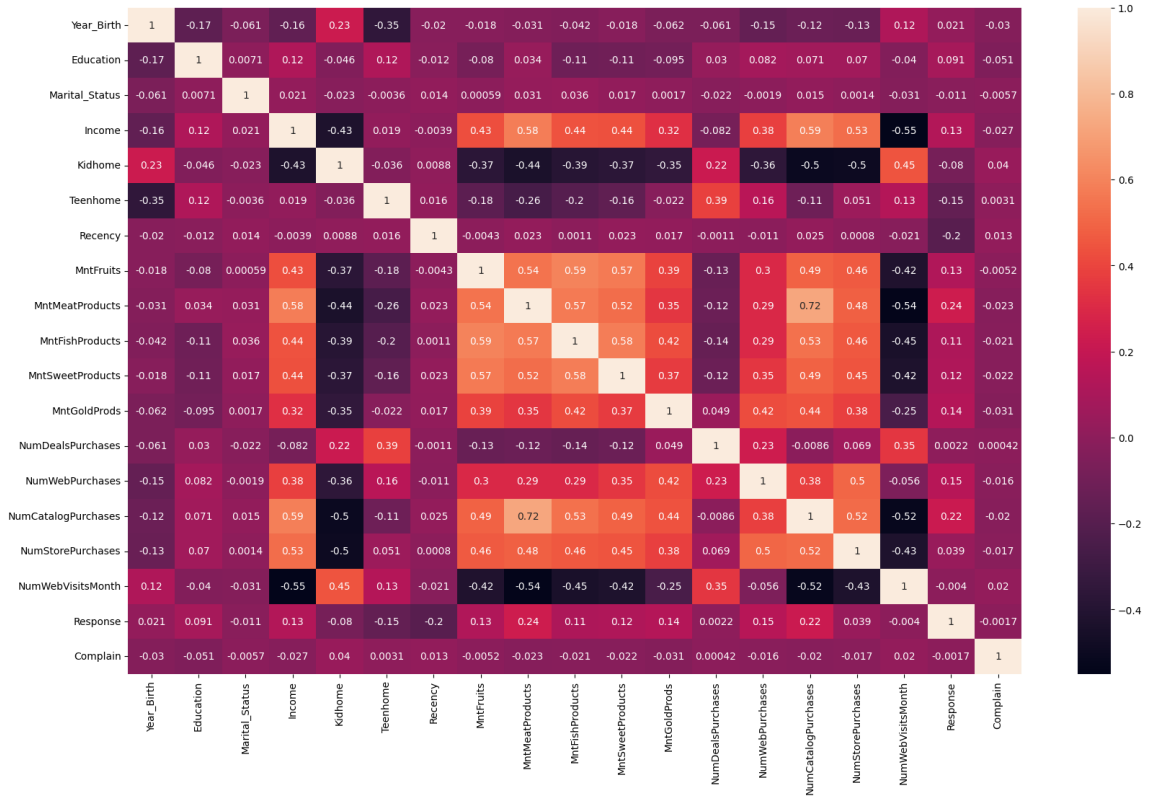
Για νευρωνικά μοντέλα αξιοποιείται ένα νευρωνικό δίκτυο προωθητικής ροής. Η κλάση MLPClassifier του scikit-learn χρησιμοποιείται για ταξινόμηση και εφαρμόζει το πολυεπίπεδο νευρωνικό δίκτυο με πλήρως συνδεδεμένα στρώματα μεταξύ των νευρώνων και με ρυθμίσεις βασισμένες σε μικρές, πλήρως συνδεδεμένες αρχιτεκτονικές ανά στρώμα. Η επιλογή αυτή κρατά ενιαίο API με τους υπόλοιπους *classifiers* και επιτρέπει ενιαία προσέγγιση για tuning, ενώ συμμορφώνεται με τις απαιτήσεις κλιμάκωσης και τακτοποίησης που αναλύθηκαν στο 2.4.3.

Η αξιολόγηση στηρίζεται στο υποσύστημα μετρικών του scikit-learn (με συναρτήσεις όπως precision, recall, F1, accuracy κ.ά.). Τα αποτελέσματα παρουσιάζονται συγκεντρωτικά στο 4ο κεφάλαιο, με σαφή αναφορά ότι προκύπτουν σε προκαθορισμένο *threshold* απόφασης, σύμφωνα με τη βιβλιογραφία που παρατέθηκε στο κεφάλαιο 2.3 περί της σημασίας των κατωφλιών και της βαθμονόμησης πιθανοτήτων.

4.1 Σχεδιασμός και δεδομένα έρευνας

Η μελέτη ακολουθεί ποσοτική προσέγγιση ταξινόμησης, από υπάρχοντα αριθμητικά δεδομένα πελατών με ετικέτα (εποπτευόμενη μάθηση). Στόχος είναι η εκτίμηση της πιθανότητας ανταπόκρισης σε προωθητική ενέργεια (Response = 1 για αποδοχή και 0 για απόρριψη) που αφορά αποδοχή προσφοράς χρυσής κάρτας μέλους για 499$ αντί 999$, με ανταπόδοση οριζόντια έκπτωση 20% για όλες τις αγορές. Ο ορισμός του στόχου συνάδει με το θεωρητικό πλαίσιο του 2.3, όπου οι αποφάσεις μάρκετινγκ μεταφράζονται σε πιθανότητες και κατώφλια κόστους και οφέλους. Η αξιολόγηση βασίζεται σε μετρικές που προτιμώνται για σπάνια θετική τάξη (precision, recall, F1), όπως επίσης τεκμηριώνεται στο 2.3.

Μονάδα ανάλυσης είναι το άτομο ή νοικοκυριό, με εγγραφές που συνδυάζουν δημογραφικά στοιχεία και αγοραστική συμπεριφορά. Η χρονολογική αναφορά των μεταβλητών αντιμετωπίζεται ως στιγμιότυπο κατά την περίοδο συλλογής, οπότε δεν εκτιμώνται δυναμικά μοντέλα σειράς χρόνου. Η διάκριση μεταξύ του παράθυρου παρατήρησης χαρακτηριστικών, το οποίο λογίζει χαρακτηριστικά μόνο από το παρελθόν, έναντι παραθύρου πρόβλεψης που αφορά τον στόχο, ακολουθεί τις αρχές του 2.3.1, ώστε να αποφεύγεται διαρροή πληροφορίας στη μοντελοποίηση.



Εικόνα 1. Συσχέτιση χαρακτηριστικών dataset

Το σύνολο δεδομένων περιλαμβάνει δημογραφικά γνωρίσματα (ηλικιακή ένδειξη, εισόδημα, σύνθεση νοικοκυριού κτλ.), δείκτες προτιμήσεων και δαπανών ανά κατηγορία (π.χ. κρασιά, κρέας, ψάρια, γλυκά), εμπειρικούς δείκτες συμπεριφοράς (πρόσφατη αλληλεπίδραση, κανάλι εξυπηρέτησης κτλ.) και τη δυαδική ετικέτα Response. Η επιλογή αυτών των αξόνων αντανακλά τα ευρήματα του 2.2 για τους καθοριστικούς παράγοντες στο σούπερ μάρκετ, όπου ως *feature* αναγνωρίζεται η συσχέτιση από δημογραφικά χαρακτηριστικά, προωθήσεις και συνήθειες καλαθιού.

Ο σχεδιασμός εκτίμησης θέτει ως βασική υπόθεση ότι οι διαθέσιμες μεταβλητές φέρουν επαρκή πληροφορία για τη διάκριση ανταποκριθέντων και μη. Η πιθανή

ανισορροπία κλάσεων αντιμετωπίζεται αργότερα με κατάλληλες μετρικές και σταθμίσεις ή επαναδειγματοληψία, όπου χρειαστεί. Η ερμηνευσιμότητα παραμένει ζητούμενο, καθώς οι γραμμικοί και δενδροειδείς ταξινομητές αναδεικνύουν σημασίες χαρακτηριστικών και κανόνες, χρήσιμους για την τεκμηρίωση των συμπερασμάτων του 2.1 και 2.2. Η δεοντολογία και η ελαχιστοποίηση δεδομένων διέπουν τον χειρισμό των χαρακτηριστικών, όπως υπαγορεύεται στο 2.3.4. Χρησιμοποιούνται μόνο τα απολύτως αναγκαία πεδία για την πρόβλεψη και τεκμηριώνονται οι ορισμοί τους.

Ακολούθως παρατίθενται συνοπτικά οι ομάδες πεδίων που τροφοδοτούν τη μοντελοποίηση. Ομαδοποίηση και περιγραφές ακολουθούν τη λογική του 2.2 (παράγοντες και προφίλ) και του 2.3 (*pipeline features*). Οι τύποι μεταβλητών που αναφέρονται είναι ενδεικτικοί και χρησιμοποιούνται για να περιγράψουν τη μορφή των πεδίων. Αυτοί είναι *αριθμητικοί* (συνεχή ή διακριτά μεγέθη όπως το εισόδημα), *κατηγορικοί* (ονόματα ή επίπεδα κατηγορίας όπως η οικογενειακή κατάσταση), *δυαδικοί* (τιμές 0 και 1 όπως το παράπονο) και *ημερομηνίες* (χρονικά πεδία, π.χ. Dt_Customer).

Δημογραφικά - Σύνθεση νοικοκυριού

- Year_Birth (αριθμητικός). Έτος γέννησης.
- Education (κατηγορηματικός). Εκπαιδευτικό επίπεδο.
- Marital_Status (κατηγορηματικός). Οικογενειακή κατάσταση.
- Income (αριθμητικός). Ετήσιο εισόδημα νοικοκυριού.
- Kidhome (αριθμητικός). Παιδιά στο νοικοκυριό.
- Teenhome (αριθμητικός). Έφηβοι στο νοικοκυριό.
- Country (κατηγορηματικός). Χώρα κατοικίας.

Χρονική συσχέτιση με το κατάστημα

- Dt_Customer (ημερομηνία). Ημερομηνία ένταξης/πρώτης καταγραφής.
- Recency (αριθμητικός). Ημέρες από την τελευταία αλληλεπίδραση (χαμηλότερο = πιο πρόσφατα).

Δαπάνες ανά κατηγορία προϊόντων (μονάδες σε νόμισμα του dataset, οριζόμενες ως αθροιστικές ποσότητες περιόδου)

- MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds (αριθμητικοί). Σωρευτικές δαπάνες ανά κατηγορία.

Συμπεριφορά καναλιών και επισκέψεων

- NumWebPurchases (αριθμητικός). Αγορές μέσω ίντερνετ.
- NumCatalogPurchases (αριθμητικός). Αγορές μέσω καταλόγου.
- NumStorePurchases (αριθμητικός). Αγορές στο κατάστημα.
- NumWebVisitsMonth (αριθμητικός). Επισκέψεις στον ιστό τον τελευταίο μήνα.

Προωθήσεις και Ανταπόκριση

- NumDealsPurchases (αριθμητικός). Αγορές με χρήση προσφορών και εκπτώσεων.
- AcceptedCmp1 - AcceptedCmp5 (δυαδικός). Αποδοχή προηγούμενων καμπανιών 1 - 5.
- Response (δυαδικός). Στήλη στόχος, αποδοχή της τρέχουσας καμπάνιας (1) ή μη (0).

Λοιπά λειτουργικά πεδία

- Complain (δυαδικός). Καταγεγραμμένο παράπονο στο πρόσφατο διάστημα.
- Z_CostContact, Z_Revenue (αριθμητικός). Σταθερές επιχειρησιακές παραδοχές, τυπικό κόστος ανά επαφή και τυπικό έσοδο ανά αποδοχή, ωστόσο δεν χρησιμοποιούνται για πρόβλεψη γιατί εισάγουν μεροληψία.

Ως προς την μονάδα ανάλυσης, η εγγραφή αντιστοιχεί σε πελάτη ή νοικοκυριό σε συγκεκριμένο χρονικό στιγμιότυπο. Δεν εκτιμάται δυναμικό πρότυπο σειράς χρόνου. Η Recency λειτουργεί ως δείκτης χρονικής εγγύτητας της τελευταίας αλληλεπίδρασης

(σε μικρή τιμή πρόσφατη επαφή, σε μεγάλη μεσολάβησε αρκετός χρόνος) και οι Mnt* ως αθροιστική συμπεριφορά (σύνολο δαπάνης ανά κατηγορία). Αυτό επιτρέπει στα μοντέλα να μάθουν σταθερά μοτίβα ανταπόκρισης χωρίς να απαιτείται ακολουθία γεγονότων. Στο επόμενο κεφάλαιο διευκρινίζεται αν μετατρέπονται οι συσσωρευμένες δαπάνες σε μερίδια, ώστε να ελεγχθεί η ετερογένεια καλαθιού, βάσει της θεωρίας που αναπτύχθηκε στο 2.2.

Αναφορικά με τον στόχο και την σπανιότητα, το Response είναι δυαδικός στόχος με δυνητικά ανισόρροπες κλάσεις. Αυτό επηρεάζει δύο κρίσιμα σημεία. Την επιλογή μετρικών αξιολόγησης που εστιάζει σε precision, recall, F1 και PRC–AUC και τη ρύθμιση των *thresholds*. Τα κατηγορικά πεδία (Education, Marital_Status, Country) δεν τροφοδοτούνται αυτούσια και απαιτούν κωδικοποίηση (one-hot encoding), ώστε να είναι συμβατά με γραμμικούς και δενδροειδής ταξινομητές. Στόχος είναι η αποφυγή τεχνητής τάξης (ψευδής ιεραρχία σε κατηγορικές αποδόσεις 0, 1, 2 κτλ) σε μη διατεταγμένες μεταβλητές. Οι κωδικοποιήσεις εφαρμόζονται στο στάδιο προεπεξεργασίας, με μέριμνα αποφυγής διαρροής.

Οι μεταβλητές έχουν ετερογενή μεγέθη (π.χ. Income σε νόμισμα, NumWebVisitsMonth ως πλήθος). Η κλιμάκωση με StandardScaler είναι απαραίτητη για ευαίσθητα μοντέλα στην κλίμακα όπως LR, SVM, KNN και NN, επειδή βασίζονται σε μεγέθη βαρών ή αποστάσεις. Στο επόμενο κεφάλαιο δηλώνεται ρητά η εφαρμογή της κλιμάκωσης μετά τον χωρισμό σε *folds* (fit στο train, transform σε train ή validation). Η πρακτική υλοποιήθηκε ακολουθώντας πλήρως τις απαιτήσεις των *classifier*, όπως αυτές ορίζονται στο 2.4.

Πεδία όπως Z_CostContact, Z_Revenue λειτουργούν ως σταθερές και δεν μεταφέρουν συμπεριφορική πληροφορία. Αποκλείονται από το μοντέλο για να αποφευχθεί τεχνητή ενίσχυση ακρίβειας ή έμμεση κωδικοποίηση του στόχου, χωρίς επιχειρησιακή γενίκευση. Ομοίως, πεδία που αποτελούν μεταγενέστερες συνέπειες της καμπάνιας δεν πρέπει να περιλαμβάνονται πριν οριστεί καθαρά το *prediction window*, καθώς υπεισέρχεται μεροληψία στο μοντέλο, όπως ορίζει το θεωρητικό υπόβαθρο στο 2.3.1.

4.2 Προεπεξεργασία και Χαρακτηριστικά

Η προεπεξεργασία ακολουθεί διαδοχικά βήματα, ώστε τα δεδομένα να είναι καθαρά, συνεπή και έτοιμα για εκπαίδευση *classifiers*. Η ροή ξεκινά με ελέγχους πληρότητας και συμφωνίας τύπων, συνεχίζει με κωδικοποίηση κατηγορικών και καταλήγει σε κλιμάκωση αριθμητικών μεταβλητών, πριν τη δημιουργία των συνόλων εκπαίδευσης και δοκιμής. Η επιλογή και η σειρά των βημάτων εκπροσωπούν την ανάγκη για σταθερότητα και δυνατότητα γενίκευσης.

Η μοναδική στήλη με ελλείπουσες τιμές είναι η Income (24/2240 εγγραφές). Οι ελλείψεις συμπληρώνονται με τον μέσο όρο της στήλης. Η συμπλήρωση πραγματοποιείται στο πλήρες σύνολο πριν από οποιονδήποτε χωρισμό, ώστε να διατηρηθεί το μέγεθος δείγματος και η συνάφεια κατανομών για τα επόμενα βήματα. Η επιλογή καταγράφεται ρητά, διότι στο 2.3 έχει ήδη επισημανθεί ότι η χρονική και δειγματοληπτική συνέπεια είναι κρίσιμη για την αποτίμηση.

```
Income
False    2216
True       24
Name: count, dtype: int64
```

Εικόνα 2. Έλλειψη 24 εγγραφών Income

Τα κατηγορικά πεδία Education και Marital_Status μετασχηματίζονται σε αριθμητική μορφή με LabelEncoder. Ο μετασχηματισμός εφαρμόζεται στο ενοποιημένο *dataframe*, ακολουθώντας τις πρακτικές που ταιριάζουν σε δενδροειδή και συνθετικά μοντέλα, όπου οι διασπάσεις δεν προϋποθέτουν μετρική ισοδυναμία μεταξύ κατηγοριών. Ο χειρισμός αυτός διατηρεί τη λιτή αναπαράσταση των μεταβλητών και επιτρέπει σύγκριση μοντέλων με ομοιόμορφη είσοδο χαρακτηριστικών.

Η κλιμάκωση των αριθμητικών χαρακτηριστικών γίνεται με StandardScaler. Ο scaler εφαρμόζεται με fit_transform στο πίνακα χαρακτηριστικών X πριν την εκπαίδευση των μοντέλων που είναι ευαίσθητα στην κλίμακα (π.χ. LR, SVM, KNN, NN), όπως τεκμηριώνεται στο 2.4. Η επιλογή αυτή εξασφαλίζει συγκρισιμότητα μεγεθών και σταθερότητα των αλγορίθμων που βασίζονται σε αποστάσεις ή μεγέθη βαρών, ενώ αφήνει ανεπηρέαστα τα *ensembles* όπου η κλίμακα δεν είναι προϋπόθεση.

Πεδία χωρίς προγνωστική αξία ή με καθαρά τεχνικό ρόλο αφαιρούνται πριν τον ορισμό του συνόλου χαρακτηριστικών. Συγκεκριμένα, τα Id, MntWines και Dt_Customer απομακρύνονται, ώστε το τελικό X να περιέχει μόνο ουσιώδεις μεταβλητές και το y να ορίζεται ως Response. Με αυτόν τον ορισμό, η είσοδος προς τα μοντέλα παραμένει συμπαγής και επικεντρωμένη σε δημογραφικά, συμπεριφορικά και δαπανηρά σήματα, κατά το θεωρητικό υπόβαθρου του 2.3.2.

Η ανισορροπία κλάσεων αντιμετωπίζεται με SMOTE μέσω sm.fit_resample(x, y), ώστε η θετική τάξη να ενισχυθεί πριν από τον τελικό χωρισμό σε εκπαίδευση και δοκιμή. Η πρακτική αυτή διασφαλίζει ότι οι ταξινομητές θα εμπλακούν με επαρκή παραδείγματα της σπάνιας κλάσης κατά τη μάθηση, στοιχείο που συνδέεται άμεσα με τους στόχους ανάκτησης (recall) και ισορροπίας σφάλματος που θα αναλυθούν στα επόμενα κεφάλαια. Μετά το oversampling, εκτελείται ο διαχωρισμός train/test με test_size = 0.20 και σταθερό random_state. Η ρύθμιση του σπόρου στηρίζει την αναπαραγωγιμότητα, ενώ η αναλογία 80/20 διασφαλίζει επαρκές δείγμα εκπαίδευσης και ένα καθαρό δείγμα ελέγχου. Η ίδια είσοδος τροφοδοτεί όλους τους ταξινομητές που θα συγκριθούν, ώστε τα αποτελέσματα να είναι άμεσα αντιπαραβαλλόμενα.

Η τελική διάταξη χαρακτηριστικών περιλαμβάνει τα δημογραφικά (Year_Birth, Education, Marital_Status, Income, Kidhome, Teenhome), τα συμπεριφορικά (NumWebVisitsMonth, Recency, δείκτες αγορών/επαφών) και δαπάνες ανά κατηγορία εξ' αυτών (Mnt πεδία, πλην MntWines που αφαιρέθηκε). Επίσης διατηρούνται ιστορικά χαρακτηριστικα ανταπόκρισης (AcceptedCmp), καθώς και δείκτες σχέσης όπως το Complain. Με αυτό το προφίλ, τα μοντέλα λαμβάνουν συμπυκνωμένη πληροφορία για τον πελάτη σε επίπεδο νοικοκυριού. Με τα παραπάνω βήματα, η ροή προεπεξεργασίας ολοκληρώνεται με σαφή ορισμό X και y, κωδικοποιημένα κατηγορικά, κλιμακωμένα αριθμητικά και με μετριασμένη ανισορροπία.

4.3 Διαδικασία tuning & validation

Στόχος του πρωτοκόλλου είναι η αμερόληπτη εκτίμηση γενίκευσης και συνεπής επιλογή υπερπαραμέτρων, σε αρμονία με το επιχειρησιακό πρόβλημα ταξινόμησης. Υιοθετείται διακριτός ρόλος για εκπαίδευση, επικύρωση και τελικό έλεγχο, με σαφή αναπαραγωγιμότητα (σταθεροί σπόροι). Ο διαχωρισμός δεδομένων υλοποιείται με train_test_split και σταθερό random_state=27. Η ρύθμιση υπερπαραμέτρων γίνεται

εντός του εκπαιδευτικού μέρους μέσω GridSearchCV, χωρίς καμία προσαρμογή στο test κατά το tuning. Στην εργασία, το GridSearchCV χρησιμοποιείται για συγκεκριμένα μοντέλα με καθορισμένα πλέγματα τιμών και προκαθορισμένο πλήθος πτυχών (cv). Όπου δεν εφαρμόστηκε GridSearchCV, ο ταξινομητής εκπαιδεύτηκε με ρητά ορισμένες ή προεπιλεγμένες ρυθμίσεις και αναφέρεται στο επόμενο κεφάλαιο.

Η επικύρωση βασίζεται σε k-fold CV ανά μοντέλο ως εξής:

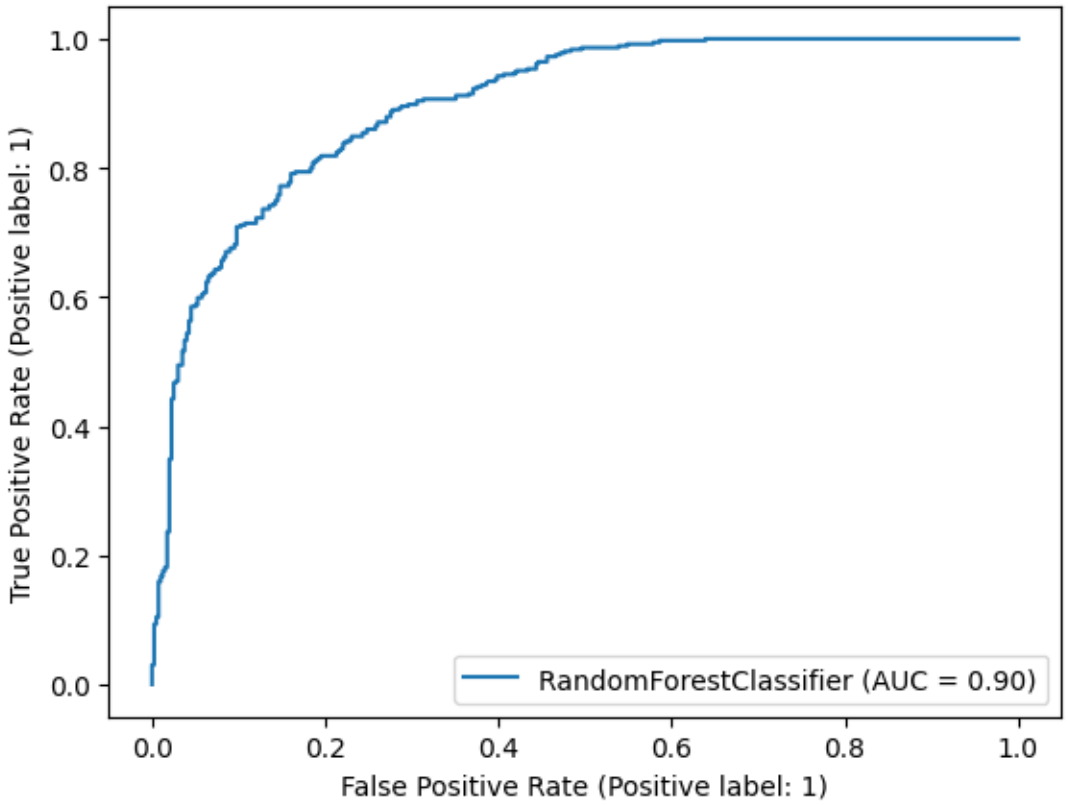
- Random Forest: cv=5 με πλέγμα υπερπαραμέτρων και κλήση GridSearchCV.
- Extra Trees: cv=5 με πλέγμα και GridSearchCV.
- KNN: cv=2 με πλέγμα n_neighbors και GridSearchCV.
- SVM: cv=2 με πλέγμα C/kernel/degree/gamma και GridSearchCV.
- AdaBoost: cv=5 με πλέγμα n_estimators/learning_rate και GridSearchCV.
- Gradient Boosting: cv=5 με πλέγμα n_estimators/learning_rate και GridSearchCV.

Όπου δεν τεκμηριώνεται GridSearchCV (π.χ. Logistic Regression, Decision Tree, Naive Bayes), η εκτέλεση γίνεται με ρητές σταθερές ρυθμίσεις και χωρίς σάρωση χώρου υπερπαραμέτρων, με συνοπτική παρουσίαση να ακολουθεί παρακάτω και τις αντίστοιχες αναφορές κελιών εκτέλεσης.

Το scoring στα παραπάνω GridSearchCV δεν ορίζεται ρητά, άρα ακολουθεί την προεπιλογή του estimator (ισοδύναμο της accuracy) για την επιλογή των βέλτιστων υπερπαραμέτρων. Οι μετρικές σύγκρισης μεταξύ μοντέλων (Precision, Recall, F1, Accuracy και καμπύλη ROC) υπολογίζονται μετά την επιλογή μοντέλου, χωρίς σάρωση κατωφλιών (το thresholding με οικονομικούς όρους τεκμηριώνεται μεθοδολογικά αλλά δεν υλοποιείται στον κώδικα). Όλες οι τυχαίες συνιστώσες φέρουν σταθερούς σπόρους, με διακριτή δήλωση (π.χ. random_state στους δενδροειδείς αλγορίθμους).

Για διαφάνεια και αναπαραγωγιμότητα, ο πίνακας που ακολουθεί συνοψίζει τους πραγματικούς χώρους υπερπαραμέτρων που σαρώθηκαν ανά μοντέλο. Οι τιμές αποτυπώνουν τις λίστες που πέρασαν στα param_grid των κλήσεων GridSearchCV,

μαζί με το πλήθος των πτυχών (cv). Όπου δεν χρησιμοποιήθηκε GridSearchCV, σημειώνεται *default* και η παρουσίαση τους συνεχίζεται στο κεφάλαιο 4.4 σε επίπεδο εκτέλεσης.



Εικόνα 3. Random Forest ROC AUC

Το εύρος των πλεγμάτων δείχνει πρακτική στόχευση σε λίγες, ασφαλείς ρυθμίσεις αντί για εξαντλητική σάρωση. Στο RF η επέκταση βάθους έως 11 με min_samples_split=2 ευνοεί ισχυρό fit, ενώ με cv=5 μετριάζεται ο κίνδυνος υπερπροσαρμογής, αλλά σε συνδυασμό με προηγηθέν leakage (scaling/SMOTE πριν το split) το βέλτιστο μπορεί να είναι αισιόδοξο. Το Extra Trees κρατά ίδια λογική με λιγότερα κουμπιά, κάτι που συνήθως δίνει σταθερότητα (λόγω τυχαιοποίησης σπασιμάτων) με μικρότερο tuning overhead. Στο KNN το πλέγμα εστιάζει σε πολύ μικρά k. Αυτό αυξάνει τη διακύμανση και εξαρτάται έντονα από το scale, άρα η απόδοση θα αντανακλά κυρίως τη θέση του scaler στη ροή. Στον SVM η διερεύνηση όλων των *kernels* με μικρό cv=2 δίνει γρήγορη εικόνα τοπολογίας, όχι όμως σταθερή εκτίμηση. Ειδικά το poly με degree έως 5 μπορεί να επηρεαστεί από θόρυβο αν η κλιμάκωση δεν είναι fold-aware. Στο AdaBoost οι πολύ υψηλές learning_rate (έως 2.5) είναι ασυνήθιστες, όταν όμως το base learner είναι ρηχό δέντρο, η μεγάλη κλίση μπορεί να δώσει γρήγορο κέρδος με τίμημα την ευαισθησία

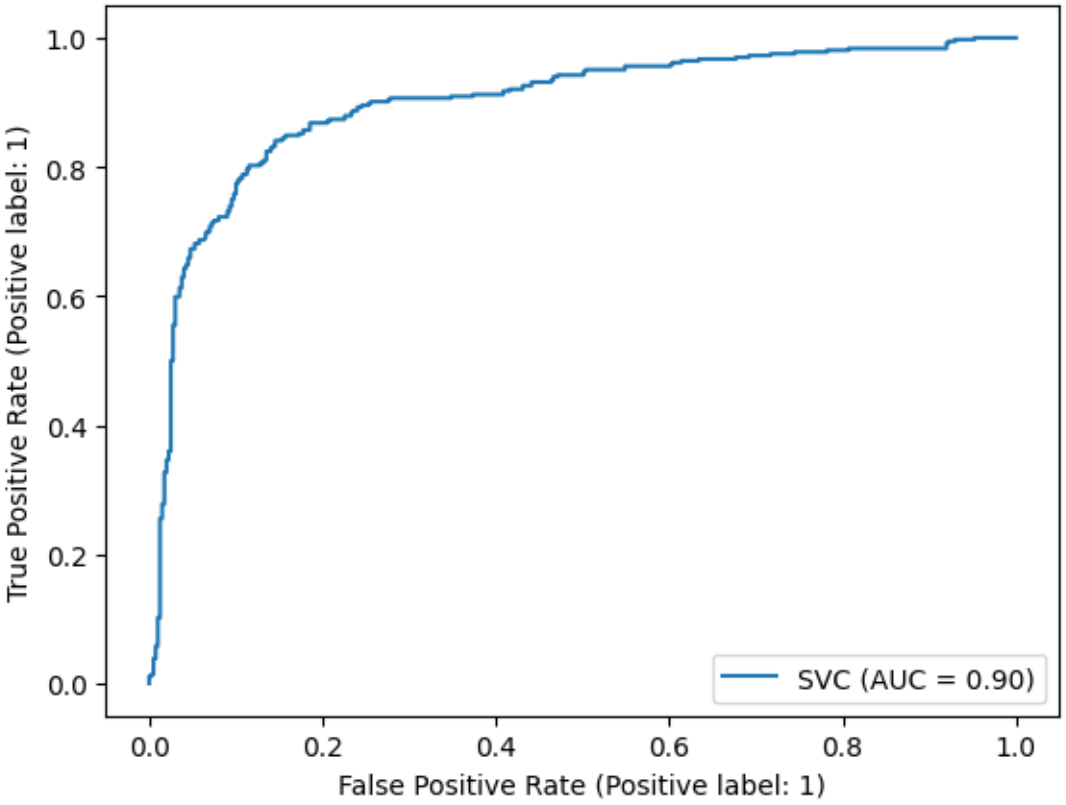
στον θόρυβο. Το Gradient Boosting δοκιμάζει ευρύ φάσμα learning_rate (0.005–0.5) με σταθερό n_estimators=300. Πρακτικά αυτό ισοδυναμεί με δοκιμή διαφορετικής έντασης τακτοποίησης, χωρίς να εξερευνώνται depth και subsample, άρα μένει ανεκμετάλλευτο μέρος του bias–variance trade-off.

Μοντέλο	Υπερπαράμετροι που σαρώθηκαν	CV
Random Forest	n_estimators: [100], criterion: ['entropy','gini'], min_samples_split: [2,3,4,5,6,7], max_depth: [3,4,5,6,7,9,11]	5
Extra Trees	n_estimators: [100], criterion: ['entropy','gini']	5
KNN	n_neighbors: [1,2,3,4,5,6,7,8,9] *(metric='minkowski', p=2)*	2
SVM	kernel: ['linear','rbf','poly','sigmoid'], C: [3,4,5], degree: [2,3,4,5], gamma: ['auto','scale']	2
AdaBoost	n_estimators: [500], learning_rate: [2.5,2.0,1.9,1.7,0.5,0.4]	5
Gradient Boosting	n_estimators: [300], learning_rate: [0.5,0.3,0.1,0.09,0.07,0.05,0.02,0.01,0.005]	5
Logistic Regression	— *(default ρυθμίσεις στην εκτέλεση/αξιολόγηση)*	—
Decision Tree	— *(default ρυθμίσεις στην εκτέλεση/αξιολόγηση)*	—
Naive Bayes	— *(default ρυθμίσεις στην εκτέλεση/αξιολόγηση)*	—

Εικόνα 4. Χώροι υπερπαραμέτρων ανά classifier

Η απουσία GridSearchCV σε LR, DT, NB τα τοποθετεί συνειδητά ως ενδεικτικούς ταξινομητές βάσης (baselines). Η LR δίνει ερμηνεύσιμη γραμμή εκκίνησης, το DT παρέχει διαφανή κανόνες και ο NB γρήγορη αναφορά σε αραιές και απλές δομές. Αυτή η ιεράρχηση κρίνεται επαρκής γιατί πρώτα ορίζεται ρεαλιστικό σημείο αναφοράς και έπειτα ελέγχεται αν τα ensembles όντως προσφέρουν ουσιώδες κέρδος. Αν επεκταθεί το πρωτόκολλο, δύο στοχευμένες βελτιώσεις θα αυξήσουν αξιοπιστία χωρίς να εκτροχιάσουν τον χρόνο. Πρώτον η εναρμόνιση scoring με F1 ή PRC–AUC, όπου η θετική κλάση είναι σπάνια, ώστε τα βέλτιστα να ευθυγραμμίζονται με τον στόχο, και

δεύτερον cv ≥ 5 για SVM και KNN, γιατί είναι τα πιο ευαίσθητα σε scale και γειτνίαση αντίστοιχα, ενώ τείνουν να παρουσιάζουν υψηλή διακύμανση με μικρό cv. Με αυτά, ο πίνακας παύει να είναι απλή λίστα ρυθμίσεων και γίνεται τεκμηριωμένη γέφυρα από το tuning στις επιχειρησιακές μετρικές.



Εικόνα 5. SVM ROC AUC

4.4 Classifiers – Εκτελεστικό Πλαίσιο

Σκοπός της ενότητας είναι να παρουσιαστεί τι ακριβώς εκτελέστηκε για κάθε ταξινομητή, με τις κρίσιμες επιλογές ροής και τις αντίστοιχες παραπομπές. Η αξιολόγηση επιδόσεων και η σύγκριση μοντέλων πραγματοποιείται στο κεφάλαιο 5. Τα χαρακτηριστικά και ο στόχος ορίζονται αφού αφαιρεθούν τεχνικά και πεδία δίχως προβλεπτική αξία (π.χ. Id, Dt_Customer) και οριστεί το Response. Έχει προηγηθεί κωδικοποίηση κατηγορικών (Education, Marital_Status) με LabelEncoder, τυπική κλιμάκωση αριθμητικών με StandardScaler και υπερδειγματοληψία μειονοτικής κλάσης με SMOTE. Συνεπώς, τα παρακάτω μοντέλα στηρίζονται σε αυτή τη ροή εκτέλεσης.

Logistic Regression

Γραμμική *baseline* με ερμηνεύσιμους συντελεστές. Χρησιμοποιείται για να δοθεί σταθερό σημείο αναφοράς και πιθανότητες αλλά απαιτεί κλιμάκωση. Οι κρίσιμες υπερπαράμετροι είναι C, penalty, max_iter. Η υλοποίηση περιλαμβάνει ορισμό, εκπαίδευση και πρόβλεψη, με παραγωγή καμπύλης ROC. Δεν εφαρμόζεται GridSearchCV στη συγκεκριμένη ακολουθία και λειτουργεί ως ταχεία γραμμή βάσης.

Random Forest

Ensemble δέντρων με δειγματοληψία παρατηρήσεων και χαρακτηριστικών, κατάλληλο για μη γραμμικότητες και αλληλεπιδράσεις. Ως κρίσιμες υπερπαράμετροι θεωρούνται n_estimators, max_depth, min_samples_split ή leaf, criterion. Ο αρχικός χώρος διερεύνησης και η αναζήτηση υπερπαραμέτρων υλοποιούνται με GridSearchCV. Έπειτα κατασκευάζεται τελικό RF με τις βέλτιστες ρυθμίσεις και εκτελείται η πρόβλεψη στο test-set. Το RF αξιοποιείται επίσης για ενδείξεις σημαντικότητας χαρακτηριστικών.

Extra Trees (Extremely Randomized Trees)

Παραλλαγή δασών με εντονότερη τυχαιοποίηση στα κατώφλια διάσπασης, χρήσιμη για γρήγορο fit και ανθεκτικότητα σε θόρυβο. Κρίσιμες ρυθμίσεις είναι n_estimators, max_depth, criterion, max_features. Η διερεύνηση υπερπαραμέτρων γίνεται με GridSearchCV και ακολουθεί εκπαίδευση τελικού μοντέλου με τις επιλεγμένες τιμές. Χρησιμοποιείται συμπληρωματικά προς RF όταν ζητείται ταχύτητα και σταθερότητα στην τυχαιοποίηση.

K-Nearest Neighbors (KNN)

Ο KNN βασίζεται στη γειτνίαση σε κλιμακωμένο χώρο χαρακτηριστικών, γι' αυτό στηρίζεται στο προηγηθέν StandardScaler. Οι κρίσιμες επιλογές είναι ο αριθμός γειτόνων (n_neighbors), η μετρική απόστασης και τα weights. Η υλοποίηση περιλαμβάνει σάρωση γειτόνων μέσω GridSearchCV και στη συνέχεια εκπαίδευση και πρόβλεψη του τελικού KNN. Χρησιμοποιείται ως απλός, μη παραμετρικός συγκριτικός άξονας απέναντι στα δενδροειδή (*ensembles*).

Support Vector Machine (SVM)

Το SVM αξιοποιεί μέγιστο περιθώριο και kernels για μη γραμμικό διαχωρισμό, με απαραίτητη κλιμάκωση εισόδων. Ορίζεται χώρος υπερπαραμέτρων για C, kernel (linear/rbf/poly), degree (για poly) και gamma (auto/scale). Η διερεύνηση και επιλογή ρυθμίσεων γίνεται με GridSearchCV, κατόπιν εκπαιδεύεται το επιλεγμένο SVM και παράγονται προβλέψεις στο test-set. Το SVM εντάσσεται ως ισχυρή μη γραμμική εναλλακτική απέναντι στα *ensembles*.

Naive Bayes (NB)

Ο NB λειτουργεί ως ταχεία γραμμή βάσης με ελάχιστη ρύθμιση. Στο notebook εκτελείται η εκπαίδευση και η πρόβλεψη της κατάλληλης παραλλαγής (Gaussian, Multinomial και Bernoulli κατά τη φύση των χαρακτηριστικών) με λιτή προεπεξεργασία. Δεν εφαρμόζεται GridSearchCV και το μοντέλο χρησιμοποιείται κυρίως ως σημείο αναφοράς για ταχείς συγκρίσεις.

AdaBoost

Ενισχυτικό σχήμα που αυξάνει τα βάρη των δύσκολων παραδειγμάτων σε διαδοχικούς ασθενείς μαθητές (μικρά δέντρα). Ο χώρος υπερπαραμέτρων περιλαμβάνει n_estimators και learning_rate. Η αναζήτηση υλοποιείται με GridSearchCV και στη συνέχεια εκπαιδεύεται το τελικό μοντέλο με τις επιλεγμένες τιμές και παράγονται οι προβλέψεις. Το AdaBoost αξιοποιείται ως ελαφρύ, ευαίσθητο σε καθαρά σήματα, όντας εναλλακτική επιλογή των βαθύτερων δασών (*boosting*).

Gradient Boosting

Διαδοχική εκμάθηση διορθώσεων στο υπόλοιπο σφάλμα, με έλεγχο πολυπλοκότητας μέσω learning_rate, n_estimators και βάθους (μεγέθους φύλλων). Η διερεύνηση υπερπαραμέτρων γίνεται με GridSearchCV και ακολουθεί εκπαίδευση του τελικού ταξινομητή και παραγωγή προβλέψεων. Το Gradient Boosting τοποθετείται ως βαθύτερο ενισχυτικό μοντέλο, ικανό να αποτυπώσει αλληλεπιδράσεις και μη γραμμικότητες.

Νευρωνικό Δίκτυο (MLP)

Πλήρως συνδεδεμένη αρχιτεκτονική (multi-layer perceptron) που απαιτεί κλιμάκωση

εισόδων και τυπικούς μηχανισμούς ρύθμισης (πλήθος και μέγεθος στρωμάτων, activation, max_iter/early stopping). Ορίζεται, εκπαιδεύεται και αξιοποιείται για προβλέψεις. Δεν εφαρμόζεται GridSearch στα παραπάνω κελιά, καθώς το MLP εντάσσεται ως ευέλικτη, μη γραμμική γραμμή σύγκρισης.

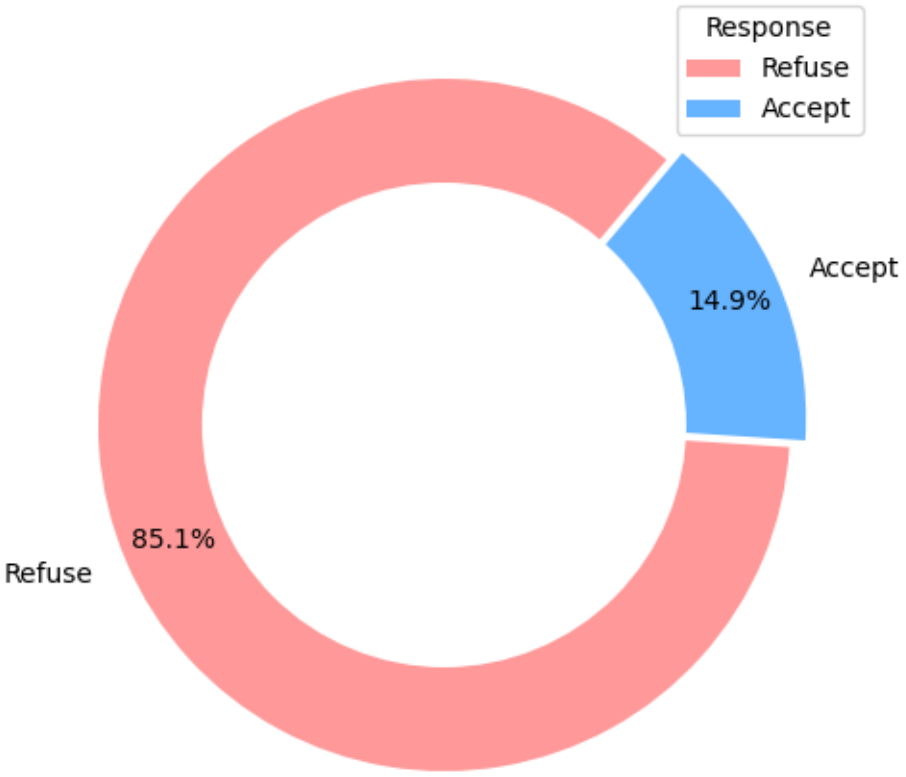
Συγκεντρωτικά, το εκτελεστικό φάσμα καλύπτει γραμμικά (LR), μοντέλα βάσει απόστασης (KNN), μη γραμμικούς ταξινομητές με περιθώριο (SVM), δενδροειδή ensembles (RF/Extra Trees) και ενισχυτικά σχήματα (AdaBoost/Gradient Boosting), καθώς και ένα νευρωνικό baseline (MLP). Όπου υλοποιήθηκε GridSearchCV (SVM, AdaBoost, Gradient Boosting, RF, Extra Trees, KNN), δηλώνεται ρητά. Τα Naive Bayes και MLP εκτελούνται με ρυθμίσεις που ορίζονται απευθείας στα αντίστοιχα κελιά. Η αποτίμηση όλων των παραπάνω ακολουθεί στο επόμενο κεφάλαιο, όπου παρουσιάζονται οι μετρικές, οι πίνακες σύγχυσης και η συγκριτική σύνοψη.

ΚΕΦΑΛΑΙΟ 5. ΑΠΟΤΕΛΕΣΜΑΤΑ & ΕΦΑΡΜΟΓΗ

5.1 Εξερεύνηση & Οπτικοποίηση

Η διερευνητική ανάλυση στοχεύει να αποτυπώσει συνοπτικά τη φύση του προβλήματος και να υποστηρίξει την ανάγνωση των συγκριτικών αποτελεσμάτων που ακολουθούν. Πρώτον, τεκμηριώνεται η ανισορροπία της θετικής τάξης. Δεύτερον, αναδεικνύονται προγνωστικά μοτίβα σε πρόσφατη δραστηριότητα και δαπάνες, που εμπλέκονται άμεσα με τα μοντέλα που θα αξιολογηθούν. Οι οπτικοποιήσεις παρατίθενται επιλεκτικά και με σαφές ερμηνευτικό μήνυμα παρακάτω.

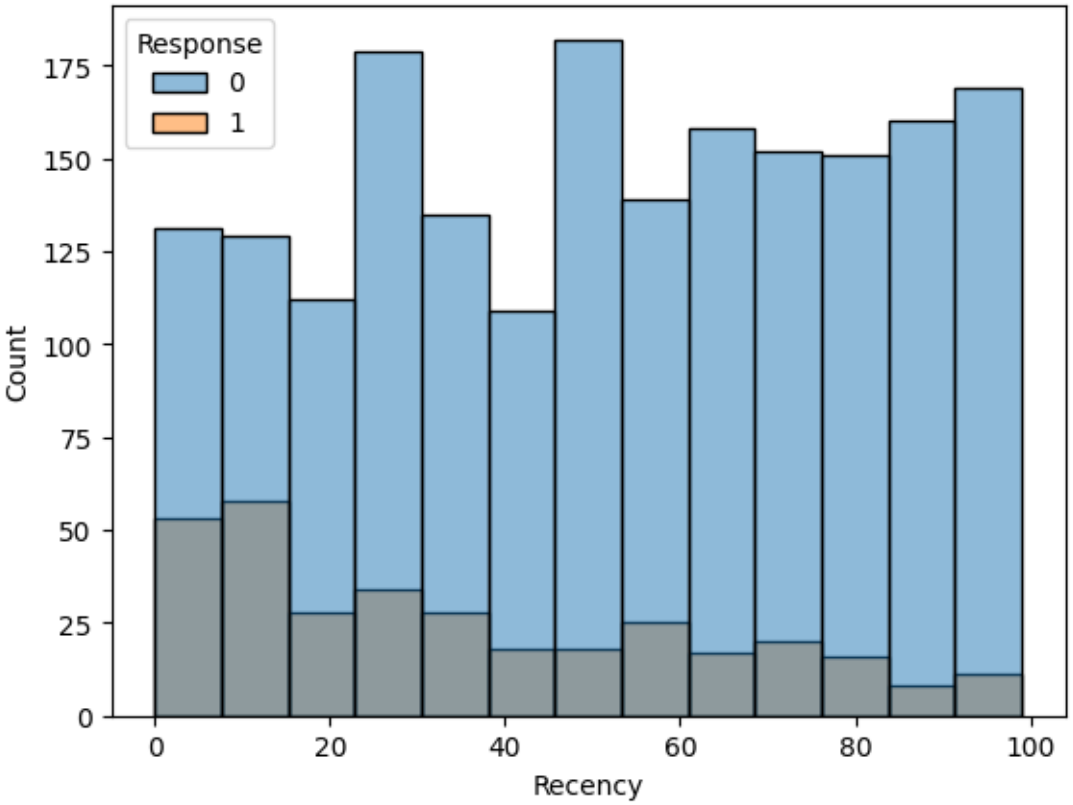
Η κατανομή του στόχου δείχνει ότι η θετική κλάση είναι μειοψηφική. Αυτό δικαιολογεί γιατί στο κεφάλαιο 6.3 δίνεται έμφαση σε μετρικές ευαίσθητες στην ανισορροπία (ιδίως F1 και Recall) και γιατί το απλό accuracy δεν επαρκεί για αξιολόγηση.



Εικόνα 6. Κατανομή Response σε γράφημα donut.

Η πρόσφατη αγοραστική δραστηριότητα του πελάτη στη μάρκα, όπως αποτυπώνεται από το Recency, συνδέεται καθαρά με την πιθανότητα αποδοχής. Χαμηλές τιμές Recency εμφανίζουν υψηλότερη αναλογία θετικών, στοιχείο που προοικονομεί ότι ταξινομητές με

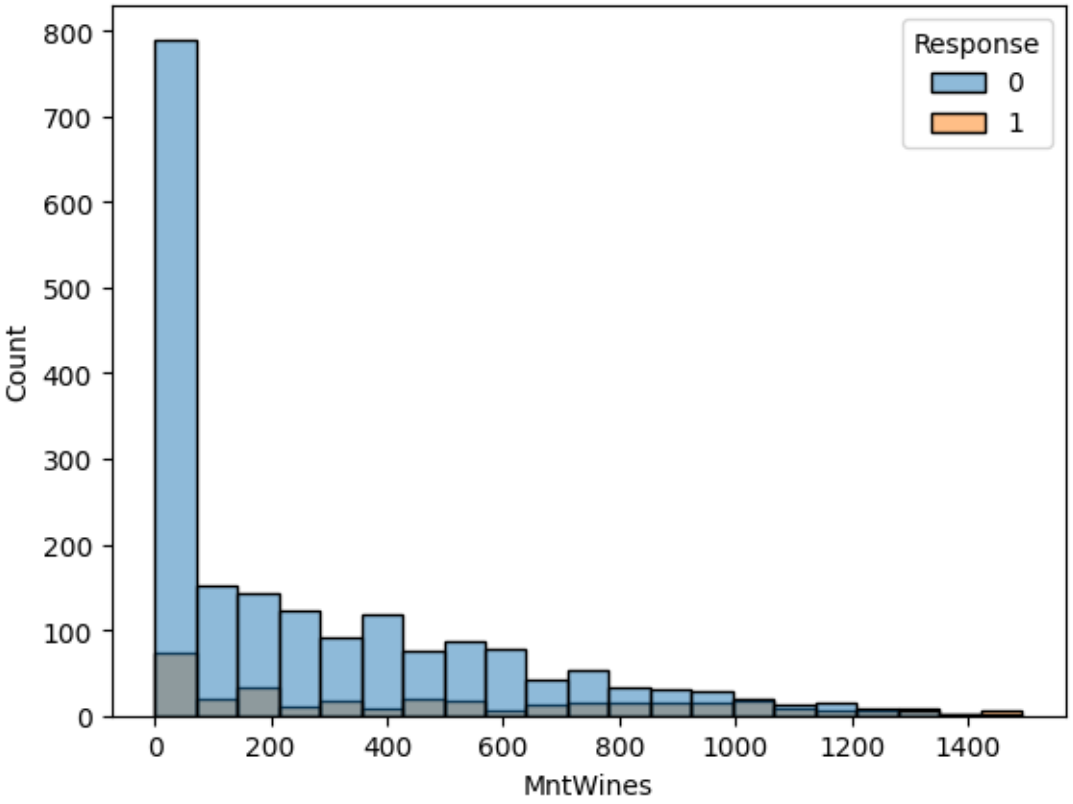
ικανότητα σύλληψης μονοτονικών και μη γραμμικών σχέσεων (π.χ. δέντρα, boosting) θα εκμεταλλευτούν αποτελεσματικά το χαρακτηριστικό αυτό.



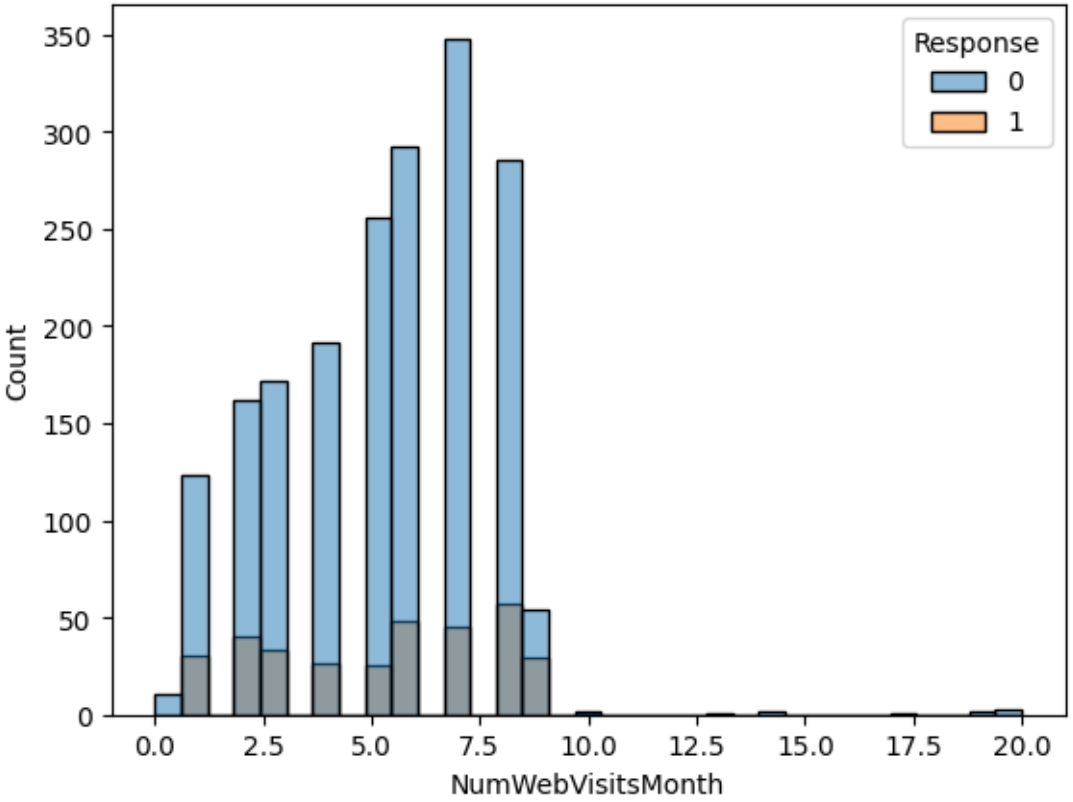
Εικόνα 7. Κατανομή Recency με συσχέτιση κατά Response.

Οι μεταβλητές δαπάνης ανά κατηγορία παρέχουν επίσης καθαρό προγνωστικό δείκτη. Η μεγαλύτερη σωρευτική δαπάνη σε ευαίσθητες κατηγορίες (ενδεικτικά κρέας ή κρασί) συσχετίζεται με υψηλότερη πιθανότητα ανταπόκρισης. Αυτό το μοτίβο είναι συνεπές με την υπόθεση εμπλοκής ή αξίας και συνήθως ενισχύει μοντέλα που αξιοποιούν αλληλεπιδράσεις μεταξύ δαπάνης και πρόσφατης δραστηριότητας.

Η ψηφιακή δραστηριότητα συμπληρώνει την εικόνα. Μέτριες έως υψηλές μηνιαίες επισκέψεις στον ιστότοπο συνδέονται με ελαφρώς αυξημένη πιθανότητα αποδοχής, χωρίς όμως να αποτελούν αυτόνομα επαρκή μετρική. Η πληροφορία αυτή είναι χρήσιμη όταν συνδυάζεται με πρόσφατη αγορά και δαπάνη, καθώς υποδηλώνει διαθεσιμότητα σε ψηφιακή επαφή.

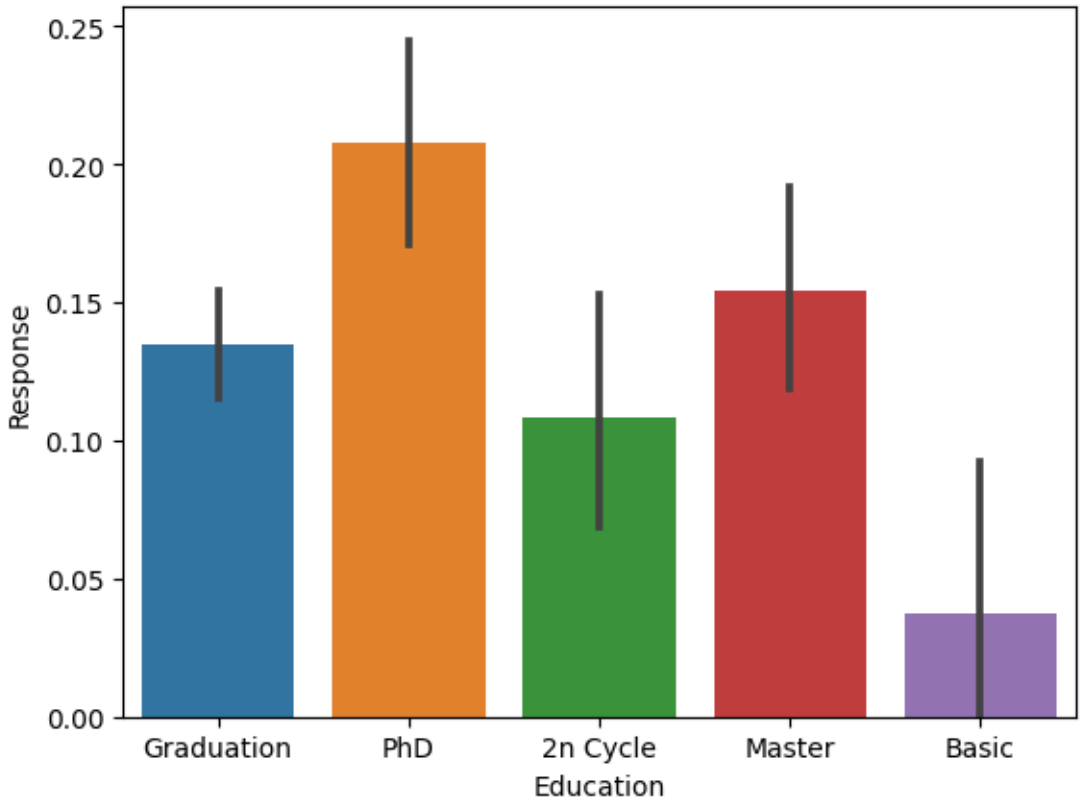


Εικόνα 8. Συσχέτιση MntWines κατά Response



Εικόνα 9. Συσχέτιση NumWebVisitsMonth κατά Response

Τα κατηγορικά προφίλ προσθέτουν ήπια διαφοροποίηση. Η κατανομή της Education ανά κλάση υποδεικνύει μικρές διαφοροποιήσεις που λειτουργούν επικουρικά έναντι των ισχυρότερων δεικτών από Recency και δαπάνες. Η ερμηνεία παραμένει περιγραφική και αξιοποιείται μόνο για συσχέτιση ως συμπτώσεις μοτίβων, όχι για αιτιότητα.



Εικόνα 10. Συσχέτιση Education κατά Response

Ως τεχνική υποσημείωση, ο έλεγχος ελλειπουσών τιμών έδειξε ότι μόνο το Income έχει 24 κενά, τα οποία συμπληρώθηκαν με μέση τιμή, όπως έχει ήδη τεκμηριωθεί στη ροή προεπεξεργασίας. Η πληροφορία καταγράφεται εδώ μόνο για λόγους πληρότητας τεκμηρίωσης και όχι ως αποτέλεσμα. Συνολικά, η EDA αναδεικνύει ορισμένα πρακτικά σημεία που θα καθοδηγήσουν την ανάγνωση των αποτελεσμάτων στο επόμενα υποκεφάλαια. Αρχικά, η σπάνια θετική τάξη επιβάλλει αξιολόγηση με F1 και Recall και όχι αποκλειστικά με Accuracy. Επίσης, οι δείκτες πρόσφατης δραστηριότητας και συνολικής δαπάνης παρέχουν στιβαρή ένδειξη για ταξινόμηση. Τέλος, οι δείκτες ψηφιακής εμπλοκής και τα κατηγορικά προφίλ λειτουργούν συμπληρωματικά. Έτσι, αναμένεται τα μοντέλα που αποτυπώνουν μη γραμμικότητες και αλληλεπιδράσεις να εμφανίσουν συγκριτικό πλεονέκτημα.

5.2 Application – Εκτέλεση μοντέλων

5.2.1 Ροή εκτέλεσης

- Αρχικοποίηση και ορισμός στόχου και χαρακτηριστικών. Οι μεταβλητές φόρτωσης και ελέγχου προετοιμάζονται, αφαιρούνται τα τεχνικά πεδία και προσδιορίζεται ο στόχος. Απομακρύνονται Id και Dt_Customer. Ορίζονται τα χαρακτηριστικά και η ετικέτα με x = data.drop("Response", axis=1) και y = data["Response"]. Η θετική κλάση είναι η αποδοχή προσφοράς, όπως έχει ήδη τεκμηριωθεί.
- Κωδικοποίηση κατηγορικών. Τα πεδία Education και Marital_Status μετατρέπονται σε αριθμητική μορφή με LabelEncoder. Η κωδικοποίηση προηγείται του υπολοίπου βήματος, ώστε όλα τα *downstream* μοντέλα να λαμβάνουν αριθμητικές εισόδους.
- Κλιμάκωση αριθμητικών. Εφαρμόζεται τυποποίηση με StandardScaler.fit_transform(X) στα συνεχή χαρακτηριστικά. Η κλιμάκωση ενιαίως προετοιμάζει το υπόστρωμα για αλγορίθμους ευαίσθητους στην κλίμακα (π.χ. KNN, SVM, MLP), χωρίς να αλλάζει τα δενδροειδή.
- Αντιμετώπιση ανισορροπίας. Υλοποιείται υπερδειγματοληψία της μειονοτικής κλάσης με SMOTE, sm.fit_resample(x, y). Η δημιουργία συνθετικών δειγμάτων στοχεύει σε βελτιωμένη κάλυψη του χώρου για τη θετική κλάση πριν τον χωρισμό.
- Διαχωρισμός εκπαίδευσης και ελέγχου. Ορίζεται τυχαίος, ικανός να αναπαραχθεί χωρισμός με train_test_split(x_sm, y_sm, test_size=0.20, random_state=27). Οι μεταβλητές x_train, x_test, y_train, y_test αποτελούν το σταθερό σημείο αναφοράς για όλα τα μοντέλα που ακολουθούν.
- Κανόνας εκτέλεσης. Κάθε ταξινομητής αρχικοποιείται, εκπαιδεύεται σε x_train, y_train και αξιολογείται περιγραφικά σε x_test, y_test. Όπου παράγεται αναφορά ταξινόμησης ή καμπύλη ROC και πίνακας σύγχυσης, παραπέμπεται το σχετικό κελί.

5.2.2 Εκτελεστικό πλαίσιο ανά classifier

Η εκτέλεση των ταξινομητών οργανώθηκε με ενιαία και συνεπή λογική, ώστε οι συγκρίσεις να βασίζονται αποκλειστικά στις ιδιότητες των αλγορίθμων και όχι σε διαφοροποιήσεις της ροής δεδομένων ή της αξιολόγησης. Κοινή αφετηρία για όλα τα

μοντέλα αποτελεί ο ίδιος πίνακας χαρακτηριστικών και ο ίδιος δυαδικός στόχος, όπως προέκυψαν μετά τα στάδια καθαρισμού, κωδικοποίησης και χειρισμού ανισορροπίας. Η εκπαίδευση και η αξιολόγηση πραγματοποιούνται στα ίδια σύνολα εκπαίδευσης και ελέγχου, με ενιαία αναπαράσταση των προβλέψεων και των μετρικών, ώστε να διασφαλίζεται πλήρης συγκρισιμότητα.

Η προεπεξεργασία προσαρμόζεται στις απαιτήσεις κάθε αλγορίθμου. Οι γραμμικοί και οι ταξινομητές βασισμένοι σε αποστάσεις (LR, KNN και SVM) χρησιμοποιούν τυποποιημένα χαρακτηριστικά, δεδομένης της ευαισθησίας τους στην κλίμακα των μεταβλητών. Αντίθετα, τα δενδροειδή μοντέλα και τα ensembles (Decision Tree, Random Forest, Extra Trees και Boosting) εκπαιδεύονται απευθείας στα αρχικά χαρακτηριστικά, καθώς είναι εγγενώς ανθεκτικά σε διαφορετικές κλίμακες. Όλα τα μοντέλα ακολουθούν την ίδια διαδικασία εκπαίδευσης και πρόβλεψης, ενώ όπου είναι διαθέσιμη η πιθανοτική έξοδος, αυτή χρησιμοποιείται για κατωφλίωση, καμπύλες ROC και συγκριτική αξιολόγηση.

Οι υπερπαράμετροι κάθε ταξινομητή επιλέγονται με συντηρητική προσέγγιση, είτε διατηρώντας τις προεπιλεγμένες τιμές είτε με περιορισμένη διερεύνηση βασικών ρυθμίσεων, με στόχο τη σταθερότητα και την αποφυγή υπερπροσαρμογής. Δεν εφαρμόζεται διαφοροποιημένη στρατηγική αξιολόγησης ανά μοντέλο. Αντιθέτως, όλα κρίνονται υπό το ίδιο πλαίσιο threshold απόφασης και τις ίδιες μετρικές απόδοσης. Με τον τρόπο αυτό, οι παρατηρούμενες διαφορές στην επίδοση αντανακλούν ουσιαστικά τις μαθησιακές ιδιότητες των αλγορίθμων και όχι αποκλίσεις στον τρόπο εκτέλεσης.

Η εκτέλεση στηρίζεται σε σταθερό σπόρο τυχαιοποίησης για κάθε τυχαίο βήμα, ώστε τα αποτελέσματα να επαναλαμβάνονται. Ο βασικός σπόρος ορίζεται στο train και test split με random_state = 27 και κληρονομείται όπου οι ταξινομητές υποστηρίζουν αντίστοιχη παράμετρο (π.χ. RF και Extra Trees). Το ίδιο ισχύει για μεθόδους που εμπεριέχουν τυχαιοποίηση (bootstrap, τυχαία διάσπαση χαρακτηριστικών).

Οι εξαρτήσεις λογισμικού και οι βασικές βιβλιοθήκες έχουν ήδη τεκμηριωθεί στο προηγούμενο κεφάλαιο, συνεπώς στο παρόν αναφέρονται οι εντολές που παράγουν πίνακες και διαγράμματα, όπως τα περιγραφικά στατιστικά και οι κατανομές describe(), histplots.

Για τον πλήρη κύκλο που αναπαράγεται από την εκπαίδευση μέχρι την πρόβλεψη και την αναφορά, κάθε μοντέλο ακολουθεί το ίδιο ζεύγος συνόλων (x_train, y_train σε x_test, y_test) και τους ίδιους κανόνες αναπαράστασης, όπως στο classification_report και RocCurveDisplay. Έτσι, οι όποιες διαφορές παρουσιάζονται μεταξύ μοντέλων

αποδίδονται καθαρά σε επιλογές αλγορίθμων και αντίστοιχων υπερπαραμέτρων και όχι σε ασυνέπειες επιλογής της εκτέλεσης. Ακόμη, όλα τα outputs που παρατίθενται στο παρόν υπολογίζονται στο default κατώφλι 0.5 των πιθανοτήτων, όπου αυτό υφίσταται (π.χ. LR, SVM με probability=True).

5.3 Συγκριτικά αποτελέσματα

Ακολούθως συγκρίνονται τα μοντέλα στο test set με Precision, Recall, F1, Accuracy στο κατώφλι 0.5, ως default threshold, και με ROC–AUC ως ανεξάρτητη μέτρηση διάκρισης. Ο κύριος πίνακας συνοψίζει τα αποτελέσματα. Επιπλέον, παρουσιάζονται ROC και PRC για να απεικονιστεί συνοπτικά η συμπεριφορά των κορυφαίων μοντέλων. Όλες οι μετρικές είναι στο test set και οι τιμές έχουν στρογγυλοποιηθεί σε 2 δεκαδικά.

Μοντέλο	**Precision**	**Recall**	**F1**	**Accuracy**	**ROC-AUC**
Logistic Regression	0.76	0.76	0.76	0.76	0.81
Decision Tree	0.91	0.91	0.91	0.91	—
Random Forest	0.96	0.96	0.96	0.96	0.90
Extra Trees	0.99	0.99	0.99	0.99	—
K-Nearest Neighbors	0.93	0.93	0.93	0.93	0.81
Support Vector Machine	0.90	0.90	0.90	0.90	0.90
Naive Bayes (Gaussian)	0.64	0.64	0.64	0.64	0.72
AdaBoost	0.91	0.91	0.91	0.91	—
Gradient Boosting	0.96	0.95	0.95	0.95	—
Neural Network (MLP)	—	—	—	0.94	—

Εικόνα 11. Συγκεντρωτική αξιολόγηση μοντέλων ταξινόμησης

Θα πρέπει να αναφερθεί πως όλες οι τιμές υπολογίστηκαν στο ίδιο split και οι καμπύλες ROC/PRC βασίζονται σε predict_proba και decision_function και για όσα μοντέλα υπολογίστηκε. Τα Precision, Recall και F1 είναι από τη weighted avg γραμμή των classification_report. Για το MLP έτρεξε μόνο το compiled metric, το οποίο ορίστηκε ως metrics=['acc'] και εκπαιδεύτηκε κατ' αυτόν τον τρόπο το μοντέλο.

Μοντέλο	*DT*	*RF*	*Extra Trees*	*KNN*	*LR*	*SVM*	*Ada-Boost*	*GB*	*NB (Gaussian)*	*K-Means*	*NN (MLP)*
F1	0.9135	0.9607	0.9869	0.9292	0.7628	0.8978	0.9096	0.9515	0.6396	0.4256	0.9353

Εικόνα 12. Επίδοση μοντέλων ταξινόμησης

Αξιολογώντας το κάθε μοντέλο, προκύπτουν συνοπτικά συγκριτικά μεγέθη μεταξύ τους. Πιο συγκεκριμένα, το LR λειτουργεί ως γραμμική βάση αναφοράς με καλή ερμηνευσιμότητα των συντελεστών, σταθερότητα και ενδείκνυται για σύγκριση και έλεγχο βαθμονόμησης πιθανοτήτων. Η απόδοση (F1 = 0.76) και η ROC-AUC = 0.81 δείχνουν ικανότητα διάκρισης, αλλά υστερεί έναντι δενδροειδών και *Boosting* σε μη γραμμικότητες. Από τα μη γραμμικά μοντέλα, το DT προσφέρει υψηλή ακρίβεια και καθαρούς κανόνες (μονοπάτια), αλλά είναι ευάλωτο σε υπερπροσαρμογή χωρίς κλάδεμα λόγω περιορισμών. Η επίδοση γύρω στο 0.91 F1 το καθιστά ανταγωνιστικό και ύποπτο για *data leakage*, με πλεονέκτημα την ερμηνευσιμότητα μονοπατιών απόφασης. Είναι κατάλληλο όταν ζητείται απλή, εξηγήσιμη λογική.

Το RF συνδυάζει δέντρα με τυχαιοποίηση και δίνει σταθερά υψηλές επιδόσεις (F1 = 0.96), με ROC-AUC = 0.90, όντας αξιόπιστο σε διάφορα σενάρια. Είναι ανθεκτικό σε θόρυβο και προσφέρει πολύτιμες σημαντικότητες χαρακτηριστικών. Μπορεί να βελτιστοποιηθεί με σωστό κατώφλι απόφασης, ειδικά όταν προτεραιότητα είναι το Recall, αποτελώντας καλή προεπιλογή γενικής χρήσης σε μικτή κλίμακα features. Αντέχει δηλαδή σε ακραίες τιμές (*outliers*) και δεν χρειάζεται απαραίτητα scaling. Το ET ομοιάζοντας του RF, συνιστά μοντέλο με περισσότερη τυχαιοποίηση στα splits και πετυχαίνει κορυφαία επίδοση (F1 = 0.99) στο τρέχον σύνολο, το οποίο παραπέμπει κατά μεγάλη πιθανότητα σε διαρροή δεδομένων. Τείνει να είναι ταχύτερο, με παρόμοια ή ελαφρώς καλύτερη απόδοση και λιγότερο ευαίσθητο σε παραμετροποίηση από το RF. Ισχυρή επιλογή όταν προέχει η ακρίβεια πρόβλεψης. Το KNN εκμεταλλεύεται τοπικές ομοιότητες και αποδίδει πολύ καλά (F1 = 0.93), με ROC-AUC = 0.81. Απαιτεί κλιμάκωση και προσοχή στη διάσταση και πυκνότητα των δεδομένων. Χρήσιμο όταν οι γειτνιάσεις έχουν σαφές νόημα επιχειρησιακά, όπως πελάτες με όμοια χαρακτηριστικά να έχουν και παρόμοια συμπεριφορά. Αξιοποιείται όταν ζητείται χαμηλό FN θυσιάζοντας λίγο από το Precision. Η ROC έχει λίγα σπασίματα επειδή τα scores είναι διακριτά (m και k), καθώς προκύπτουν από ποσοστά ψήφων των *k* γειτόνων, χωρίς να υπάρχουν έτσι πολλά ενδιάμεσα thresholds.

Το SVM διατηρεί υψηλή ισορροπία precision–recall (F1 = 0.90) και ισχυρή, ομαλή διάκριση (ROC-AUC = 0.90). Επωφελείται από σωστό scaling και ρύθμιση C/γ, χωρίς tuning κατωφλιού. Κατάλληλο σε μεσαίου μεγέθους δεδομένα με καθαρά περιθώρια διαχωρισμού. Το Gaussian NB αποτελεί ένα ταχύ μοντέλο βάσης με χαμηλότερη επίδοση (F1 = 0.64, ROC-AUC = 0.72) λόγω υπόθεσης ανεξαρτησίας. Παραμένει χρήσιμο για

γρήγορο σημείο αναφοράς και όταν τα χαρακτηριστικά είναι περίπου στα πρότυπα της καμπύλης του Gauss, ταιριάζοντας έτσι τα features στην υπόθεση ανεξαρτησίας. Μπορεί να συνυπάρξει σε στοίχιση (*stacking*) ως ελαφρύς μαθητής του μοντέλου.

Το AB, ως ensemble αδύναμων δένδρων με εστίαση στα λάθη προηγούμενων γύρων, ανυψώνει δύσκολα δείγματα και επιτυγχάνει ισορροπημένη και ανέλπιστα υψηλά επίδοση εδώ, εγείροντας ανησυχίες για *data leakage* (F1 = 0.91). Λειτουργεί καλά όταν υπάρχουν καθαρές μετρικές (χωρίς outliers) και μέτριος θόρυβος, παρουσιάζει καλό precision με αξιοπρεπές recall, ενώ τείνει να βελτιώνεται με tuning. Ωστόσο, απαιτεί προσεκτική ρύθμιση εκτιμητών και learning rate για να αποφευχθεί overfitting. Στο ίδιο πλαίσιο, το GB συλλαμβάνει περίπλοκες μη γραμμικότητες και αλληλεπιδράσεις, με πολύ υψηλή απόδοση (F1 = 0.95). Προσφέρει λεπτό έλεγχο πολυπλοκότητας (learning rate, depth, subsample) και αποτελεί ισχυρή επιλογή όταν ζητείται μέγιστη προγνωστική ικανότητα, σε μη τεράστια σύνολα, αφού παρουσιάζει υψηλά F1 και AUC. Τέλος, το NN (MLP) παρέχει υψηλή ακρίβεια (Accuracy = 0.94) όταν η προεπεξεργασία και η κλιμάκωση είναι σωστές. Θέλει προσεκτικό tuning (αρχιτεκτονική, regularization, dropout, early stopping, neurons, batch size) και επαρκή δεδομένα. Χρήσιμο όταν αναμένονται σύνθετα, μη γραμμικά πρότυπα. Στην παρούσα βρίσκεται μόνο το accuracy λόγω της *compile* ρύθμισης.

5.4 Επιλογή τελικού μοντέλου

Τα κριτήρια της επιλογής του τελικού μοντέλου, βασίζονται σε ισορροπία μεταξύ των recall, precision και F1 στο test set (κατώφλι 0,5), με υψηλή προτεραιότητα στην ανάκτηση θετικών περιπτώσεων (*marketing reach*) υπό αποδεκτό κόστος επαφής. Η ROC–AUC αξιοποιείται συμπληρωματικά ως δείκτης ικανότητας διάκρισης ανεξαρτήτως κατωφλίου, ενώ όταν υπάρχει, η PR–AUC θεωρείται πιο ενδεδειγμένη σε περιβάλλον σπάνιας θετικής τάξης. Παράλληλα, συνεκτιμώνται πρακτικά κριτήρια όπως σταθερότητα σε διαφορετικούς σπόρους ή splits, χρόνος πρόβλεψης (batch scoring), απαιτήσεις συντήρησης (tuning, retraining κτλ) και ερμηνευσιμότητα για τεκμηρίωση πολιτικών στόχευσης.

Βάσει του 5.3, τα υποψήφια καταληκτικά μοντέλα προς επιλογή είναι το SVM, το RF και το KNN. Το SVM παρουσιάζει πολλή ισορροπημένη συμπεριφορά, με το precision να είναι περίπου ίσο με το recall και το F1 και το ROC-AUC να είναι περίπου στο 0.90, με ομαλή καμπύλη, άρα και προβλέψιμη απόκριση σε μεταβολές κατωφλίου. Το RF δείχνει

ισχυρό ranking (ROC–AUC = 0,90), συχνά βελτιώνεται ουσιαστικά με ρύθμιση κατωφλίου προς υψηλότερο recall και προσφέρει feature importances, το οποίο παίζει πολύ βοηθητικό ρόλο. Το KNN, ως τρίτη επιλογή, έχει υψηλό recall και είναι χρήσιμο όταν προέχει η μείωση FN, να μη χάνονται δηλαδή δυνητικά αποδεκτοί, με πιθανό κόστος σε precision.

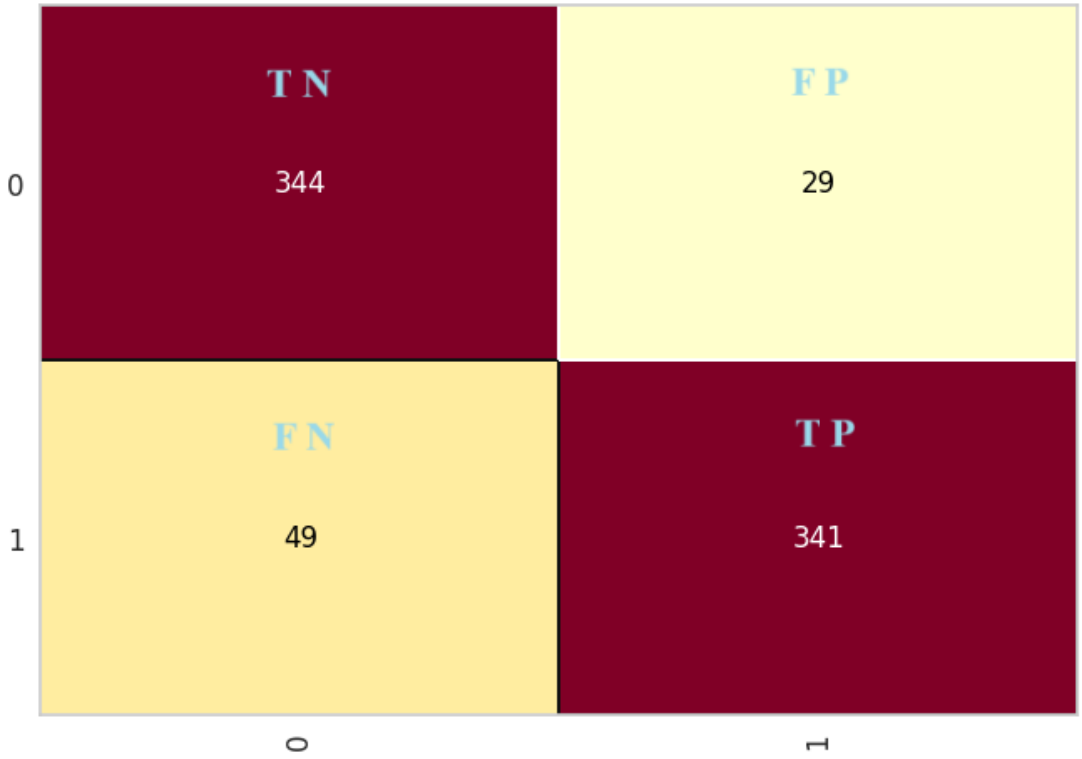
Συμπερασματικά, επιλέγεται το SVM ως τελικό μοντέλο διότι προσφέρει την καλύτερη ισορροπία precision και recall στο default κατώφλι, υψηλή AUC, ομαλή συμπεριφορά στο φάσμα κατωφλιών και σταθερότητα χωρίς βαριά παραμετροποίηση. Επιπλέον, υποστηρίζει probability outputs (με probability = True στην εκπαίδευση), διευκολύνοντας κατωφλίωση και δοσολογία καμπάνιας (τμηματοποίηση ανά decile score). Αυτό γιατί αφενός μπορεί να επιλεγεί κατώφλι διαφορετικό από το 0.5, ανάλογα με στόχους κόστους–οφέλους και αφετέρου το SVM μπορεί να δώσει πιθανότητες και όχι απλώς προβλέψεις, εφόσον οριστεί με probability = True. Έτσι, οι πελάτες μπορούν να ομαδοποιηθούν με βάση πόσο πιθανό είναι να ανταποκριθούν (π.χ. top 10%, δεύτερο 10% κ.ο.κ.).

Ως εναλλακτική, όταν προέχει το recall και ο επιχειρησιακός στόχος είναι να μειωθούν τα FN (π.χ. σε καμπάνια όπου η απώλεια ενός θετικού πελάτη επιβαρύνει το churn rate ή μειώνει CLV), τότε προκρίνεται είτε ο KNN (λόγω υψηλού recall για κατώφλι 0.5) είτε ο RF με μετατόπιση κατωφλίου προς χαμηλότερες τιμές πιθανότητας για να αυξηθεί η κάλυψη (με *frequency capping*, δηλαδή περιορισμό της συχνότητας επικοινωνίας με κάθε άτομο για να ελεγχθεί το κόστος FP). Ωστόσο, η τελική απόφαση δεν θα παρθεί με κατώφλι 0.5, αλλά θα συνοδεύεται από την βέλτιστη τιμή του, η οποία προκύπτει από σάρωση σε διάστημα [0,1] και μεγιστοποιεί κάποιο προκαθορισμένο κριτήριο. Η επιλογή αυτή του κριτηρίου μπορεί να είναι το F1 όταν ζητείται συνολική ισορροπία, είτε το αναμενόμενο καθαρό κέρδος ανά επαφή που υπολογίζεται από τον τύπο Profit = $TP \cdot \pi - FP \cdot c - Contacts \cdot c_{channel}$ με το π να αναπαριστά το οριακό περιθώριο κέρδους από TP, c το κόστος λανθασμένης επαφής και $c_{channel}$ το κόστος καναλιού.

Η ερμηνεία κόστους καναλιού μπορεί να προσεγγιστεί με email, με sms και με τηλεφωνική κλήση. Το email είναι οικονομικά αμελητέο ανά επαφή, αλλά υπάρχει κόπωση (*contact fatigue*), απεγγραφή (*unsubscribe*) και αρνητικό *brand sentiment* αν αυξηθούν τα FP (επηρεάζοντας την πιστότητα και ικανοποίηση του πελάτη και μελλοντικό engagement rate). Μπορεί επίσης να γίνει με sms, όπου το κόστος είναι μέτριο, αλλά απαιτείται αυστηρότερο κατώφλι. Υψηλότερο κόστος ωστόσο παρουσιάζει

η τηλεφωνική κλήση, για αυτό κατά κανόνα στοχεύει σε υψηλά deciles (top 10% - 20%) πελατών. Στην πράξη, το SVM παράγει πιθανότητες, γίνεται σάρωση κατωφλιών (π.χ. βήμα 0,01) και επιλέγεται βέλτιστο κατώφλι με βάση πίνακα κόστους–οφέλους και περιορισμούς χωρητικότητας (π.χ. μέγιστος ημερήσιος όγκος επαφών).

Για παραγωγική χρήση θα πρέπει να προηγηθεί βαθμονόμηση, ώστε οι πιθανότητες του SVM να παραλληλιστούν με τις πραγματικές συχνότητες. Η βαθμονόμηση βελτιώνει την ιεράρχηση στόχων, την ακρίβεια του βέλτιστου κατωφλίου και τη δοσολογία ανά κανάλι ή κύμα (π.χ. top–decile σήμερα, επόμενο decile αύριο). Στην περίπτωση των κλήσεων, τα μεγέθη των decile συμβάλλουν και στην αξιολόγηση της επιτυχίας των υπαλλήλων, αφού αποτελούν προσανατολισμό για την αναμενόμενη θετική ή μη ανταπόκριση. Ωστόσο, παρότι το SVM είναι λιγότερο διαφανές από ένα DT, η χρήση *permutation importances* (δείχνει χαρακτηριστικά που επηρεάζουν το μοντέλο συνολικά) ή SHAP (δείχνει πως κάθε χαρακτηριστικό επηρεάζει την πρόβλεψη για κάθε μεμονωμένο άτομο), σε παραγωγικό περιβάλλον, παρέχει *global* και *local* εξηγήσεις για έλεγχο μεροληψίας και κατανόηση μοτίβων και παραγόντων που αυξάνουν την πιθανότητα θετικής απόκρισης (π.χ. Recency, επίπεδα δαπάνης, web activity). Όπου απαιτείται πολιτική διαφάνειας, μπορεί να διατηρηθεί ως *«shadow model»* ένα RF για να δείχνει πιο ξεκάθαρα τα μοτίβα και σύγκλιση αποφάσεων, ώστε να επιβεβαιώνει ότι το SVM παίρνει λογικές αποφάσεις. Συνοπτικά, η διαδικασία απόφασης ξεκινάει με το φιλτράρισμα των μοντέλων που ικανοποιούν ελάχιστο recall και όπου είναι διαθέσιμη, PR-AUC. Έπειτα κρατούνται 2-3 υποψήφια με υψηλότερο F1 και σταθερότητα. Στο επικρατέστερο εξ' αυτών (SVM) εφαρμόζεται *calibration* και σάρωση κατωφλιού. Ακολούθως υπολογίζεται καθαρό κέρδος ανά κατώφλι με περιορισμούς (budget, channel cost, capacity κτλ). Η διαδικασία κλείνει με επιλογή ζεύγους (μοντέλο, βέλτιστο κατώφλι) με μέγιστο όφελος και λειτουργική καταλληλότητα (batch scoring, χρόνος πρόβλεψης).



Εικόνα 13. SVM Confusion Matrix

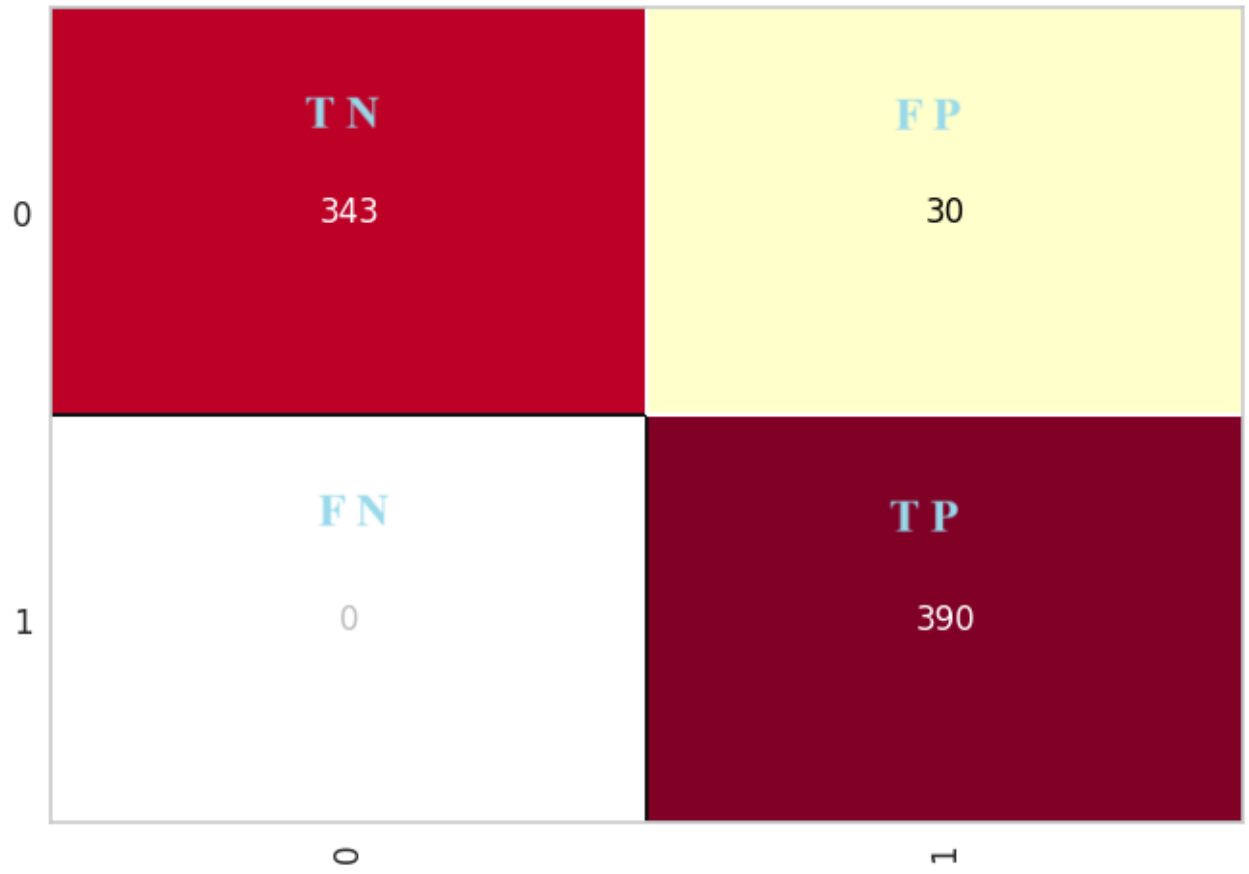
Το confusion matrix του SVM στο επιλεγμένο κατώφλι (για την εργασία παραμένει στο 0,5) αποτυπώνει το trade-off FP/FN.

	precision	recall	f1-score	support
0	0.85	0.85	0.85	399
1	0.84	0.84	0.84	364
accuracy			0.85	763
macro avg	0.85	0.85	0.85	763
weighted avg	0.85	0.85	0.85	763

Εικόνα 14. SVM Score

Το μοντέλο χάνει 49 πραγματικούς θετικούς (FN), συνεπώς έχει απολεσθέντα πιθανά έσοδα από τους δυνητικούς αποδέκτες που χάθηκαν με κίνδυνο αύξησης του churn rate αν δεν λάβουν έγκαιρα την προσφορά. Επίσης, στοχεύει 29 αρνητικούς (FP), το οποίο αυξάνει το κόστος λόγω των άσκοπων επαφών (κόστος καναλιού, κορεσμός και μείωση engagement, αύξηση unsubscribe rate). Ισορροπεί καλά precision με recall ωστόσο, ενώ

είναι λίγο πιο «σφιχτό» από πλευράς recall σε σχέση με RF.



Εικόνα 15. RF Confusion Matrix

	precision	recall	f1-score	support
0	0.84	0.78	0.81	399
1	0.77	0.83	0.80	364
accuracy			0.80	763
macro avg	0.81	0.81	0.80	763
weighted avg	0.81	0.80	0.80	763

Εικόνα 16. RF Score

Το θετικό με το RF είναι ότι δεν χάνει κανέναν θετικό (FN=0), άρα παρέχει μέγιστη κάλυψη με τίμημα 30 FP (περισσότερες άσκοπες επαφές). Είναι ιδανικό όταν το κόστος FN (να χάσεις πρόθυμο πελάτη) είναι πολύ ακριβό. Ενώ το SVM έχει λιγότερα FP, δέχεται μερικά FN, το οποίο το καθιστά πιο συντηρητικό αφού μειώνει άσκοπες επαφές,

ρισκάροντας έτσι να χάσει κάποιους πρόθυμους. Το RF από την άλλη έχοντας μηδενικά FN και περισσότερα FP, μπορεί να χαρακτηριστεί ως περισσότερο επιθετικό, καθώς πιάνει όλους τους πρόθυμους, πληρώνοντας το τίμημα δεχομένο παραπάνω άσκοπες επαφές.

Το SVM προκρίνεται ως τελικό μοντέλο λόγω ισορροπίας, υψηλού AUC και προβλέψιμης συμπεριφοράς στο φάσμα κατωφλιών. Η επιχειρησιακή απόφαση θα οριστικοποιηθεί με επιλογή κατωφλιού βάσει της σχέσης κόστους-οφέλους που θα οριστεί, ενδεχόμενη βαθμονόμηση πιθανοτήτων και τους κανόνες καναλιού ή της συχνότητας που διασφαλίζουν κέρδος, χαμηλό churn και βιώσιμο engagement.

Ολοκληρώνοντας με την λειτουργική ετοιμότητα και τον έλεγχο του μοντέλου, ώστε να υλοποιηθεί η χρήση του, προτού πραγματοποιηθεί η μετάβαση σε εφαρμογή θα πρέπει να γίνουν ορισμένα διαδικασίες. Decile targeting (top–k) και batch scoring σε τακτά κύματα. Ακόμη πρέπει να ακολουθήσουν *testing* ομάδων ή γεωγραφικά πειράματα τμηματοποίησης για *incremental lift,* όπως αναλύθηκε η χρησιμότητα κατά τη βιβλιογραφική ανασκόπηση. Ομοίως, παρακολούθηση του *data drift* (π.χ. με τα scores του μοντέλου ή Population Stability Index - PSI) και *retraining triggers*, ώστε το μοντέλο να εκπαιδεύεται εκ νέου, βάσει του κατωφλιού ποιότητας που έχουν τεθεί. Θα πρέπει ακόμη να γίνονται *fairness checks* ανά υποομάδα πελατών με ηλικιακά, χωροθετικά, εισοδηματικά κριτήρια κ.ά. (π.χ. έλεγχος αν αδικείται μία ομάδα στην επιλογή και εξισορρόπηση FN rate), ενώ οι κανόνες διαχείρισης επικοινωνίας με τους πελάτες θα πρέπει να πληρούν νομιμότητας, ηθικής και να μην κουράζουν τον πελάτη. Συγκεκριμένα, για να προχωρήσει η εφαρμογή ο πελάτης θα πρέπει να έχει συμμόρφωση με το *opt-in*, να έχει εύκολο *opt-out*, να τοποθετείται κατ' επιθυμία σε *suppression lists* και να τηρούνται τα *fatigue rules*. Εν ολίγοις, να υφίσταται ο πελάτης σε νόμιμο πλαίσιο που τον σέβεται και δεν τον «βομβαρδίζει» με παραπάνω δεδομένα από όσα του είναι ευχάριστο.

ΚΕΦΑΛΑΙΟ 6. ΣΥΖΗΤΗΣΗ & ΕΠΙΠΤΩΣΕΙΣ

6.1 Ερμηνεία αποτελεσμάτων

Τα ευρήματα συγκλίνουν σε ένα καθαρό μοτίβο. Τα μοντέλα με ομαλά, συνεχούς score (SVM, RF) πέτυχαν την καλύτερη ικανότητα διάκρισης (ROC–AUC) και ισορροπημένο precision–recall στο default κατώφλι. Τεκμήριο αποτελούν οι καμπύλες ROC/PR του 5.3, όπου οι τροχιές τους παραμένουν ομαλές σε όλο το φάσμα κατωφλιών, σε αντίθεση με μοντέλα που παράγουν scores με σκαλοπάτια (π.χ. KNN). Συνεπώς, η καλή κατάταξη σε όλο το φάσμα κατωφλιών υποδεικνύει ότι αυξάνεται ακόμη περισσότερο το κέρδος με ρύθμιση κατωφλίου στο 5.4. Με άλλα λόγια, το πρότυπο επίδοσης δείχνει υπεροχή μοντέλων με πλούσιο score (SVM/RF), ενώ το KNN ευνοεί υψηλό recall εις βάρος του precision, στοιχείο χρήσιμο όταν το κόστος FN είναι υψηλό.

Αναλύοντας το ισοζύγιο precision–recall ανά μοντέλο, το SVM ξεχωρίζει για την ισορροπία του, καθώς παρουσιάζει σύγκλιση των δύο μετρικών και ομαλή ROC, άρα σταθερότητα στις αποφάσεις υπό μεταβαλλόμενα κατώφλια. Το Random Forest επιδεικνύει πολύ καλή ικανότητα διάκρισης. Από επιχειρησιακή σκοπιά, ένα κατώφλι που μετατοπίζει τη βαρύτητα προς το recall, μπορεί να το καταστήσει ιδανικό όταν το να χαθούν πρόθυμοι πελάτες (FN) είναι ακριβό. Το KNN τείνει να προσφέρει υψηλότερο recall με χαμηλότερο precision, πολύτιμο σε εκστρατείες κάλυψης (reach) όπου επιτρέπονται περισσότερα FP. Η LR λειτουργεί ως σταθερό baseline και σημείο αναφοράς για πιθανότητες και καλιμπράρισμα, ενώ ο Gaussian Naive Bayes παραμένει χρήσιμος για τον ίδιο λόγο και λόγω ταχύτητας, αλλά υποχωρεί όταν οι εξαρτήσεις μεταξύ χαρακτηριστικών είναι ουσιώδεις.

Σε επίπεδο γενικής κατεύθυνσης, τα ευρήματα της EDA και οι συμπεριφορές των μοντέλων είναι συνεκτικά. Η Recency αναδείχθηκε κομβική, με το μικρότερο διάστημα από την τελευταία αγορά να συνδέεται με μεγαλύτερη πιθανότητα ανταπόκρισης, και τα δενδροειδή ενισχυτικά μοντέλα πιάνουν αυτό το πρότυπο μέσω μη γραμμικών ορίων. Οι δείκτες συμπεριφορικής δαπάνης και έντασης αγορών (π.χ. Mnt* και Num*Purchases) προσφέρουν ισχυρό τεκμήριο που αξιοποιείται ιδιαίτερα από RF και Gradient Boosting, τα οποία αποτυπώνουν αλληλεπιδράσεις συσχέτισης τύπου τιμή-προώθηση-εποχικότητα. Η web activity (NumWebVisitsMonth και NumWebPurchases) συνεισφέρει θετικά και πρακτικά ενημερώνει επιλογές καναλιού

στόχευσης. Τα δημογραφικά (όπως Education) εμφανίζονται ασθενέστερα και δευτερεύοντα. Χρησιμοποιούνται με προσοχή, τόσο για αποφυγή μεροληψίας όσο και για να μην επισκιάσουν δυναμικές ενδείξεις συμπεριφοράς. Όπου διατίθενται importances (π.χ. από RF), οι σημαντικότερες μεταβλητές στοιχίζονται με τα ευρήματα EDA (Recency, δαπάνες), ενισχύοντας την ερμηνευσιμότητα του τελικού σχήματος κατάταξης. Ωστόσο, ελλείψει τοπικών εξηγήσεων, προτείνεται εφαρμογή SHAP σε επόμενο βήμα.

Οι μήτρες σφαλμάτων φωτίζουν το επιχειρησιακό trade-off. Στο SVM παρατηρείται συντηρητικό προφίλ με περιορισμένα FP και μετρήσιμα FN, το οποίο είναι ωφέλιμο όταν το κόστος μιας άστοχης επαφής δεν είναι αμελητέο. Αντιστρόφως, σε πιο επιθετικές ρυθμίσεις δασικών μοντέλων το FN μπορεί να περιοριστεί με τίμημα περισσότερα FP, κάτι που ταιριάζει σε σενάρια όπου το κόστος χαμένου πρόθυμου πελάτη υπερβαίνει το κόστος επαφής. Συνεπώς, η επιλογή κατωφλίου θα πρέπει να γίνει με πίνακα κόστους–οφέλους, λαμβάνοντας υπόψη τόσο το άμεσο κόστος ή έσοδο, όσο και δευτερογενείς επιπτώσεις (όπως κόπωση καναλιού, αύξηση unsubscribe, μεταβολή churn rate κ.ά.).

Εντούτοις, επειδή οι αποφάσεις θα στηριχθούν σε πιθανότητες (thresholding, κατάταξη, δοσολογία καμπάνιας), προτείνεται calibration για βελτίωση της αξιοπιστίας των scores, ειδικά αν εφαρμοστεί cost-sensitive thresholding. Η καλύτερη αντιστοίχιση πιθανότητας και συχνότητας κάνει συνεπέστερη την ιεράρχηση στόχων και επιτρέπει κανόνες όπως top-k υπό περιορισμό προϋπολογισμού ή χωρητικότητας καναλιού. Συνολικά, τα αποτελέσματα είναι συνεπή με τα μοτίβα της EDA και αναδεικνύουν τον ρόλο της ρύθμισης κατωφλίου ως καταλύτη επιχειρησιακής αξίας.

6.2 Ακρίβεια, σταθερότητα & περιορισμοί

Σε επίπεδο συνολικής εικόνας, οι επιδόσεις είναι συνεπείς μεταξύ διαφορετικών ταξινομητών. SVM και RF διατηρούν υψηλή ικανότητα διάκρισης με ισορροπημένο precision–recall, ενώ η συμπεριφορά τους παραμένει σταθερή σε μικρές μεταβολές κατωφλίου. Αντίθετα, μοντέλα όπως το KNN εμφανίζουν μεγαλύτερη διακύμανση λόγω διακριτών scores και ευαισθησίας στην κλίμακα. Η ακρίβεια που παρατηρείται στο συγκεκριμένο test split δηλώνει αντοχή σε αυτό το δείγμα, χωρίς να συνεπάγεται αυτομάτως μεταφερσιμότητα σε άλλα χρονικά παράθυρα ή πληθυσμούς.

Όσο για την διαρροή πληροφοριών (data leakage), αυτή είναι πολύ πιθανή καθώς μερικά βήματα προεπεξεργασίας (π.χ. oversampling) εφαρμόστηκαν πριν τον διαχωρισμό σε train και test, κάτι που μπορεί να αισιοδοξεί τις μετρικές, αφού το test βλέπει πληροφορία από το train. Επίσης, αναφορικά με την ανισορροπία και SMOTE εκτός ροής CV, η χρήση SMOTE πριν το split μεταβάλλει την κατανομή του test και επηρεάζει τις συγκρίσεις (ιδίως σε recall και precision).

Η εξάρτηση από το κατώφλι είναι ακόμη ένας κίνδυνος αποσταθεροποίησης. Οι τιμές F1, precision και recall αναφέρονται στο default 0.5. Διαφορετικές επιχειρησιακές ρυθμίσεις κατωφλίου μπορούν να αντιστρέψουν τη σειρά κατάταξης μοντέλων. Ομοίως, όπου δεν χρησιμοποιήθηκε nested CV ή αυστηρός διαχωρισμός tuning–αποτίμησης, αυξάνεται ο κίνδυνος αισιοδοξίας (overfitting στο validation). Για αυτό τον λόγο η βαθμονόμηση πιθανοτήτων είναι ζωτικής σημασίας. Οι πιθανότητες ενδέχεται να μην είναι καλιμπραρισμένες, περιορίζοντας την αξιοπιστία των αποφάσεων κόστους–οφέλους.

Ανησυχίες σχετικά με την αντιπροσωπευτικότητα του δείγματος έχουν γερή βάση, καθώς τα δεδομένα προέρχονται από συγκεκριμένο περιβάλλον και περίοδο, συνεπώς υπάρχει εμφανές ρίσκο *dataset shift* (seasonality, προωθητικές ενέργειες, κανάλια). Εφόσον δεν εφαρμόστηκαν time-based splits, τίθεται θέμα χρονικής συνοχής, καθώς υποβόσκει ο κίνδυνος temporal leakage, όταν τα χαρακτηριστικά ενσωματώνουν γνώση μεταγενέστερη του prediction window. Η απουσία PRC–AUC για όλα τα μοντέλα και η έλλειψη cost curves περιορίζουν την επιχειρησιακή βαρύτητα των συγκρίσεων. Ακόμη, η χρήση δημογραφικών χωρίς *fairness checks* ανά υποομάδα ενέχει κίνδυνο bias.

Η απουσία ενιαίου pipeline και πλήρους seed control μπορεί να επιφέρει μικρές αποκλίσεις σε επαναλήψεις, διακυβεύοντας την αναπαραγωγιμότητα του μοντέλου. Είναι απαραίτητη η σωστή ροή στο pipeline, ώστε encoders, scaler και SMOTE να εκπαιδεύονται μόνο από το train (ανά fold). Nested cross-validation ή εναλλακτικά σαφής αλυσίδα train–validation–test με stratified k-fold στο train, είναι κρίσιμη για αυστηρότερη επικύρωση. Έπειτα ακολουθεί η σάρωση κατωφλιών και επιλογή με profit curve / expected cost αντί για fixed 0.5, για βελτιστοποίηση κατωφλίου με κόστος–όφελος.

Ακόμη, calibration πιθανοτήτων και πληρέστερες μετρικές, είναι εξίσου σημαντικές.

Συστηματική PRC–AUC σε όλα τα μοντέλα και cost ή benefit curves για αποφάσεις marketing, είναι χρήσιμα εργαλεία λήψης αποφάσεων. Έλεγχοι fairness και σταθερότητας υλοποιούνται με τη λήψη μετρικών ανά υποομάδα (π.χ. FNR parity), drift monitoring (PSI και KS, performance drift) και retraining triggers. Η χρονική εγκυρότητα με time-based splits και εφαρμογή rolling CV, όπου υφίσταται χρονική εξάρτηση, είναι μία αξιόπιστη μέθοδος ελέγχου. Σταθεροί seeds, καταγραφή εκδόσεων βιβλιοθηκών και ντετερμινιστικές ρυθμίσεις όπου είναι εφικτό, διασφαλίζουν την ομαλή αναπαραγωγιμότητα του μοντέλου μελλοντικά.

Ως προς την αξιοπιστία των μοντέλων πάντως, είναι υπεύθυνο να σημειωθεί πως τα δύο matrixes έχουν πραγματικές τάξεις 373/390, δηλαδή σχεδόν ισορροπημένες. Αυτό προκύπτει από τη ροή όπου έγινε SMOTE με προεπεξεργασία πριν το split. Σε γενική εικόνα, τα ευρήματα είναι ενθαρρυντικά, αλλά πρέπει να συνοδευτούν υπό τους ανωτέρω περιορισμούς. Η μετάβαση σε ενιαίο pipeline, αυστηρότερο validation, καλιμπράρισμα και κατωφλίωση βάσει κόστους–οφέλους, εκτιμάται πως θα προσδώσει πιο αξιόπιστες και επιχειρησιακά συνεπείς επιδόσεις.

6.3 Σύγκριση με βιβλιογραφία

Τα πρότυπα επίδοσης που παρατηρήθηκαν εναρμονίζονται με όσα αναδεικνύει η σχετική βιβλιογραφία για ταξινόμηση σε δομημένα και μικτά δεδομένα λιανικής. Μοντέλα με πλούσιο και ομαλό score, ιδίως SVM και RF, εμφανίζονται συστηματικά ισχυρά σε ικανότητα διάκρισης και σταθερότητα, ενώ η LR επιβεβαιώνει τον ρόλο της στιβαρής γραμμής βάσης με πλεονέκτημα ερμηνευσιμότητας. Η εικόνα αυτή είναι σύμφωνη με μελέτες που δείχνουν ότι οι δενδροειδείς μέθοδοι και τα ensembles πιάνουν μη γραμμικότητες και αλληλεπιδράσεις ανάμεσα σε συμπεριφορικές μεταβλητές (π.χ. δαπάνη-συχνότητα-κανάλι), την ώρα που τα γραμμικά μοντέλα προσφέρουν διαφάνεια και αξιόπιστη βαθμονόμηση πιθανοτήτων ως σημείο αναφοράς.

Σε περιβάλλον σπάνιας θετικής τάξης, τυπικό στο response modeling, η βιβλιογραφία προκρίνει την PRC–AUC για να αναδειχθεί το trade-off precision–recall, αντί της αποκλειστικής εστίασης στη ROC. Η παρούσα χρήση precision, recall και F1 (και όπου διατέθηκε ROC–AUC) συμβαδίζει με αυτή τη γραμμή, ενώ η έμφαση στην ανάκτηση θετικών περιπτώσεων (recall) είναι συμβατή με πρακτικές *direct marketing* που

αποτιμούν το κόστος ευκαιρίας ενός FN (χαμένος δυνητικός αποδέκτης). Υπό αυτό το πρίσμα, η παρατηρούμενη ισορροπία του SVM και το ισχυρό ranking των δενδροειδών ensembles είναι ακριβώς ό,τι αναμένει κανείς από την διεθνή βιβλιογραφία.

Αναλυτικότερα, τα δενδροειδή και ensembles τείνουν να υπερέχουν σε AUC έναντι μοντέλων πρώτης γραμμής (LR, NB), καθώς αποτυπώνουν αλληλεπιδράσεις χωρίς ρητή μηχανική χαρακτηριστικών. Το ET αναφέρεται συχνά ως ταχύτερο, με απόδοση συγκρίσιμη του RF, μοτίβο που παρατηρήθηκε εν μέρει. Το SVM, ιδίως σε μεσαίο μέγεθος δειγμάτων και με ορθή κλιμάκωση, επιτυγχάνει ομαλές ROC και ισορροπημένο precision–recall, επαληθεύοντας το βιβλιογραφικό του πλεονέκτημα. Ο KNN από την άλλη, τείνει να γέρνει προς υψηλότερο recall εις βάρος του precision σε δεδομένα με σαφή τοπικές δομές, γνωστό αποτέλεσμα της γειτνίασης και της εξάρτησης από την κλίμακα των χαρακτηριστικών. Ο NB λειτουργεί προβλέψιμα ως γρήγορο baseline, αλλά υπολείπεται όταν οι υπό συνθήκη ανεξαρτησίες ή κανονικότητες δεν τηρούνται, όπως συμβαίνει συχνά σε ετερογενή λιανικά σύνολα. Τα boosting μοντέλα AB και GB επιβεβαιώνουν τη βιβλιογραφική τους φήμη ως ισχυρές επιλογές σε μη γραμμικότητες υπό προσεκτικό tuning, ενώ τα νευρωνικά διατηρούν συμπληρωματικό ρόλο όταν ο όγκος και η ποικιλία χαρακτηριστικών δεν επαρκεί για να υπερκεράσουν τα δασικά ensembles.

Ως προς τους παράγοντες που οδηγούν τις προβλέψεις, η βιβλιογραφία σε RFM και omni-channel δεδομένα, συνδέει σταθερά την ανταπόκριση με δείκτες πρόσφατης δραστηριότητας (recency), επίπεδα δαπάνης και μετρικές εμπλοκής σε web κανάλια. Οι παρούσες οπτικές ενδείξεις EDA κινούνται στην ίδια κατεύθυνση. Δημογραφικά στοιχεία τείνουν να έχουν δευτερεύον ρόλο ή να απαιτούν προσεκτική χρήση για λόγους fairness, σημείο που επίσης αναδεικνύεται στη βιβλιογραφία και τοποθετείται ρητά στα επιχειρησιακά ζητήματα.

Τα σημεία απόκλισης εντοπίζονται στην εμφάνιση υπερβολικά υψηλών σκορ σε εναλλακτική ροή αξιολόγησης. Η βιβλιογραφία είναι σαφής, καθώς strict validation (nested CV και καθαρό hold-out) και ενιαία pipelines που αποτρέπουν διαρροή είναι προϋποθέσεις αξιοπιστίας. Εφόσον τα πολύ υψηλά σκορ συμπίπτουν με ροές όπου προεπεξεργασία και oversampling γίνονται εκτός fold ή πριν το split, η πιο εύλογη ερμηνεία είναι μεθοδολογική και όχι ουσιαστική υπεροχή μοντέλου, ακριβώς όπως προειδοποιούν οι καλύτερες πρακτικές εφαρμογής.

Καταλήγοντας, η διεθνής συζήτηση μετακινείται από απλή πιθανότητα απόκρισης σε uplift και causal μοντελοποίηση για εκτίμηση καθαρής επίδρασης επαφής. Αρκετές μελέτες το συνιστούν, αλλά σπανίως εφαρμόζεται σε πραγματικά retail datasets. Η παρούσα εργασία παραμένει προγνωστική, στοιχείο που τη φέρνει σε μερική απόκλιση από την αιτιώδη κατεύθυνση και αναδεικνύει ως επόμενο βήμα την ενσωμάτωση cost-sensitive κατωφλιών, calibration και πειραματικών επικυρώσεων. Συνολικά, το μοτίβο αποτελεσμάτων είναι συνεπές με τη βιβλιογραφία για retail response. Οι όποιες αποκλίσεις εξηγούνται από ρυθμίσεις επικύρωσης και ενισχύουν την ανάγκη για αυστηρό pipeline και αξιολόγηση προσανατολισμένη σε κόστος–όφελος.

6.4 Επιχειρησιακές συνέπειες

Το προβλεπτικό σκορ $\hat{p}$ ερμηνεύεται ως πιθανότητα ανταπόκρισης και η πολιτική στόχευσης μεταφράζει το σκορ σε πράξη. Η επιχειρησιακή απόφαση λαμβάνεται με δύο ισοδύναμους τρόπους. Είτε με τον κανόνα κατωφλίου, όπου επιβεβαιώνεται η επικοινωνίας με τον πελάτη όταν $\hat{p} \geq \tau$, ή με την top-K προσέγγιση, όπου ιεραρχούνται όλοι οι πελάτες κατά $\hat{p}$ και στοχεύονται οι K κορυφαίοι βάσει διαθέσιμου budget. Και στις δύο περιπτώσεις προηγούνται αποκλεισμοί από τη διαδικασία (opt-out, suppression lists, «μην ενοχλείτε», πρόσφατα service cases) και fatigue rules (π.χ. μέγιστο 1 επαφή ανά 7 ημέρες ή κανάλι, έως 2 τον μήνα συνολικά κ.ο.κ.), ώστε να περιορίζεται η φθορά μάρκας και ο κίνδυνος churn από υπερβολική προσπάθεια επικοινωνίας. Η τμηματοποίηση (δημογραφική ή συμπεριφορική) λειτουργεί επικουρικά. Το ίδιο $\hat{p}$ μπορεί να οδηγεί σε διαφορετικό μήνυμα και προσφορά ανά τμήμα (π.χ. value-seekers vs premium explorers).

Το κατώφλι απόφασης με κόστος – όφελος, είθισται να ορίζει και την επιχειρηματική απόφαση. Χωρίς πραγματικές τιμές κόστους ωστόσο, η επιλογή κατωφλιού τ δεν μπορεί να πάρει τιμή εδώ. Πάρα ταύτα ο κανόνας είναι σαφής. Αρχικά ορίζεται ανά επαφή

Benefit (B), ως προσδοκώμενο καθαρό όφελος από έναν αληθώς θετικό (αγορά μείον έκπτωση & λειτουργικά).

Contact cost (C), ως κόστος καναλιού (email = αμελητέο ανά μονάδα, αλλά με κρυφό κόστος brand fatigue, ενώ SMS και τηλέφωνο είναι ακριβότερα).

Για οποιοδήποτε κατώφλι τ υπολογίζεται από τις προβλέψεις test TP(τ), FP(τ), #Contacted(τ) = TP + FP και εκτιμάται το Profit=TP·π−FP·c−Contacts·$c_{channel}$

Σαρώνεται το κατώφλι σε διάστημα [0,1] όπου το τ∈[0,1] (ή ισοδύναμα ποσοστιαία Top-K), επιλέγοντας τ που μεγιστοποιεί το κέρδος ή ικανοποιεί περιορισμούς (π.χ. FP-rate ≤ στόχος καναλιού). Εάν το FN είναι ακριβό (χαμένος δυνητικός τζίρος ή CLV), επιλέγεται χαμηλότερο τ (υψηλότερο recall). Αν το FP είναι ακριβό (τηλεφωνικό κανάλι, συμφόρηση contact center), αυξάνεται το τ (υψηλότερο precision). Η επιλογή τ γίνεται αφότου ελεγχθεί η βαθμονόμηση πιθανοτήτων, αφού κακώς καλιμπραρισμένα scores οδηγούν σε λάθος οικονομικές αποφάσεις.

Ό,τι αφορά την επιλογή καναλιού προώθησης της προσφοράς και τους κανόνες επικοινωνίας που το διέπουν, αυτά ιεραρχούνται από φθηνά ή ευρείας κάλυψης σε ακριβά ή υψηλού lift. Ενδεικτικό κανάλι ανά ζώνη με βάση το $\hat{p}$ είναι το ακόλουθο:

$0.60 \leq \hat{p} < 0.75$. Email/Push (χαμηλό C, έλεγχος συχνότητας, A/B σε subject και ώρα αποστολής).

$0.75 \leq \hat{p} < 0.85$. SMS (μεσαίο C, μεσαίο lift, περιορισμός συχνότητας).

$\hat{p} \geq 0.85$. Τηλέφωνο ή assisted sales (υψηλό C, υψηλό lift, αυστηρός όγκος).

Για τη συνολική εμπειρία πελάτη εφαρμόζονται κανόνες frequency capping και «ψύξης» (π.χ. retargeting floor, αν $\hat{p} < 0.35$ μετά από 2 μη αποδοτικές επαφές, παύση 30 ημερών). Έτσι μειώνεται ο κορεσμός του πελάτη και προστατεύονται NPS και άλλα brand metrics.

Για το μέγεθος της παρτίδας τα πράγματα είναι σχετικά απλά. Με budget B και κόστος καναλιού C, ο μέγιστος όγκος επαφών είναι K = B/C. Η διαδικασία ορίζει ταξινόμηση βάσει $\hat{p}$ στοχεύοντας στους top-k. Παρακολουθείτε το marginal ROI ανά decile του $\hat{p}$ και σταματά όταν το οριακό ROI προσεγγίζει το μηδέν. Στην πράξη προτιμώνται κύματα, δηλαδή επιμερισμός των deciles ανά εβδομάδα (π.χ. top-10% $1^{η}$ εβδομάδα, 10–20% $2^{η}$ εβδομάδα), ώστε να προκύπτει αξιολόγηση και να γίνεται μάθηση από μετατροπή του πελάτη σε πραγματικό χρόνο (*live conversion*) και να αναπροσαρμόζεται το κατώφλι και το κανάλι επίσης σε πραγματικό χρόνο.

Σαφώς θα πρέπει να τηρείται η ηθική δεοντολογία κατά την διαδικασία προώθησης.

Πριν από κάθε κύκλο εκτέλεσης θα πρέπει να προηγούνται opt-in και consent checks, ενημερωμένες suppression lists, έλεγχος privacy. Σε κάθε έκδοση μοντέλου γίνονται fairness checks ανά υποομάδα (π.χ. ηλικιακά κλιμάκια, σύνθεση νοικοκυριού). Έτσι, συγκρίνονται precision, recall και FNR parity και αναθεωρούνται features, κατώφλι και το κανάλι αν εντοπίζονται ασυμμετρίες. Για αποφάσεις αποκλεισμού και προνομίων, φυλάσσονται εξηγήσεις (global importances, τοπικά SHAP) για εσωτερικούς ελέγχους και συμμόρφωση. Με αυτόν τον τρόπο αποφεύγονται αθέλητες διακρίσεις και περιορίζεται ο κίνδυνος νομικής ή δυσφημιστικής έκθεσης.

Όπως αναφέρθηκε και παραπάνω, η μέτρηση της απόδοσης είναι κρίσιμη πριν τρέξει σε πραγματικό χρόνο το εγχείρημα, οπότε είναι απαραίτητη η πιλοτική ενεργοποίησή του. Η πρώτη διάθεση γίνεται πειραματικά με Control (κανόνας βάσης) vs Model-based βάσει κατωφλίου ή Top-k υποψηφίους. Μετράται uplift (Δ-conversion), net profit per contacted, CPA και spillover ανά κανάλι ή χρόνο. Τίθεται Stop και Go κανόνες με ελάχιστο δείγμα για ισχύ και όπου απαιτείται, sequential testing για γρήγορες αποφάσεις όταν το σήμα είναι ισχυρό. Παράλληλα, παρακολουθείται brand KPIs (π.χ. unsubscribe, complaint rate κτλ.), καθώς ακόμη και φθηνά κανάλια όπως το email, αν και αμελητέα σε λογιστικό κόστος ανά επαφή, φέρουν στρατηγικό κόστος σε brand fatigue, αρνητική προδιάθεση και μελλοντική propensity to churn όταν γίνεται κατάχρηση.

6.5 Μελλοντική ενασχόληση

Για να μετατραπεί το τρέχον διαβλητό concept σε ανθεκτικό, επιχειρησιακό σύστημα πρόβλεψης ανταπόκρισης, προτείνεται ένα πλάνο τεσσάρων επιπέδων. Πρώτο επίπεδο είναι η αυστηροποίηση μεθοδολογίας και επικύρωσης. Δέυτερο η βελτίωση πιθανοτήτων και κανόνα απόφασης. Τρίτον η ενίσχυση μοντελοποίησης και χαρακτηριστικών. Τέταρτον η παραγωγή, διακυβέρνηση και πειραματισμός.

Κατά την αυστηροποίηση της μεθοδολογίας, η ροή πρέπει να αναδιαταχθεί σε ενιαίο pipeline με ColumnTransformer για όλους τους μετασχηματισμούς και SMOTE μόνο στο training μέσω imblearn.Pipeline. Έτσι εξαλείφεται η πιθανότητα leakage από προεπεξεργασία εκτός fold. Η επιλογή υπερπαραμέτρων να αποσυνδεθεί από την τελική αποτίμηση με nested cross-validation. Όπου υπάρχουν χρονικές εξαρτήσεις, να χρησιμοποιηθεί time-based split και rolling CV. Κάθε μοντέλο να συνοδεύεται

συστηματικά από PRC–AUC, Brier score και calibration curves, ώστε να αξιολογείται τόσο η διάκριση όσο και η αξιοπιστία πιθανοτήτων.

Για τη βελτίωση πιθανοτήτων και κατωφλίου, θα πρέπει πριν από οποιαδήποτε επιχειρησιακή κατωφλίωση, να εφαρμόζεται calibration σε ανεξάρτητο σύνολο. Ο κανόνας απόφασης να οριστικοποιείται με σάρωση κατωφλιών ή Top-K και profit curve βάσει πίνακα κόστους–οφέλους (TP benefit, contact cost, ενδεχομένως penalty για FP). Να εξεταστούν cost-sensitive τεχνικές (class weights, utility-aware losses) και να παρουσιάζονται decision και cost curves για διαφάνεια στην επιλογή τ.

Τα μοντέλα και οι συνδυασμοί, πέρα από SVM και RF, θα πρέπει να δοκιμαστούν και άλλα ευρέως εφαρμοσμένα μοντέλα Gradient Boosting όπως XGBoost / LightGBM / CatBoost με προσεκτικό tuning και stacking (π.χ. LR ως meta-learner πάνω σε SVM, RF και GB). Για νευρωνικά, να υιοθετηθεί binary head (sigmoid) με dropout και regularization, early stopping και learning-rate προγράμματα. Όπου η απόφαση είναι ευαίσθητη στο ρίσκο, να διερευνηθεί τεχνική ποσοτικοποίησης της αβεβαιότητας (π.χ. MC-dropout) για κανόνες τύπου «επικοινωνώ μόνο όταν η αβεβαιότητα είναι χαμηλή». Σαφώς και αποτελεσματικοί συνδυασμοί μοντέλων σε περισσότερο προχωρημένο επίπεδο δύνανται να δώσουν ένα ευστοχότερο αποτέλεσμα.

Για την περαιτέρω αξιοποίηση των χαρακτηριστικών και των δεδομένων, θα πρέπει να εμπλουτιστεί το σύνολο με χρονικά features (lags, rolling windows), RFM και LRFMP και δείκτες όπως κανάλι-συχνότητα-αξία. Να ενσωματωθούν εξωτερικές μετρικές (όπως ημερολόγιο, εποχικότητα, περίοδοι προσφορών κ.ά.) και επιλεγμένες αλληλεπιδράσεις (π.χ. εισόδημα-κατηγορία δαπάνης). Οποιαδήποτε παραγωγή παραγώγων θα πρέπει να περνά από έλεγχο leakage. Παράλληλα, να εξεταστεί segmentation-first (μοντέλα ανά τμήμα, π.χ. high-value vs low-value) όταν εντοπίζονται ετερογένειες συμπεριφοράς.

Άξια μελλοντικής ενασχόλησης επίσης αποτελεί η αιτιώδης επίδραση και uplift. Για να αποδοθεί σωστά το incremental κέρδος στο μοντέλο (και όχι σε επιλογές καναλιού και προσφοράς), να μετακινηθεί ο στόχος από response σε uplift. Προτείνονται uplift learners (όπως οι T-, X-learner, causal forests) και πειραματισμοί ομάδων A/B ή ισοδύναμα σχήματα (geo-split, staggered rollout). Έτσι βελτιστοποιείται απευθείας η αυξητική αποτελεσματικότητα της καμπάνιας.

Ακόμη προτείνεται να οργανωθεί μοντελοθήκη με versioning (κώδικας–δεδομένα–μοντέλο), audit logs και model cards. Σε παραγωγή να παρακολουθείται drift (data, label, PSI, KS, performance drift) με alerts και ρητούς re-training triggers (π.χ. μηνιαίος κύκλος ή όταν PRC–AUC < στόχος). Να υλοποιηθεί CI και CD για ασφαλή αναβαθμίσεις. Σε κάθε έκδοση προτείνεται επίσης να εκτελούνται fairness checks ανά υποομάδα και να τεκμηριώνονται global και local εξηγήσεις (SHAP). Όπου τα δημογραφικά δεν είναι αναγκαία, να περιορίζεται η χρήση τους ή να εφαρμόζονται fairness constraints. Έτσι θα πρέπει να θεσπιστούν σταθεροί κανόνες καναλιών (fatigue, opt-out, suppression lists) και privacy by design.

Για την επιχειρησιακή βελτιστοποίηση επιπρόσθετα, καλούνται να κλειδώσουν κανόνες καναλιού βάσει ROI, batch sizing από budget και χωρητικότητα και sequential testing για ταχύτερες αποφάσεις. Στο ίδιο πλαίσιο, κρίνεται σκόπιμο να υλοποιηθούν dashboards με marginal ROI ανά decile πρόβλεψης και calibration-aware thresholding σε πραγματικό χρόνο. Με την υιοθέτηση όλων των παραπάνω, η επόμενη έκδοση του συστήματος θα διαθέτει επιστημονική αυστηρότητα (χωρίς leakage, με σωστή επικύρωση και αξιόπιστες πιθανότητες) και επιχειρησιακή αποτελεσματικότητα (κατώφλι βάσει κέρδους, κανόνες καναλιού, A/B μέτρηση, MLOps). Έτσι, η πρόβλεψη μετατρέπεται σε βιώσιμη απόφαση με μετρήσιμο αυξανόμενο αντίκτυπο, αναβαθμίζοντας το μοντέλο στο σύνολό του.

ΚΕΦΑΛΑΙΟ 7. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η παρούσα εργασία ξεκίνησε από ένα πρακτικό ερώτημα του λιανικού εμπορίου, που αναζητά απάντηση στο πώς μετατρέπεται η ιστορία συναλλαγών και τα δημογραφικά στοιχεία σε αξιόπιστη πρόβλεψη ανταπόκρισης, ικανή να καθοδηγήσει στοχευμένες καμπάνιες με μετρήσιμο όφελος. Με μια καθαρή, επαναλήψιμη μεθοδολογία καθαρισμού, κωδικοποίησης, κλιμάκωσης και εκτέλεσης πολλαπλών ταξινομητών υπό κοινές μετρικές και συγκριτική αποτίμηση, απεδείχθη ότι η πρόβλεψη είναι όχι μόνο εφικτή αλλά και επιχειρησιακά χρήσιμη, όταν συνδέεται ρητά με κανόνες λήψης απόφασης (κατώφλια, Top-K, επιλογή καναλιού).

Σε επίπεδο μοντελοποίησης, τα μοντέλα με ομαλή και πλούσια συνάρτηση score, ιδίως SVM και δενδροειδή ensembles, ανέδειξαν υψηλή ικανότητα διάκρισης και ισορροπημένο precision recall στο test set, επιβεβαιώνοντας ότι η μη γραμμική δόμηση αλληλεπιδράσεων (π.χ. πρόσφατη δραστηριότητα-ένταση δαπάνης) προσφέρει ουσιαστικό στατιστικό μέγεθος έναντι της γραμμικής βάσης. Ως τελικό μοντέλο προκρίθηκε το SVM λόγω της σταθερότητάς του, της καλής συμπεριφοράς σε όλο το φάσμα κατωφλιών και της άμεσης παραγωγικής αξιοποίησης. Η διερευνητική ανάλυση και οι συγκριτικές επιδόσεις συγκλίνουν στο ότι μεταβλητές όπως η Recency και οι δείκτες δαπάνης και δραστηριότητας λειτουργούν ως βασικοί οδηγοί της πρόβλεψης, παρέχοντας παράλληλα ξεκάθαρα σημεία παρέμβασης για το μάρκετινγκ.

Τα συμπεράσματα αυτά συνοδεύονται από νηφάλια αποτίμηση περιορισμών. Η ύπαρξη βημάτων προεπεξεργασίας και oversampling πριν τον διαχωρισμό train και test επιφέρει κίνδυνο αισιόδοξης εκτίμησης (leakage), ενώ η χρήση του default κατωφλίου 0.5 δεν αποτυπώνει πλήρως τα επιχειρησιακά trade-offs. Η απουσία συστηματικής βαθμονόμησης πιθανοτήτων και πλήρους PRC–AUC για όλα τα μοντέλα περιορίζει την ακριβή μετάφραση πιθανοτήτων σε κανόνες απόφασης. Ακόμη, η γενίκευση των ευρημάτων σε άλλες περιόδους ή γεωγραφικές προσεγγίσεις απαιτεί προσοχή στη μετατόπιση των δεδομένων.

Παρά ταύτα, η εργασία προσφέρει ένα χρηστικό πλαίσιο εφαρμογής. Οι προβλεπόμενες πιθανότητες μετατρέπονται σε στοχευμένες λίστες (threshold και Top-K), οι καμπάνιες χαρτογραφούνται σε κανάλια με διαβαθμισμένο κόστος (email, SMS, τηλέφωνο) και ορίζονται κανόνες διακυβέρνησης (consent, suppression, frequency

capping) ώστε να προστατεύεται η εμπειρία πελάτη και να ελαχιστοποιείται η φθορά μάρκας. Η πρακτική αξία ενισχύεται από ένα σαφές *playbook* ενεργοποίησης που περιλαμβάνει πιλοτική μέτρηση (A/B), παρακολούθηση marginal ROI ανά decile πρόβλεψης και περιοδική επαναπροσαρμογή κατωφλίου σύμφωνα με τον προϋπολογισμό και το κόστος καναλιού.

Η φυσική συνέχεια του έργου είναι μεθοδολογική ωρίμανση και επιχειρησιακή ενσωμάτωση. Αυτό προϋποθέτει ενιαίο pipeline χωρίς διαρροή, nested και χρονική επικύρωση, calibration πιθανοτήτων και επιλογή κατωφλίου βάσει κέρδους, καθώς και διερεύνηση uplift και αιτιώδους μοντελοποίησης ώστε η στόχευση να βελτιστοποιεί απευθείας το incremental lift. Σε επίπεδο λειτουργίας, απαιτείται πειθαρχημένο MLOps (monitoring drift, triggers re-training, versioning) και τακτικοί έλεγχοι δικαιοσύνης και εξηγησιμότητας ανά υποομάδα πελατών.

Η κατακλείδα είναι πως η μελέτη αποδεικνύει ότι η πρόβλεψη ανταπόκρισης μπορεί να αποτελέσει στιβαρό άξονα λήψης αποφάσεων στο σούπερ μάρκετ, όταν συνδυάζεται με αυστηρή μεθοδολογία και κανόνες ενεργοποίησης προσανατολισμένους στη σχέση κόστους με όφελος. Με την προτεινόμενη κατεύθυνση βελτίωσης, το μοντέλο μετατρέπεται από αλγοριθμική άσκηση σε λειτουργικό εργαλείο αύξησης εσόδων, μετριασμού κόστους και αναβάθμισης της πελατειακής εμπειρίας. Όπως κάθε εργαλείο έτσι και αυτό, με ορθή χρήση και τακτική συντήρηση δύναται να δώσει απαντήσεις στις σημαντικότερες προκλήσεις και να υποστηρίξει ουσιαστική λήψη αποφάσεων στην επιχειρησιακή ανέλιξη των διαφόρων τομέων του λιανικού εμπορίου και όχι μόνο του μάρκετινγκ.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ajiga, D. I., Ndubuisi, N. L., Asuzu, O. F., Owolabi, O. R., Tubokirifuruar, T. S., & Adeleye, R. A. (2024). AI-driven predictive analytics in retail: A review of emerging trends and customer engagement strategies. *International Journal of Management & Entrepreneurship Research, 6*(2), 307–321. https://doi.org/10.51594/ijmer.v6i2.772

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2013). Classification with class imbalance problem: A review. *International Journal of Advance Soft Computing and Applications, 5*(3). https://www.researchgate.net/publication/288228469_Classification_with_class_imbalance_problem_A_review

Allaway, A. W., Huddleston, P., Whipple, J., & Ellinger, A. (2011). Customer-based brand equity, equity drivers, and customer loyalty in the supermarket industry. *Journal of Product & Brand Management, 20*(3), 190–204. https://doi.org/10.1108/10610421111134923

Assemlali, H., Bouhsissin, S., & Sael, N. (2025). Deep learning-driven CNN model for detection and classification of dynamic obstacles. *Green Energy and Intelligent Transportation*. Advance online publication (Article 100334). https://doi.org/10.1016/j.geits.2025.100334

Bhandari, S., Lencastre, P., Denisov, S., Bystryk, Y. S., & Lind, P. G. (2025). IntLevPy: A Python library to classify and model intermittent and Lévy processes. *SoftwareX, 31*, 102334. https://doi.org/10.1016/j.softx.2025.102334

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research, 13*, 1063–1095. https://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks, 106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

de Mooij, M., & Hofstede, G. (2002). Convergence and divergence in consumer behavior: Implications for international retailing. *Journal of Retailing, 78*(1), 61–69. https://doi.org/10.1016/S0022-4359(01)00067-7

Elreedy, D., Atiya, A. F., & Kamalov, F. (2024). A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE). *Machine Learning, 113*, 4903–4923. https://doi.org/10.1007/s10994-022-06296-4

Esfandiari Bahraseman, S., Dehghani Dashtabi, M., Firoozzare, A., Boccia, F., Pakook, S., & Covino, D. (2025). Understanding consumer behavior in the choice of healthy food retail outlets: An examination of information types and the interplay between institutional trust and social recommendations. *Economic Analysis and Policy, 86*, 2070–2094. https://doi.org/10.1016/j.eap.2025.05.035

García-Martín, D., Larocca, M., & Cerezo, M. (2025). Quantum neural networks form Gaussian processes. *Nature Physics, 21*, 1153–1159. https://doi.org/10.1038/s41567-025-02883-z

Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., & Japkowicz, N. (2024). The class imbalance problem in deep learning. *Machine Learning, 113*, 4845–4901. https://doi.org/10.1007/s10994-022-06268-8

Goldmann, S. H., Machado, M. R., & Osterrieder, J. R. (2025). Advancing credit risk assessment in the retail banking industry: A hybrid approach using time series and supervised learning models. *Data & Knowledge Engineering, 160*, 102490. https://doi.org/10.1016/j.datak.2025.102490

Hagberg, J., Sundström, M., & Egels-Zandén, N. (2016). The digitalization of retailing: An exploratory framework. *International Journal of Retail & Distribution Management*. https://doi.org/10.1108/IJRDM-09-2015-0140

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing

K-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications. *Journal of Big Data, 11*, 113. https://doi.org/10.1186/s40537-024-00973-y

Hardin, D. P., Tsamardinos, I., & Aliferis, C. F. (2004). A theoretical characterization of linear SVM-based feature selection. *Conference paper*. https://doi.org/10.1145/1015330.1015421

Herzallah, F., Abosamaha, A. J., Salameh, S. M., & Alhayek, M. (2023). Social commerce attributes, customer engagement and repurchase intention in social commerce platforms: A stimulus–organism–response approach. *Information & Management, 60*, 103808. https://doi.org/10.1016/j.im.2023.103808

Heydarian, M., Doyle, T. E., & Samavi, R. (2022). MLCM: Multi-label confusion matrix. *IEEE Access, 10*, 19083–19094. https://doi.org/10.1109/ACCESS.2022.3151048

Htike, Z. H. M. (2017). Efficient determination of the number of weak learners in AdaBoost. *Journal of Experimental & Theoretical Artificial Intelligence, 29*(5), 967–982. https://doi.org/10.1080/0952813X.2016.1266038

Imakura, A., Kihira, M., Okada, Y., & Sakurai, T. (2023). Another use of SMOTE for interpretable data collaboration analysis. *Expert Systems with Applications, 228*, 120385. https://doi.org/10.1016/j.eswa.2023.120385

Jiang, L., Cai, Z., Zhang, H., & Wang, D. (2013). Naive Bayes text classifiers: A locally weighted learning approach. *Journal of Experimental & Theoretical Artificial Intelligence, 25*(2), 273–286. https://doi.org/10.1080/0952813X.2012.721010

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data, 6*, 27. https://doi.org/10.1186/s40537-019-0192-5

Jung, Y., Qian, C., Barnett-Neefs, C., Ivanek, R., & Wiedmann, M. (2024). Developing an agent-based model that predicts *Listeria* spp. transmission to assess *Listeria* control

strategies in retail stores. *Journal of Food Protection, 87*, 100337. https://doi.org/10.1016/j.jfp.2024.100337

Klabunde, M., Schumacher, T., Strohmaier, M., & Lemmerich, F. (2025). Similarity of neural network models: A survey of functional and representational measures. *ACM Computing Surveys, 57*(9), Article 242. https://doi.org/10.1145/3728458

Khan, T. A., Sadiq, R., Shahid, Z., Alam, M. M., & Mohd Su'ud, M. B. (2024). Sentiment analysis using support vector machine and random forest. *Journal of Informatics and Web Engineering, 3*(1), 68–75. https://doi.org/10.33093/jiwe.2024.3.1.5

Kumar, R., Ramesh, J. V. N., Mohanty, S. N., Rafiq, M., Shernazarov, I., & Khan, M. I. (2025). Evaluating consumers' benefits in electronic-commerce using fuzzy TOPSIS. *Results in Control and Optimization, 18*, 100535. https://doi.org/10.1016/j.rico.2025.100535

Li, J. (2024). Area under the ROC curve has the most consistent evaluation for binary classification. *PLOS ONE, 19*(12), e0316019. https://doi.org/10.1371/journal.pone.0316019

Li, Y., García-de-Frutos, N., & Ortega-Egea, J. M. (2025). Impulse buying in live streaming e-commerce: A systematic literature review and future research agenda. *Computers in Human Behavior Reports, 19*, 100676. https://doi.org/10.1016/j.chbr.2025.100676

Ma, Y., Zhang, Y., & Zheng, H. (2025). Data-driven and people-centric operations: Research opportunities in retail operations. *Fundamental Research*. Advance online publication. https://doi.org/10.1016/j.fmre.2024.11.028

Madden, W. G., Jin, W., Lopman, B., Zufle, A., Dalziel, B. D., Metcalf, C. J. E., Grenfell, B. T., & Lau, M. S. Y. (2024). Deep neural networks for endemic measles dynamics: Comparative analysis and integration with mechanistic models. *PLOS Computational Biology, 20*(11), e1012616. https://doi.org/10.1371/journal.pcbi.1012616

Mani, K., & Shenoy, A. K. B. (2025). Machine learning models in web applications: A comprehensive review. *ICT Express*. Advance online publication. https://doi.org/10.1016/j.icte.2025.09.006

Min, H. (2006). Developing the profiles of supermarket customers through data mining. *The Service Industries Journal*, 26(7), 747–763. https://doi.org/10.1080/02642060600898252

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics, 7*, 21. https://doi.org/10.3389/fnbot.2013.00021

Noyan, F., & Golbasi Simsek, G. (2012). A partial least squares path model of repurchase intention of supermarket customers. *Procedia – Social and Behavioral Sciences, 62*, 921–926. https://doi.org/10.1016/j.sbspro.2012.09.156

Ogeawuchi, J. C., Onifade, A. Y., Abayomi, A. A., Agboola, O. A., Dosumu, R. E., & George, O. O. (2022). Systematic review of predictive modeling for marketing funnel optimization in B2B and B2C systems. *IRE Journals, 6*(3). https://www.researchgate.net/publication/392708464_Systematic_Review_of_Predictive_Modeling_for_Marketing_Funnel_Optimization_in_B2B_and_B2C_Systems

Padmanabhan, R., Hadid, M., Alkhiyami, H. W., Elomri, A., & Kerbache, L. (2025). Inventory management of perishable foods in omnichannel retail using predictive analytics: A case study. *IFAC PapersOnLine, 59*(10), 691–696. https://doi.org/10.1016/j.ifacol.2025.09.118

Passos, D., & Mishra, P. (2022). A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks. *Chemometrics and Intelligent Laboratory Systems, 223*, 104520. https://doi.org/10.1016/j.chemolab.2022.104520

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3–14.

https://doi.org/10.1080/00220670209598786

Peretz, O., Koren, M., & Koren, O. (2024). Naive Bayes classifier – An ensemble procedure for recall and precision enrichment. *Engineering Applications of Artificial Intelligence, 136*, 108972. https://doi.org/10.1016/j.engappai.2024.108972

Prasetiyowati, M. I., Maulidevi, N. U., & Surendro, K. (2021). Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest. *Journal of Big Data, 8*, 84. https://doi.org/10.1186/s40537-021-00472-4

Sharma, N., Acquila-Natale, E., Dutta, N., & Hernández-García, Á. (2025). A tale of stores and screens: Unveiling consumer behaviour in omnichannel retailing through the lens of behavioural reasoning. *Electronic Commerce Research and Applications, 70*, 101480. https://doi.org/10.1016/j.elerap.2025.101480

Singh, T., Ali, R., & Tyagi, V. (2025). An efficient identity-based multi-proxy multi-signcryption scheme for electronic commerce using bilinear pairing. *Procedia Computer Science, 259*, 1592–1601. https://doi.org/10.1016/j.procs.2025.04.114

Silverio, F., Cantalupo, M., Custode, L. L., & Iacca, G. (2025). A customer behavior-driven clustering method in the planogram design domain. *Applied Soft Computing, 172*, 112836. https://doi.org/10.1016/j.asoc.2025.112836

Strani, L., Cocchi, M., Tanzilli, D., Biancolillo, A., & Marini, F. (2025). One class classification (class modelling): State of the art and perspectives. *Trends in Analytical Chemistry, 183*, 118117. https://doi.org/10.1016/j.trac.2024.118117

Stylianou, T., & Pantelidou, A. (2025). A machine learning approach to consumer behavior in supermarket analytics. *Decision Analytics Journal, 16,* 100600. https://doi.org/10.1016/j.dajour.2025.100600

Szymański, P., & Kajdanowicz, T. (2019). scikit-multilearn: A scikit-based Python environment for performing multi-label classification. *Journal of Machine Learning*

Research, 20, 1–22. https://www.jmlr.org/papers/volume20/17-100/17-100.pdf

Theodoridis, P. K., & Chatzipanagiotou, K. C. (2009). Store image attributes and customer satisfaction across different customer profiles within the supermarket sector in Greece. *European Journal of Marketing, 43*(5–6), 708–734. https://doi.org/10.1108/03090560910947016

Tsay, R. S. (1998). Testing and modeling multivariate threshold models. *Journal of the American Statistical Association, 93*(443), 1188–1202. https://doi.org/10.1080/01621459.1998.10473779

Wang, L., Pertheban, T. R. A. L., Li, T., & Zhao, L. (2024). Application of business intelligence based on big data in E-commerce data evaluation. *Heliyon, 10*, e38768. https://doi.org/10.1016/j.heliyon.2024.e38768

Wang, S., Dai, Y., Shen, J., & Xuan, J. (2021). Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports*, *11*, 24039. https://doi.org/10.1038/s41598-021-03430-5

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information, 14*, 54. https://doi.org/10.3390/info14010054

Xu, X., Luo, C., Luo, X. (Robert), & Wang, Z. (2024). Examining how emotions affect online audience retention: Empirical evidence from livestreaming electronic commerce platforms. *Information & Management, 61*, 104031 https://doi.org/10.1016/j.im.2024.104031

Yokoyama, N., Azuma, N., & Kim, W. (2022). Moderating effect of customer's retail format perception on customer satisfaction formation: An empirical study of mini-supermarkets in an urban retail market setting. *Journal of Retailing and Consumer Services, 66*, 102935. https://doi.org/10.1016/j.jretconser.2022.102935

Zaidi, A., & Al Luhayb, A. S. M. (2023). Two statistical approaches to justify the use of the logistic function in binary logistic regression. *Mathematical Problems in Engineering, 2023*, Article 5525675. https://doi.org/10.1155/2023/5525675

Zhang, H., & Su, J. (2008). Naive Bayes for optimal ranking. *Journal of Experimental & Theoretical Artificial Intelligence, 20*(2), 79–93. https://doi.org/10.1080/09528130701476391