

**ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ – ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ****Πρόγραμμα Μεταπτυχιακών Σπουδών****«Κυβερνοασφάλεια και Επιστήμη Δεδομένων»****Μεταπτυχιακή Διατριβή**

Τίτλος Διατριβής	Επιχειρησιακή Αναλυτική στην Πράξη: Τμηματοποίηση και Πρόβλεψη Αποχώρησης Πελατών στον Κλάδο Ηλεκτρονικής Διανομής Φαγητού Business Analytics in Practice: Customer Segmentation and Churn Prediction in the Online Food Delivery Industry
Ονοματεπώνυμο Φοιτητή	Βασίλειος Μητρόπουλος
Πατρώνυμο	Κωνσταντίνος
Αριθμός Μητρώου	ΜΠΚΕΔ2325
Επιβλέπων	Διονύσιος Σωτηρόπουλος, Αναπληρωτής Καθηγητής

Ημερομηνία Παράδοσης: **Μάιος 2026**

Στην ολοκλήρωση της παρούσας μεταπτυχιακής διατριβής, ιδιαίτερα σημαντική ήταν η συμβολή του Διδάσκοντα του ΠΜΣ κ. Ανδρέα Ζάρα, ο οποίος προσέφερε επιστημονική και συμβουλευτική καθοδήγηση σε όλα τα στάδια εκπόνησής της.

Τριμελής Εξεταστική Επιτροπή

Διονύσιος Σωτηρόπουλος,
Αναπληρωτής Καθηγητής

Άγγελος Πικράκης,
Αναπληρωτής Καθηγητής

Δημήτρης Αποστόλου,
Καθηγητής

Ευχαριστίες

Θα ήθελα να εκφράσω την ειλικρινή μου εκτίμηση προς τον κ. Ανδρέα Ζάρα για την καθοδήγηση, τα σχόλια και την εξειδίκευσή του σε όλη τη διαδικασία έρευνας και συγγραφής. Η αφοσίωσή του στην προώθηση της κριτικής σκέψης, της επεξηγηματικής προσέγγισης και της ακαδημαϊκής αριστείας έπαιξε καθοριστικό ρόλο στη διαμόρφωση της ποιότητας της παρούσας μεταπτυχιακής διατριβής.

Τέλος, θα ήθελα να αφιερώσω την παρούσα εργασία στην οικογένειά μου, η οποία στάθηκε δίπλα μου σε κάθε στάδιο αυτής της διαδρομής. Η στήριξη, η υπομονή και η πίστη τους σε εμένα αποτέλεσαν καθοριστικό παράγοντα για την ολοκλήρωση αυτής της προσπάθειας. Χωρίς τη δική τους παρουσία και ενθάρρυνση, η επίτευξη αυτού του στόχου δεν θα ήταν δυνατή.

Περίληψη

Η παρούσα μεταπτυχιακή διατριβή επικεντρώνεται στην εφαρμογή τεχνικών ανάλυσης δεδομένων σε πρακτικό επιχειρησιακό πλαίσιο, μελετώντας την περίπτωση μιας διαδικτυακής πλατφόρμας παράδοσης φαγητού. Στόχος της εργασίας είναι η εξαγωγή αξιοποιήσιμων συμπερασμάτων για τη βελτίωση της διαχείρισης συνεργατών και της συνολικής λειτουργικής απόδοσης, μέσω της ανάλυσης δεδομένων που αφορούν καταστήματα και την εμπορική τους δραστηριότητα στην πλατφόρμα. Στο πλαίσιο αυτό εφαρμόζονται μέθοδοι μηχανικής μάθησης για (α) την ομαδοποίηση καταστημάτων, με σκοπό την ανάδειξη διακριτών προφίλ και προτύπων συμπεριφοράς/απόδοσης, και (β) την πρόβλεψη αποχώρησης καταστημάτων, ώστε να εντοπίζονται έγκαιρα συνεργάτες αυξημένου κινδύνου και να υποστηρίζονται στοχευμένες δράσεις διατήρησης. Η ανάλυση αναδεικνύει τα χαρακτηριστικά που διαφοροποιούν τα καταστήματα μεταξύ των ομάδων και τους παράγοντες που σχετίζονται με την πιθανότητα αποχώρησης, παρέχοντας προτάσεις για προσαρμογή πολιτικών προμηθειών, προωθητικών ενεργειών και πρακτικών υποστήριξης συνεργατών. Επιπλέον, συζητούνται ζητήματα ποιότητας και συνέπειας δεδομένων, η ερμηνευσιμότητα των μοντέλων και οι ηθικές διαστάσεις της αξιοποίησης δεδομένων σε ψηφιακά περιβάλλοντα. Τα ευρήματα της εργασίας υπογραμμίζουν τη σημασία της αναλυτικής προσέγγισης στη λήψη αποφάσεων για πλατφόρμες εστίασης, παρέχοντας ένα πλαίσιο για την πρακτική εφαρμογή της ομαδοποίησης και της πρόβλεψης αποχώρησης καταστημάτων με στόχο την ενίσχυση της απόδοσης και της ανταγωνιστικότητας.

Abstract

This thesis focuses on the practical application of data analytics techniques in the context of an online food delivery platform. The study aims to derive actionable insights that support partner management and overall operational efficiency by analyzing data related to stores and their commercial activity on the platform. In this context, machine learning methods are applied for (a) shop clustering, in order to identify distinct store profiles and performance/behavior patterns, and (b) shop churn prediction, to proactively detect partners at increased risk of leaving the platform and enable targeted retention actions. The analysis highlights the key characteristics that differentiate stores across clusters and the factors associated with churn likelihood, providing recommendations for tailoring commission policies, promotional strategies, and partner support practices. Furthermore, the thesis discusses issues related to data quality and consistency, model interpretability, and ethical considerations in data-driven decision-making within digital environments. The findings underscore the value of analytics for strategic decision-making in food service platforms, offering a practical framework for leveraging clustering and store churn prediction to improve performance and competitiveness.

Πίνακας Περιεχομένων

Περίληψη	4
Abstract	4
1. Εισαγωγή	7
2. Θεωρητικό Υπόβαθρο και Βασικές Έννοιες	8
2.1 Μεγάλος Όγκος Δεδομένων (Big Data)	8
2.2 Επιχειρησιακή Ευφυΐα (Business Intelligence) και Επιχειρησιακή Αναλυτική (Business Analytics)	11
2.3 Τεχνητή Νοημοσύνη (Artificial Intelligence)	13
2.4 Πλαίσιο Εφαρμογής: Online Food Delivery και Shop Analytics	14
3. Ανασκόπηση Βιβλιογραφίας (Literature Review)	15
3.1 Τμηματοποίηση Καταστημάτων (Shop Clustering)	15
3.2 Πρόβλεψη Αποχώρησης (Churn Prediction)	16
4. Μεθοδολογικό Πλαίσιο και Αναλυτική Προσέγγιση	17
4.1 Προεπεξεργασία και Προετοιμασία Δεδομένων	17
4.2 Τεχνικές Τμηματοποίησης	18
4.2.1 Αλγόριθμοι τμηματοποίησης	19
4.2.2 Επιλογή αριθμού συστάδων και αξιολόγηση ποιότητας	21
4.3 Μοντέλα Πρόβλεψης Αποχώρησης Καταστημάτων	21
4.3.1 Μοντέλα ταξινόμησης για churn prediction	22
4.3.2 Αξιολόγηση, ανισορροπία κλάσεων, επιλογή κατωφλίου και βαθμονόμηση	26
5. Μελέτη Περίπτωσης μιας Ηλεκτρονικής Πλατφόρμας Διανομής Φαγητού	28
5.1 Περιγραφή και Δομή Συνόλου Δεδομένων	28
5.2 Εξερευνητική Ανάλυση Δεδομένων (EDA)	29
6. Τμηματοποίηση Καταστημάτων (Shop Clustering)	38
6.1 Πειραματική Διαδικασία	38
6.1.1 Υποψήφιες μεταβλητές και κριτήρια επιλογής	38
6.1.2 Correlation analysis & απόρριψη “πλεοναστικών” μεταβλητών	39
6.1.3 Καθαρισμός δεδομένων & χειρισμός ακραίων τιμών (outliers)	44
6.2 Αποτελέσματα Τμηματοποίησης Καταστημάτων	46

6.2.1 Διαδικασία υλοποίησης στο SAS Model Studio	47
6.2.2 Ανάλυση και ερμηνεία των τελικών clusters	49
6.2.3 Επιχειρησιακές προτάσεις ανά cluster καταστημάτων	51
7. Πρόβλεψη Αποχώρησης Καταστημάτων (Shop Churn Prediction)	54
7.1 Πειραματική Διαδικασία	54
7.1.1 Ορισμός της μεταβλητής-στόχου και διαμόρφωση τελικού συνόλου μοντελοποίησης	55
7.1.2 Διαδικασία υλοποίησης στο SAS Model Studio	56
7.2 Αποτελέσματα Πρόβλεψης Αποχώρησης Καταστημάτων	59
7.2.1 Σύγκριση μοντέλων και επιλογή champion model	59
7.2.2 Αξιολόγηση του champion model	61
7.2.3 Σημαντικότερες μεταβλητές πρόβλεψης churn	63
7.2.4 Επιχειρησιακή αξιοποίηση των αποτελεσμάτων	65
8. Συμπεράσματα – Προτάσεις	66
Βιβλιογραφία	69

1. Εισαγωγή

Οι πλατφόρμες online παραγγελίας και διανομής φαγητού αποτελούν πλέον βασικό κανάλι διασύνδεσης των καταναλωτών με τα καταστήματα εστίασης. Η ανάπτυξή τους έχει ενισχύσει σημαντικά την ψηφιοποίηση της αγοράς, καθώς οι χρήστες μπορούν να αναζητούν, να συγκρίνουν και να παραγγέλλουν εύκολα από μεγάλο αριθμό διαθέσιμων επιλογών. Παράλληλα, η λειτουργία αυτών των πλατφορμών παράγει μεγάλο όγκο δεδομένων, τα οποία αποτυπώνουν πραγματικές συναλλαγές και επιχειρησιακές διαδικασίες και μπορούν να αξιοποιηθούν για υποστήριξη αποφάσεων.

Κεντρικό στοιχείο για την επιτυχία μιας πλατφόρμας δεν είναι μόνο η εξυπηρέτηση των καταναλωτών, αλλά και η αποτελεσματική διαχείριση του δικτύου συνεργαζόμενων καταστημάτων. Τα καταστήματα διαφέρουν σημαντικά ως προς το είδος, το μέγεθος, τη δυναμικότητα, την απόδοση και τη σταθερότητα της συνεργασίας τους. Σε ένα περιβάλλον με χιλιάδες συνεργάτες, η χειροκίνητη παρακολούθηση και αξιολόγηση της πορείας κάθε καταστήματος είναι πρακτικά δύσκολη, επομένως καθίσταται απαραίτητη η αξιοποίηση δεδομένων και αναλυτικών μεθόδων για τη συστηματική κατανόηση της συμπεριφοράς και της απόδοσης των συνεργατών.

Ιδιαίτερο ενδιαφέρον παρουσιάζει το φαινόμενο της αποχώρησης καταστημάτων από την πλατφόρμα ή, γενικότερα, της σημαντικής αποδυνάμωσης της δραστηριότητάς τους. Η εξέλιξη αυτή επηρεάζει αρνητικά τόσο το διαθέσιμο εύρος επιλογών για τους καταναλωτές όσο και τα έσοδα και τη συνολική ανταγωνιστικότητα της πλατφόρμας. Για τον λόγο αυτό, η δυνατότητα έγκαιρης αναγνώρισης συνεργατών που εμφανίζουν αυξημένη πιθανότητα αποχώρησης αποκτά ιδιαίτερη επιχειρησιακή αξία, καθώς επιτρέπει τη λήψη προληπτικών ενεργειών διατήρησης. Στο εμπειρικό μέρος της παρούσας εργασίας, το churn δεν παρατηρείται μέσω ρητής ένδειξης τερματισμού συνεργασίας, αλλά προσεγγίζεται λειτουργικά μέσω συμπεριφορικού proxy βασισμένου στην αδράνεια του καταστήματος στην πλατφόρμα.

Παράλληλα, εξίσου σημαντική είναι η κατανόηση της ετερογένειας των καταστημάτων μέσα από ομαδοποίηση, με σκοπό τον εντοπισμό διακριτών προφίλ συνεργατών. Η ομαδοποίηση καταστημάτων μέσω τεχνικών clustering μπορεί να αναδείξει ομάδες με κοινά χαρακτηριστικά ή παρόμοια πρότυπα απόδοσης, δημιουργώντας ένα πρακτικό πλαίσιο τμηματοποίησης που διευκολύνει στοχευμένες εμπορικές και λειτουργικές αποφάσεις. Επιπλέον, η ανάλυση σε επίπεδο ομάδων μπορεί να βοηθήσει στην ερμηνεία του κινδύνου αποχώρησης, καθώς ορισμένα προφίλ ενδέχεται να συνδέονται με αυξημένη πιθανότητα churn.

Στο πλαίσιο αυτό, η παρούσα μεταπτυχιακή διατριβή εξετάζει την εφαρμογή τεχνικών ανάλυσης δεδομένων και μηχανικής μάθησης σε δεδομένα συνεργαζόμενων καταστημάτων από ηλεκτρονική πλατφόρμα διανομής φαγητού. Η εργασία εστιάζει σε περίπου 17.000 καταστήματα. Τα δεδομένα περιλαμβάνουν μεταβλητές που αποτυπώνουν τη δραστηριότητα και την απόδοση των συνεργατών, όπως χαρακτηριστικά παραγγελιών, αξιολογήσεις, χρονικά και επιχειρησιακά μεγέθη, καθώς και παράγωγα χαρακτηριστικά που προκύπτουν από τη διερευνητική ανάλυση.

Οι βασικοί στόχοι της εργασίας είναι δύο. Ο πρώτος στόχος είναι η τμηματοποίηση καταστημάτων (shop clustering) με σκοπό την ανάδειξη διακριτών προφίλ συνεργατών και την επιχειρησιακή ερμηνεία των ομάδων που προκύπτουν. Ο δεύτερος στόχος είναι η πρόβλεψη αποχώρησης καταστημάτων (shop churn prediction), ώστε να κατασκευαστεί μοντέλο που εκτιμά την πιθανότητα αποχώρησης ή ουσιαστικής αποδυνάμωσης της συνεργασίας και να εντοπιστούν παράγοντες που σχετίζονται με τον κίνδυνο churn.

Η συμβολή της εργασίας έγκειται στη διαμόρφωση ενός πρακτικού πλαισίου shop analytics που συνδυάζει τμηματοποίηση συνεργατών και πρόβλεψη αποχώρησης, με στόχο την υποστήριξη αποφάσεων για στοχευμένες ενέργειες διαχείρισης, βελτίωσης απόδοσης και διατήρησης συνεργασιών. Η δομή της διατριβής είναι η εξής: Στο Κεφάλαιο 2 παρουσιάζεται το θεωρητικό

υπόβαθρο και το πλαίσιο εφαρμογής των πλατφορμών online food delivery και των shop analytics. Στο Κεφάλαιο 3 ακολουθεί ανασκόπηση της σχετικής βιβλιογραφίας. Στα επόμενα κεφάλαια περιγράφονται τα δεδομένα, η μεθοδολογία και τα πειράματα για το clustering και την πρόβλεψη churn, και τέλος συνοψίζονται τα συμπεράσματα και προτείνονται κατευθύνσεις για μελλοντική εργασία.

2. Θεωρητικό Υπόβαθρο και Βασικές Έννοιες

2.1 Μεγάλος Όγκος Δεδομένων (Big Data)

Ορισμός και εξέλιξη της έννοιας των Big Data

Στη σύγχρονη ψηφιακή εποχή παράγονται καθημερινά τεράστιοι όγκοι δεδομένων από πλήθος πηγών, όπως συναλλαγές, κοινωνικά δίκτυα, έξυπνες συσκευές και διαδικτυακές υπηρεσίες. Ο όρος Big Data αναφέρεται σε αυτά τα μεγάλα και σύνθετα σύνολα δεδομένων, τα οποία δεν μπορούν να αποθηκευτούν ή να αναλυθούν αποτελεσματικά με τις παραδοσιακές μεθόδους. Η ανάγκη για νέους τρόπους διαχείρισης και αξιοποίησης αυτών των δεδομένων οδήγησε στη δημιουργία ενός ολόκληρου πεδίου γνώσης γύρω από τα Big Data.

Αν και η διαχείριση μεγάλων ποσοτήτων δεδομένων απασχολεί τις επιχειρήσεις εδώ και δεκαετίες, η έννοια των Big Data καθιερώθηκε στις αρχές του 21ου αιώνα. Η εξάπλωση του διαδικτύου, των κινητών συσκευών και των κοινωνικών μέσων πολλαπλασίασε τον όγκο των παραγόμενων πληροφοριών, οδηγώντας σε μια νέα ψηφιακή εποχή. Οι σύγχρονες δραστηριότητες, από τις online συναλλαγές έως τους αισθητήρες του Internet of Things, παράγουν αδιάκοπα δεδομένα που συνθέτουν ένα τεράστιο οικοσύστημα πληροφορίας.

Η σημασία των Big Data δεν περιορίζεται στον όγκο των δεδομένων, αλλά επεκτείνεται στις τεχνολογίες και τις μεθοδολογίες που τα συνοδεύουν. Αρχικά η έμφαση δινόταν στη συλλογή και αποθήκευση πληροφοριών, ενώ σήμερα το ενδιαφέρον στρέφεται στην ανάλυση και αξιοποίησή τους. Οι επιχειρήσεις συνειδητοποίησαν ότι τα δεδομένα έχουν αξία μόνο όταν μετατρέπονται σε χρήσιμη γνώση μέσα από διαδικασίες καθαρισμού, συνδυασμού και ανάλυσης.

Η ραγδαία ανάπτυξη των Big Data οδήγησε στην υιοθέτηση προηγμένων τεχνικών ανάλυσης, όπως η μηχανική μάθηση και η τεχνητή νοημοσύνη, καθώς και νέων υποδομών αποθήκευσης όπως το cloud και οι βάσεις NoSQL. Το πεδίο αυτό επιτρέπει πλέον στις επιχειρήσεις να αντλούν πολύτιμες πληροφορίες από τεράστιους όγκους δεδομένων, να προβλέπουν τάσεις και να λαμβάνουν αποφάσεις βασισμένες σε πραγματικά στοιχεία. Έτσι, τα Big Data δεν αποτελούν απλώς τεχνολογική καινοτομία, αλλά θεμέλιο της ψηφιακής επιχειρηματικής στρατηγικής που επιτρέπει στις οργανώσεις να μετατρέπουν την πληροφορία σε ανταγωνιστικό πλεονέκτημα.

Κύρια χαρακτηριστικά των Big Data (5Vs)

Για να περιγραφούν πληρέστερα οι προκλήσεις και οι δυνατότητες των Big Data, έχει επικρατήσει το μοντέλο των «5V», που προέρχεται από τις λέξεις Volume, Variety, Velocity, Veracity και Value. Κάθε μία από αυτές τις διαστάσεις αντιστοιχεί σε ένα βασικό χαρακτηριστικό της έννοιας και συμβάλλει στην πληρέστερη κατανόηση του φαινομένου.

Το πρώτο χαρακτηριστικό είναι ο Όγκος (Volume), ο οποίος αναφέρεται στην τεράστια ποσότητα δεδομένων που δημιουργούνται καθημερινά. Οι επιχειρήσεις σήμερα συλλέγουν δεδομένα σε κλίμακες terabyte ή και petabyte, από εκατομμύρια συναλλαγές, αισθητήρες και ψηφιακές αλληλεπιδράσεις. Ένας τόσο μεγάλος όγκος πληροφορίας απαιτεί εξειδικευμένες λύσεις αποθήκευσης και διαχείρισης, όπως υποδομές cloud storage και καταμεμημένα συστήματα δεδομένων, που εξασφαλίζουν αποτελεσματική επεξεργασία και ανάλυση.

Η Ποικιλία (Variety) αναφέρεται στους διαφορετικούς τύπους και μορφές δεδομένων που περιλαμβάνονται στα Big Data. Εκτός από τα δομημένα δεδομένα, όπως πίνακες πελατών ή συναλλαγές, σήμερα υπάρχουν και ημιδομημένα ή αδόμητα δεδομένα, όπως κείμενα, εικόνες, βίντεο

και πληροφορίες από τα μέσα κοινωνικής δικτύωσης ή τους αισθητήρες του Internet of Things. Η ποικιλία αυτή αυξάνει την πολυπλοκότητα της επεξεργασίας, απαιτώντας εργαλεία ικανά να συνθέτουν και να αναλύουν δεδομένα διαφορετικής φύσης με τρόπο ενιαίο και αξιόπιστο.

Η Ταχύτητα (Velocity) αφορά τον ρυθμό με τον οποίο δημιουργούνται, μεταδίδονται και αναλύονται τα δεδομένα. Στο σύγχρονο επιχειρησιακό περιβάλλον, η αξία των πληροφοριών μειώνεται αν δεν αξιοποιηθούν άμεσα. Για παράδειγμα, δεδομένα σε πραγματικό χρόνο από συναλλαγές ή από τη συμπεριφορά των χρηστών στο διαδίκτυο πρέπει να αναλυθούν άμεσα για να υποστηρίξουν έγκαιρες αποφάσεις. Οι τεχνολογίες Big Data επιτρέπουν τη ροή και την ανάλυση δεδομένων υψηλής ταχύτητας (streaming data), δίνοντας στις επιχειρήσεις τη δυνατότητα να αντιδρούν σε πραγματικό χρόνο σε αλλαγές ή ευκαιρίες της αγοράς.

Η Αξιοπιστία (Veracity) σχετίζεται με την ποιότητα και την ακρίβεια των δεδομένων. Τα Big Data προέρχονται από πολλές πηγές, όχι πάντα αξιόπιστες, και συχνά περιλαμβάνουν λάθη, ελλείψεις ή προκαταλήψεις. Για να διασφαλιστεί η εγκυρότητα και η συνέπειά τους, απαιτούνται διαδικασίες καθαρισμού και διασταύρωσης δεδομένων. Η αξιοπιστία είναι κρίσιμη προϋπόθεση ώστε τα αποτελέσματα των αναλύσεων να είναι έγκυρα και να οδηγούν σε ορθές αποφάσεις.

Τέλος, η Αξία (Value) αποτελεί το σημαντικότερο χαρακτηριστικό από επιχειρηματική σκοπιά. Τα δεδομένα έχουν πραγματική σημασία μόνο όταν μετατρέπονται σε χρήσιμες πληροφορίες που συμβάλλουν στην επίτευξη των στόχων ενός οργανισμού. Μέσα από τις αναλύσεις Big Data, οι επιχειρήσεις μπορούν να εντοπίζουν νέες ευκαιρίες, να προβλέπουν τάσεις, να βελτιώνουν την εμπειρία των πελατών και να αυξάνουν την αποδοτικότητά τους. Η αξία, λοιπόν, των Big Data δεν έγκειται απλώς στην ποσότητα ή στην τεχνολογία, αλλά στη δυνατότητα μετατροπής της πληροφορίας σε δράση και σε μετρήσιμα αποτελέσματα που ενισχύουν το ανταγωνιστικό πλεονέκτημα.

Παραδείγματα πηγών δεδομένων στις επιχειρήσεις

Οι σύγχρονες επιχειρήσεις αντλούν δεδομένα από ένα ευρύ φάσμα πηγών, τόσο εσωτερικών όσο και εξωτερικών, δημιουργώντας τη βάση των Big Data τους. Οι πηγές αυτές ποικίλλουν ανάλογα με τον κλάδο, τη δραστηριότητα και το επιχειρησιακό μοντέλο κάθε οργανισμού, ωστόσο όλες συμβάλλουν καθοριστικά στη διαμόρφωση μιας πλήρους και πολυδιάστατης εικόνας του επιχειρηματικού περιβάλλοντος.

Μια βασική κατηγορία αφορά τα συναλλακτικά και επιχειρησιακά δεδομένα που παράγονται από τις καθημερινές λειτουργίες μιας εταιρείας. Σε αυτά περιλαμβάνονται οι εγγραφές συναλλαγών, τα στοιχεία παραγωγής και εφοδιαστικής αλυσίδας, τα δεδομένα αποθήκης και αποθεμάτων, καθώς και τα οικονομικά αρχεία. Για παράδειγμα, μια αλυσίδα λιανικής συλλέγει εκατομμύρια καταγραφές πωλήσεων από τα σημεία εξυπηρέτησής της, ενώ μια τράπεζα επεξεργάζεται τεράστιους όγκους δεδομένων από τις ημερήσιες τραπεζικές συναλλαγές των πελατών της.

Εξίσου σημαντικά είναι τα δεδομένα πελατών και marketing, τα οποία τροφοδοτούν τη λήψη αποφάσεων και τις στρατηγικές μάρκετινγκ. Οι επιχειρήσεις αξιοποιούν συστήματα Customer Relationship Management (CRM) για τη συγκέντρωση πληροφοριών, όπως δημογραφικά στοιχεία, ιστορικό αγορών, επίπεδο ικανοποίησης και συχνότητα επικοινωνίας. Αυτά τα δεδομένα προσφέρουν πολύτιμη γνώση σχετικά με τις προτιμήσεις, τις ανάγκες και τη συμπεριφορά των πελατών, επιτρέποντας στις επιχειρήσεις να προσαρμόζουν καλύτερα τα προϊόντα και τις υπηρεσίες τους.

Μία ακόμη πηγή προέρχεται από το διαδίκτυο και τα μέσα κοινωνικής δικτύωσης. Οι επιχειρήσεις συλλέγουν δεδομένα από ιστοσελίδες, πλατφόρμες και εφαρμογές, όπως αρχεία επισκέψεων (web logs), αναζητήσεις, κλικ, καθώς και δεδομένα από τα κοινωνικά δίκτυα (Facebook, Instagram, X, TikTok). Οι αναρτήσεις, τα σχόλια και οι αντιδράσεις των χρηστών αποτελούν πηγή πολύτιμης πληροφόρησης σχετικά με τη φήμη της εταιρείας, τις τάσεις της αγοράς και τα ενδιαφέροντα του κοινού. Επιπλέον, δεδομένα από μηχανές αναζήτησης ή άλλες ψηφιακές πλατφόρμες μπορούν να προσφέρουν εικόνα για τη συμπεριφορά και τις προτιμήσεις των καταναλωτών σε πραγματικό χρόνο.

Τα τελευταία χρόνια, ιδιαίτερη σημασία αποκτούν και τα δεδομένα από αισθητήρες και συσκευές του Internet of Things (IoT). Στις βιομηχανίες, αισθητήρες σε μηχανήματα παραγωγής καταγράφουν θερμοκρασία, απόδοση ή πιθανές δυσλειτουργίες, ενώ στις μεταφορές συσκευές GPS μεταδίδουν δεδομένα θέσης και ταχύτητας. Αντίστοιχα, στον τομέα της υγείας, φορητές συσκευές (wearables) και ιατρικοί αισθητήρες παρακολουθούν ζωτικά δεδομένα ασθενών σε πραγματικό χρόνο. Η ανάλυση τέτοιων δεδομένων βοηθά τις επιχειρήσεις να βελτιστοποιούν τις λειτουργίες τους, να προβλέπουν βλάβες εξοπλισμού και να βελτιώνουν την εμπειρία των χρηστών.

Τέλος, σημαντικό ρόλο παίζουν οι εξωτερικές βάσεις δεδομένων και τα ανοικτά δεδομένα. Πολλοί οργανισμοί αξιοποιούν δημόσια διαθέσιμα δεδομένα (open data) από κυβερνητικούς ή ερευνητικούς φορείς, όπως δημογραφικά στοιχεία, στατιστικά οικονομικά δεδομένα ή πληροφορίες για τον καιρό. Παράλληλα, εμπορικές βάσεις δεδομένων παρέχουν πληροφορίες για τις αγορές, τις συνήθειες καταναλωτών και τον ανταγωνισμό. Συνδυάζοντας εσωτερικές και εξωτερικές πηγές δεδομένων, οι επιχειρήσεις αποκτούν μια πιο ολοκληρωμένη εικόνα του περιβάλλοντός τους και μπορούν να λαμβάνουν τεκμηριωμένες αποφάσεις.

Συνολικά, οι πηγές δεδομένων στις επιχειρήσεις είναι πολυάριθμες και ετερογενείς, καλύπτοντας κάθε πτυχή της επιχειρηματικής δραστηριότητας. Η πρόκληση έγκειται στη σύνθεση αυτών των ποικίλων ροών πληροφορίας και στη μετατροπή τους σε συνεκτική και χρήσιμη γνώση που υποστηρίζει τη στρατηγική λήψη αποφάσεων και ενισχύει το ανταγωνιστικό πλεονέκτημα.

Επιρροή των Big Data στη στρατηγική, τη λήψη αποφάσεων, την καινοτομία και την ανταγωνιστικότητα

Η δυνατότητα αξιοποίησης των Big Data έχει επιφέρει σημαντικές αλλαγές στον τρόπο με τον οποίο οι επιχειρήσεις λαμβάνουν αποφάσεις, καινοτομούν και ανταγωνίζονται στην αγορά. Τα δεδομένα, όταν αναλυθούν σωστά, μετατρέπονται σε στρατηγικό πλεονέκτημα, προσφέροντας διορατικότητα, ταχύτητα και ακρίβεια στη λήψη αποφάσεων.

Πρώτα απ' όλα, τα Big Data έχουν μεταμορφώσει τη διαδικασία λήψης αποφάσεων. Παραδοσιακά, οι διοικητικές αποφάσεις βασίζονταν σε διαίσθηση, εμπειρική γνώση ή περιορισμένα δείγματα πληροφορίας. Σήμερα, όμως, η ανάλυση μεγάλων δεδομένων επιτρέπει στις επιχειρήσεις να λαμβάνουν τεκμηριωμένες, έγκαιρες και πιο αντικειμενικές αποφάσεις, καθιερώνοντας την προσέγγιση του data-driven decision making. Μέσω προηγμένων εργαλείων ανάλυσης, οι οργανισμοί επεξεργάζονται τεράστιους όγκους πληροφοριών και αποκτούν πλήρη εικόνα των λειτουργιών τους και των μεταβολών της αγοράς σε πραγματικό χρόνο. Ενδεικτικά, εταιρείες λιανικής παρακολουθούν συνεχώς τις πωλήσεις και τα αποθέματά τους για να προσαρμόζουν δυναμικά την εφοδιαστική αλυσίδα, ενώ τα τμήματα marketing μπορούν να αξιολογούν άμεσα την απόδοση μιας καμπάνιας και να ανακατανέμουν πόρους για μέγιστη αποτελεσματικότητα. Επιπλέον, οι τεχνικές προγνωστικής ανάλυσης (predictive analytics) δίνουν τη δυνατότητα πρόβλεψης μελλοντικών τάσεων και κινήσεων της αγοράς, ενισχύοντας την ικανότητα των επιχειρήσεων να προετοιμάζονται στρατηγικά. Έτσι, οι αποφάσεις γίνονται όχι μόνο πιο ακριβείς, αλλά και ταχύτερες, προσδίδοντας στις εταιρείες ευελιξία και προδραστικότητα.

Παράλληλα, τα Big Data λειτουργούν ως ισχυρός καταλύτης καινοτομίας. Αναλύοντας μεγάλα και ετερογενή σύνολα δεδομένων, οι επιχειρήσεις εντοπίζουν πρότυπα, τάσεις και συσχετίσεις που δεν θα μπορούσαν να ανακαλύψουν διαφορετικά. Αυτή η γνώση οδηγεί σε νέα ή βελτιωμένα προϊόντα, υπηρεσίες και επιχειρηματικά μοντέλα. Εταιρείες ψυχαγωγίας, για παράδειγμα, αξιοποιούν δεδομένα θέασης εκατομμυρίων χρηστών για να παράγουν εξατομικευμένο περιεχόμενο που ανταποκρίνεται στις προτιμήσεις του κοινού. Αντίστοιχα, στη βιομηχανία, η ανάλυση δεδομένων από αισθητήρες εξοπλισμού επιτρέπει την ανάπτυξη «έξυπνων» γραμμών παραγωγής που αυτορρυθμίζονται για μέγιστη αποδοτικότητα. Ακόμη και σε παραδοσιακούς τομείς, όπως η γεωργία, τα Big Data έχουν φέρει επανάσταση μέσω της «έξυπνης καλλιέργειας», όπου δεδομένα για το έδαφος, τον καιρό και τη χρήση πόρων οδηγούν σε στοχευμένες παρεμβάσεις και αυξημένη παραγωγικότητα. Η σε βάθος κατανόηση των πελατών και των αναγκών τους εμπνέει επίσης νέες ιδέες και προϊόντα, καθιστώντας την καινοτομία διαρκή και μετρήσιμη διαδικασία.

Η αξιοποίηση των Big Data ενισχύει επίσης τη λειτουργική αποδοτικότητα και βελτιώνει τις επιχειρησιακές διαδικασίες. Μέσα από την ανάλυση δεδομένων λειτουργίας, οι εταιρείες εντοπίζουν σημεία σπατάλης και βελτίωσης, προχωρούν σε προληπτική συντήρηση εξοπλισμού και βελτιώνουν τη διαχείριση πόρων. Στη βιομηχανία, για παράδειγμα, η παρακολούθηση δεδομένων μηχανών σε πραγματικό χρόνο επιτρέπει την έγκαιρη ανίχνευση φθοράς και τη μείωση του χρόνου διακοπής λειτουργίας. Στην εφοδιαστική αλυσίδα, η ανάλυση Big Data επιτρέπει τη βελτιστοποίηση προμηθειών και αποθεμάτων, αποτρέποντας καθυστερήσεις και περιττά κόστη. Η χρήση δεδομένων σε επίπεδο ανθρώπινου δυναμικού (HR analytics) συμβάλλει επίσης στην αποδοτικότερη αξιοποίηση του προσωπικού και στην αύξηση της παραγωγικότητας.

Τέλος, ο σημαντικότερος ίσως τομέας επιρροής των Big Data είναι η ενίσχυση της ανταγωνιστικότητας. Σε αγορές όπου οι τεχνολογικές δυνατότητες είναι ευρέως διαθέσιμες, η ικανότητα μιας επιχείρησης να αξιοποιεί τα δεδομένα της αποτελεί κρίσιμο παράγοντα διαφοροποίησης. Οι επιχειρήσεις που υιοθετούν κουλτούρα ανάλυσης δεδομένων αποκτούν βαθύτερη κατανόηση του πελατειακού τους κοινού και των τάσεων της αγοράς, με αποτέλεσμα να προσφέρουν προϊόντα και υπηρεσίες που ανταποκρίνονται καλύτερα στις ανάγκες των πελατών. Παράλληλα, η συνεχής μέτρηση και αξιολόγηση της απόδοσης οδηγεί σε ταχύτερη προσαρμογή και συνεχή βελτίωση. Πολλές μελέτες δείχνουν ότι οι εταιρείες που αξιοποιούν αποτελεσματικά τα Big Data Analytics καταγράφουν αυξημένα έσοδα, μερίδιο αγοράς και ικανοποίηση πελατών σε σχέση με τους ανταγωνιστές τους. Η ταχύτητα και η προγνωστική δύναμη που προσφέρουν τα δεδομένα μεταφράζονται σε στρατηγικό πλεονέκτημα που δύσκολα αντιγράφεται.

Συνολικά, τα Big Data έχουν περάσει από το στάδιο της τεχνολογικής καινοτομίας στο στάδιο του απαραίτητου στρατηγικού εργαλείου. Αποτελούν τη βάση της επιχειρηματικής γνώσης του 21ου αιώνα, προσφέροντας στις επιχειρήσεις τη δυνατότητα να κατανοούν το περιβάλλον τους, να καινοτομούν, να βελτιώνουν τις λειτουργίες τους και να λαμβάνουν αποφάσεις με αυτοπεποίθηση, στηριζόμενες σε στοιχεία. Σε μια εποχή όπου η πληροφορία αποτελεί το πολυτιμότερο κεφάλαιο, η ικανότητα φιλτραρίσματος και αξιοποίησης των δεδομένων μεταφράζεται σε επιχειρηματική επιτυχία και βιώσιμη ανάπτυξη.

2.2 Επιχειρησιακή Ευφυΐα (Business Intelligence) και Επιχειρησιακή Αναλυτική (Business Analytics)

Στη σύγχρονη επιχειρηματική σκηνή, η αξιοποίηση των δεδομένων είναι ζωτικής σημασίας για τη διατήρηση ανταγωνιστικού πλεονεκτήματος. Καθημερινά, οι οργανισμοί συλλέγουν τεράστιους όγκους πληροφοριών από τις λειτουργίες τους, τις αγορές και τους πελάτες. Οι έννοιες της Επιχειρησιακής Ευφυΐας (Business Intelligence - BI) και των Επιχειρησιακών Αναλύσεων (Business Analytics - BA) έχουν αναδειχθεί ως κεντρικά εργαλεία που βοηθούν τις εταιρείες να μετατρέπουν αυτά τα ακατέργαστα δεδομένα σε χρήσιμες γνώσεις. Παρότι οι δύο όροι συχνά χρησιμοποιούνται εναλλακτικά, αναφέρονται σε διακριτές – αλλά αλληλένδετες – προσεγγίσεις στην ανάλυση των επιχειρηματικών δεδομένων.

Ορισμός της Business Intelligence (BI)

Η Business Intelligence (Επιχειρησιακή Ευφυΐα) ορίζεται ως το σύνολο των μεθοδολογιών, διαδικασιών και τεχνολογικών εργαλείων που χρησιμοποιεί ένας οργανισμός για τη συλλογή, ενοποίηση, ανάλυση και παρουσίαση επιχειρησιακών δεδομένων. Μέσω της BI, τα ακατέργαστα δεδομένα μετατρέπονται σε ουσιαστικές πληροφορίες που μπορούν να υποστηρίξουν τη λήψη τεκμηριωμένων αποφάσεων. Στην πράξη, η Business Intelligence περιλαμβάνει δραστηριότητες όπως τη δημιουργία αποθηκών δεδομένων (data warehouses), την εξόρυξη δεδομένων (data mining), τη δημιουργία αναφορών και τη χρήση δυναμικών πινάκων ελέγχου (dashboards). Το αποτέλεσμα αυτών των διαδικασιών είναι η παραγωγή δεικτών απόδοσης και γραφικών απεικονίσεων που βοηθούν στη σαφή κατανόηση της παρούσας και παρελθούσας επιχειρησιακής κατάστασης. Με απλά λόγια, η BI εστιάζει στην περιγραφική ανάλυση, βοηθώντας να απαντηθούν ερωτήματα όπως «Τι συνέβη;» και «Πού βρισκόμαστε τώρα;».

Ορισμός των Business Analytics (BA)

Τα Business Analytics (Επιχειρησιακές Αναλύσεις) αφορούν την εκτενέστερη χρήση δεδομένων, στατιστικών μεθόδων και μαθηματικών μοντέλων για την εξαγωγή βαθύτερων γνώσεων, την πρόβλεψη τάσεων και τη λήψη πιο αποτελεσματικών αποφάσεων για το μέλλον. Τα BA περιλαμβάνουν προχωρημένες τεχνικές ανάλυσης όπως η διαγνωστική ανάλυση (διερεύνηση του γιατί συνέβη κάτι), η προγνωστική ανάλυση (πρόβλεψη του τι πιθανόν θα συμβεί) και η κανονιστική ανάλυση (προτάσεις ενεργειών για τη βελτιστοποίηση των αποτελεσμάτων). Ενώ η BI επικεντρώνεται κυρίως στην παρουσίαση των δεδομένων, τα BA εστιάζουν στην ερμηνεία και διερεύνηση των αιτιών πίσω από τα ευρήματα, αξιοποιώντας εργαλεία μηχανικής μάθησης και αλγοριθμικά μοντέλα. Με αυτόν τον τρόπο, τα Business Analytics δίνουν τη δυνατότητα στις επιχειρήσεις να προβλέπουν μελλοντικές εξελίξεις και να λαμβάνουν τεκμηριωμένες αποφάσεις, απαντώντας σε ερωτήματα όπως «Γιατί έγινε αυτό;» και «Τι πρέπει να κάνουμε στη συνέχεια;».

Διαφορές και σύνδεση μεταξύ Business Intelligence και Business Analytics

Παρόλο που η Business Intelligence και τα Business Analytics μοιράζονται τον κοινό στόχο της αξιοποίησης δεδομένων προς όφελος της επιχείρησης, διαφέρουν ως προς την εστίαση, τις τεχνικές και τον σκοπό τους. Οι δύο έννοιες λειτουργούν συμπληρωματικά, καλύπτοντας διαφορετικά στάδια στη διαδικασία ανάλυσης και λήψης αποφάσεων.

Η Business Intelligence εστιάζει κυρίως στο παρελθόν και στο παρόν, παρέχοντας μια περιγραφική αποτύπωση της επιχειρησιακής πραγματικότητας. Μέσω αναφορών, δεικτών και οπτικοποιήσεων, βοηθά τα στελέχη να κατανοήσουν τι έχει συμβεί και ποια είναι η τρέχουσα κατάσταση της εταιρείας. Αντίθετα, τα Business Analytics έχουν προοπτική προς το μέλλον. Χρησιμοποιούν τα υπάρχοντα δεδομένα για να εντοπίσουν συσχετίσεις, να προβλέψουν μελλοντικές τάσεις και να υποδείξουν ενέργειες που θα οδηγήσουν σε βελτίωση της απόδοσης.

Αναφορικά με τον τύπο της ανάλυσης, η BI περιορίζεται κυρίως σε περιγραφικά εργαλεία που απεικονίζουν την υπάρχουσα κατάσταση, χωρίς απαραίτητα να εμβαθύνουν στα αίτια. Από την άλλη, τα BA εφαρμόζουν στατιστικές αναλύσεις, μοντέλα πρόβλεψης και τεχνικές μηχανικής μάθησης για να αποκαλύψουν το «γιατί» πίσω από τα δεδομένα και να προτείνουν συγκεκριμένες δράσεις.

Όσον αφορά τον σκοπό και τη χρήση, η Business Intelligence υποστηρίζει την καθημερινή λειτουργία των οργανισμών, παρέχοντας έγκαιρη πληροφόρηση και διευκολύνοντας την παρακολούθηση της απόδοσης. Τα Business Analytics επικεντρώνονται περισσότερο στη βελτιστοποίηση και την καινοτομία, χρησιμοποιώντας τα δεδομένα για να καθοδηγήσουν στρατηγικές αποφάσεις και να εντοπίσουν ευκαιρίες βελτίωσης. Έτσι, ενώ η BI απαντά στο ερώτημα «πού βρισκόμαστε τώρα;», τα BA συμβάλλουν στην απάντηση του ερωτήματος «πού πρέπει να πάμε;».

Παρά τις διαφορές τους, οι δύο προσεγγίσεις είναι στενά συνδεδεμένες και αλληλοσυμπληρούμενες. Η Business Intelligence προσφέρει τη βάση αξιόπιστων δεδομένων και αναφορών, πάνω στην οποία οικοδομείται η Business Analytics. Σε πολλά επιχειρησιακά περιβάλλοντα, οι δύο τεχνολογίες συνυπάρχουν σε ενιαίες πλατφόρμες ανάλυσης, παρέχοντας μια ολοκληρωμένη εικόνα της επιχείρησης. Ο συνδυασμός τους επιτρέπει στις οργανώσεις να περιγράψουν με ακρίβεια την παρούσα κατάσταση, να προβλέπουν τις μελλοντικές εξελίξεις και να λαμβάνουν αποφάσεις βασισμένες σε πραγματικά στοιχεία.

Συνεισφορά στη λήψη στρατηγικών και λειτουργικών αποφάσεων

Η λήψη αποφάσεων σε μια επιχείρηση διακρίνεται γενικά σε στρατηγική, η οποία αφορά μακροπρόθεσμες επιλογές σε επίπεδο διοίκησης, και σε λειτουργική ή επιχειρησιακή, που σχετίζεται με τις βραχυπρόθεσμες, καθημερινές λειτουργίες. Τόσο η Business Intelligence όσο και τα Business Analytics διαδραματίζουν κρίσιμο ρόλο και στα δύο αυτά επίπεδα, επιτρέποντας στις επιχειρήσεις να στηρίζουν τις αποφάσεις τους σε δεδομένα αντί σε υποθέσεις.

Σε στρατηγικό επίπεδο, τα δεδομένα που παρέχει η Business Intelligence συμβάλλουν στη διαμόρφωση της κατεύθυνσης και του οράματος του οργανισμού. Μέσω της ανάλυσης ιστορικών τάσεων πωλήσεων, απόδοσης προϊόντων ή γεωγραφικών αγορών, καθώς και βασικών οικονομικών δεικτών, τα εργαλεία BI προσφέρουν πολύτιμη πληροφόρηση στη διοίκηση για τον καθορισμό στόχων και τη χάραξη μακροπρόθεσμης στρατηγικής. Τα Business Analytics ενισχύουν αυτή τη διαδικασία προσθέτοντας την προγνωστική διάσταση: μέσω μοντέλων προσομοίωσης και στατιστικών προβλέψεων, μπορούν να υποδείξουν πώς ενδέχεται να επηρεαστούν τα έσοδα ή η ζήτηση από αλλαγές στην πολιτική τιμολόγησης, σε νέα προϊόντα ή σε επενδυτικές αποφάσεις. Με αυτόν τον τρόπο, οι ηγέτες λαμβάνουν τεκμηριωμένες αποφάσεις, μειώνουν την αβεβαιότητα και εναρμονίζουν τις ενέργειες του οργανισμού με δεδομένα πραγματικής αξίας. Εντέλει, η Business Intelligence και τα Business Analytics λειτουργούν συνδυαστικά, παρέχοντας μια πλήρη εικόνα για την αγορά και καθιστώντας τη διαδικασία χάραξης στρατηγικής πιο αντικειμενική και στοχευμένη.

Σε λειτουργικό επίπεδο, η Business Intelligence παρέχει συνεχή ορατότητα της επιχειρησιακής δραστηριότητας, προσφέροντας στους διαχειριστές ενημερωμένες πληροφορίες σε πραγματικό χρόνο. Μέσω πινάκων ελέγχου και δυναμικών αναφορών, οι υπεύθυνοι μπορούν να παρακολουθούν πωλήσεις, αποθέματα ή δείκτες απόδοσης και να λαμβάνουν άμεσες αποφάσεις, όπως την αναπλήρωση προϊόντων που παρουσιάζουν αυξημένη ζήτηση. Παράλληλα, τα Business Analytics ενισχύουν τη λειτουργική αποτελεσματικότητα μέσω προβλεπτικών και βελτιστοποιητικών τεχνικών. Για παράδειγμα, προγνωστικά μοντέλα μπορούν να προβλέψουν πότε ένας εξοπλισμός χρειάζεται συντήρηση, αποτρέποντας βλάβες και καθυστερήσεις, ή να εντοπίσουν έγκαιρα μη φυσιολογικά μοτίβα συναλλαγών που ενδέχεται να υποδηλώνουν απάτη. Έτσι, οι καθημερινές αποφάσεις λαμβάνονται ταχύτερα και με μεγαλύτερη ακρίβεια, βασισμένες όχι μόνο στην εμπειρία των στελεχών, αλλά και σε τεκμηριωμένα δεδομένα.

Συνολικά, η συνεισφορά της Business Intelligence και των Business Analytics στη διαδικασία λήψης αποφάσεων είναι καθοριστική. Μετατρέπουν την πληθώρα των δεδομένων σε γνώση και επιτρέπουν στις επιχειρήσεις να αντιδρούν άμεσα στις μεταβαλλόμενες συνθήκες, να σχεδιάζουν πιο αποτελεσματικές στρατηγικές και να ενισχύουν την ανταγωνιστικότητά τους τόσο σε επιχειρησιακό όσο και σε στρατηγικό επίπεδο.

2.3 Τεχνητή Νοημοσύνη (Artificial Intelligence)

Η Τεχνητή Νοημοσύνη (Artificial Intelligence – AI) αποτελεί κλάδο της επιστήμης των υπολογιστών που ασχολείται με την ανάπτυξη συστημάτων ικανών να εκτελούν πολύπλοκες εργασίες προσομοιώνοντας ανθρώπινες γνωστικές λειτουργίες. Με τη χρήση μεγάλων όγκων δεδομένων και αλγορίθμων μηχανικής μάθησης, τα συστήματα AI μαθαίνουν να αναγνωρίζουν πρότυπα, να συναγάγουν συμπεράσματα και να λαμβάνουν αποφάσεις χωρίς άμεση ανθρώπινη παρέμβαση. Σε πρακτικό επίπεδο, η τεχνητή νοημοσύνη επιτρέπει σε υπολογιστικά συστήματα να αναγνωρίζουν ομιλία, να εντοπίζουν μοτίβα, να επιλύουν προβλήματα και να προβλέπουν μελλοντικές τάσεις.

Στον επιχειρηματικό τομέα, η AI έχει αναδειχθεί σε καθοριστικό παράγοντα βελτίωσης της αποδοτικότητας και της εμπειρίας πελατών. Πολλές εταιρείες την ενσωματώνουν στις λειτουργίες τους, αξιοποιώντας τη για αυτοματοποίηση, ανάλυση δεδομένων και λήψη αποφάσεων. Ενδεικτικά, ορισμένες από τις πιο χαρακτηριστικές εφαρμογές της είναι οι εξής:

- Εξυπηρέτηση πελατών: Τα chatbots και οι ψηφιακοί βοηθοί χρησιμοποιούν επεξεργασία φυσικής γλώσσας (Natural Language Processing – NLP) για να επικοινωνούν με τους πελάτες, παρέχοντας άμεσες και προσωποποιημένες απαντήσεις σε ερωτήματα ή αιτήματα. Με αυτόν τον τρόπο, οι επιχειρήσεις προσφέρουν 24ωρη εξυπηρέτηση, μειώνοντας παράλληλα τον φόρτο εργασίας του ανθρώπινου προσωπικού.
- Πρόβλεψη ζήτησης: Μέσω αλγορίθμων μηχανικής μάθησης (Machine Learning), η AI αναλύει ιστορικά δεδομένα πωλήσεων και παράγοντες της αγοράς, όπως εποχικότητα και καταναλωτικές τάσεις, για να προβλέψει μελλοντική ζήτηση. Αυτό βοηθά τις επιχειρήσεις να προγραμματίζουν αποτελεσματικότερα την παραγωγή και τη διαχείριση αποθεμάτων, αποφεύγοντας ελλείψεις ή υπερβολές.

- Αυτοματοποίηση διεργασιών: Η Robotic Process Automation (RPA) αξιοποιεί τεχνητή νοημοσύνη για την εκτέλεση επαναλαμβανόμενων εργασιών, όπως καταχώριση δεδομένων ή επεξεργασία τιμολογίων. Η εφαρμογή της μειώνει τα σφάλματα, εξοικονομεί χρόνο και επιτρέπει στους εργαζόμενους να επικεντρώνονται σε δραστηριότητες μεγαλύτερης προστιθέμενης αξίας.
- Ανάλυση δεδομένων και επιχειρηματική ευφυΐα: Τα συστήματα AI επεξεργάζονται τεράστιους όγκους δεδομένων για να εντοπίζουν τάσεις και σχέσεις που δύσκολα θα μπορούσε να αναγνωρίσει ο άνθρωπος. Συνδυάζοντας την τεχνητή νοημοσύνη με εργαλεία Business Intelligence, οι επιχειρήσεις αποκτούν βαθύτερη κατανόηση των λειτουργιών τους και λαμβάνουν καλύτερα τεκμηριωμένες αποφάσεις σχετικά με την αγορά, τους πελάτες και τις στρατηγικές τους κινήσεις.

Η σύνδεση της Artificial Intelligence με τα Business Analytics είναι άμεση και αμφίδρομη. Τα Business Analytics αξιοποιούν στατιστικές μεθόδους και ψηφιακά εργαλεία για την ανάλυση επιχειρηματικών δεδομένων με στόχο τη λήψη τεκμηριωμένων αποφάσεων. Η AI ενισχύει αυτή τη διαδικασία, εισάγοντας προηγμένες δυνατότητες όπως προβλεπτική ανάλυση (predictive analytics) και αυτοματοποίηση. Μέσω αλγορίθμων μηχανικής μάθησης δημιουργούνται προγνωστικά μοντέλα που εντοπίζουν μοτίβα στα δεδομένα και επιτρέπουν την έγκαιρη ανίχνευση τάσεων ή κινδύνων. Έτσι, οι επιχειρήσεις μπορούν να ανταποκρίνονται γρήγορα στις αλλαγές της αγοράς και να σχεδιάζουν με στρατηγικό τρόπο τα επόμενα βήματά τους.

2.4 Πλαίσιο Εφαρμογής: Online Food Delivery και Shop Analytics

Οι πλατφόρμες online food delivery λειτουργούν ως ψηφιακοί διαμεσολαβητές που συνδέουν καταναλωτές και επιχειρήσεις εστίασης, ενώ σε πολλές περιπτώσεις υποστηρίζουν και τη διαδικασία διανομής. Στην Ελλάδα δραστηριοποιούνται πλατφόρμες όπως η Efood, η Wolt και η BOX, ενώ διεθνώς αντίστοιχα παραδείγματα αποτελούν οι Uber Eats, DoorDash, Deliveroo και Grubhub. Στο πλαίσιο της παρούσας διατριβής, η εμπειρική ανάλυση βασίζεται αποκλειστικά σε δεδομένα ηλεκτρονικής πλατφόρμας διανομής φαγητού. Η μονάδα ανάλυσης της εργασίας είναι το συνεργαζόμενο κατάστημα. Η πλατφόρμα διατηρεί πληροφορία που αποτυπώνει τη δραστηριότητα και την απόδοση του κάθε καταστήματος στην πάροδο του χρόνου, επιτρέποντας την αξιολόγηση της συμπεριφοράς, την τμηματοποίηση συνεργατών και την ανάπτυξη προγνωστικών μοντέλων.

Shop Analytics: Κατηγορίες δεδομένων και επιχειρησιακοί δείκτες

Με τον όρο Shop Analytics νοείται η ανάλυση δεδομένων που σχετίζονται με τη λειτουργία και την απόδοση των συνεργαζόμενων καταστημάτων στην πλατφόρμα, με στόχο την υποστήριξη της λήψης αποφάσεων. Τα διαθέσιμα δεδομένα περιλαμβάνουν μετρικές ζήτησης και δραστηριότητας (π.χ. πλήθος και συχνότητα παραγγελιών), οικονομικές μετρικές (π.χ. αξία παραγγελιών), δείκτες ποιότητας (π.χ. αξιολογήσεις), λειτουργικές μεταβλητές (π.χ. χρόνοι προετοιμασίας ή παράδοσης, ακυρώσεις), καθώς και χρονικά χαρακτηριστικά που περιγράφουν την εξέλιξη ενός καταστήματος (π.χ. διάρκεια συνεργασίας, μεταβολές στη δραστηριότητα, εποχικότητα). Η αξιοποίηση των παραπάνω μετρικών επιτρέπει την παρακολούθηση της απόδοσης, τη σύγκριση καταστημάτων με βάση κοινά κριτήρια και τον εντοπισμό μοτίβων ή αποκλίσεων που σχετίζονται με την ποιότητα υπηρεσίας, την εμπορική δυναμική και τη σταθερότητα της συνεργασίας.

Αναλυτικοί στόχοι: Segmentation και Churn Prediction

Με δεδομένη την ετερογένεια των συνεργαζόμενων καταστημάτων, η ανάλυση αποκλειστικά σε επίπεδο μεμονωμένων περιπτώσεων δεν είναι πάντα αποδοτική σε κλίμακα. Για τον λόγο αυτό, η παρούσα εργασία εστιάζει σε δύο συμπληρωματικούς αναλυτικούς στόχους.

Ο πρώτος στόχος είναι η τμηματοποίηση καταστημάτων μέσω τεχνικών μη επιβλεπόμενης μάθησης, ώστε να προκύψουν ομάδες με παρόμοια χαρακτηριστικά ή πρότυπα απόδοσης. Οι ομάδες

αυτές επιδιώκεται να είναι επιχειρησιακά ερμηνεύσιμες, ώστε να μπορούν να αξιοποιηθούν ως προφίλ συνεργατών για πιο στοχευμένη διαχείριση.

Ο δεύτερος στόχος είναι η πρόβλεψη αποχώρησης καταστημάτων. Στην εργασία, η αποχώρηση προσεγγίζεται ως τερματισμός ή ουσιαστική αδράνεια της συνεργασίας με την πλατφόρμα. Η πρόβλεψη αντιμετωπίζεται ως πρόβλημα επιβλεπόμενης μάθησης, με στόχο την εκτίμηση της πιθανότητας αποχώρησης και τον εντοπισμό μεταβλητών που σχετίζονται με αυξημένο κίνδυνο.

Συνοψίζοντας, η παρούσα ενότητα οριοθετεί το περιβάλλον εφαρμογής, τη μονάδα ανάλυσης και τους βασικούς αναλυτικούς στόχους της διατριβής. Στα επόμενα κεφάλαια ακολουθεί η ανασκόπηση της σχετικής βιβλιογραφίας και στη συνέχεια παρουσιάζονται η μεθοδολογία και τα πειράματα που υλοποιούνται στα διαθέσιμα δεδομένα της μελέτης.

3. Ανασκόπηση Βιβλιογραφίας (Literature Review)

3.1 Τμηματοποίηση Καταστημάτων (Shop Clustering)

Η τμηματοποίηση καταστημάτων (shop clustering) αποτελεί μία από τις πιο κρίσιμες τεχνικές ανάλυσης δεδομένων στο πλαίσιο των πλατφορμών ηλεκτρονικής παραγγελίας φαγητού. Μέσω αυτής, επιδιώκεται η ομαδοποίηση καταστημάτων με παρόμοια χαρακτηριστικά, όπως όγκοι πωλήσεων, ρυθμοί ανάπτυξης, επίπεδα εξυπηρέτησης και γεωγραφικές συνθήκες λειτουργίας. Η προσέγγιση αυτή συμβάλλει ουσιαστικά στη βελτίωση των επιχειρησιακών αποφάσεων, καθώς επιτρέπει την προσαρμογή στρατηγικών μάρκετινγκ, προωθητικών ενεργειών και πολιτικών συνεργασίας ανά κατηγορία καταστημάτων. Στο πλαίσιο της παρούσας εργασίας, η κατανόηση των μεθόδων και εφαρμογών τμηματοποίησης είναι καθοριστικής σημασίας, καθώς καθοδηγεί την ανάλυση των δεδομένων της πλατφόρμας και τη δημιουργία στοχευμένων επιχειρησιακών μοντέλων.

Οι Agarwal et al. (2019) εξετάζουν πώς μπορούν να τμηματοποιηθούν πάνω από 5.000 φυσικά καταστήματα ενός μεγάλου λιανεμπορικού ομίλου, με στόχο τη βελτίωση στοχευμένων στρατηγικών μάρκετινγκ. Ως δεδομένα χρησιμοποιούνται χαρακτηριστικά όπως αριθμός συναλλαγών, μέσο ποσό συναλλαγής, πλήθος επισκεπτών, κυρίαρχες κατηγορίες προϊόντων, μέγεθος και διάταξη καταστήματος, ανταπόκριση σε προσφορές και γεωγραφικά δεδομένα. Μεθοδολογικά, συγκρίνουν τέσσερις τεχνικές clustering (ιεραρχική ομαδοποίηση, Self-Organizing Maps – SOM, Gaussian Mixture Models – GMM και Fuzzy C-Means), επιλέγοντας το πλήθος των clusters βάσει δεικτών όπως ο συντελεστής Silhouette ή τα κριτήρια AIC/BIC. Τα ευρήματα δείχνουν ότι οι προχωρημένες μη-παραμετρικές μέθοδοι (SOM, GMM, FCM) παράγουν πιο διακριτά προφίλ ομάδων από τις κλασικές μεθόδους, επιτρέποντας την οριοθέτηση καταστημάτων με ειδικά προφίλ πωλήσεων (π.χ. υψηλές πωλήσεις αλλά λίγα καταστήματα έναντι πολλών καταστημάτων με χαμηλότερες πωλήσεις). Η ανάλυση υποδεικνύει ότι κάθε cluster απαιτεί ξεχωριστή επιχειρηματική στρατηγική (π.χ. διαφορετικές προσφορές ή διαρρύθμιση).

Η μελέτη των Prajapati et al. (2021) προτείνει μια καινοτόμο προσέγγιση τμηματοποίησης για έναν πολυεθνικό όμιλο λιανικής (Ινδία) με περίπου 2.500 καταστήματα που έχουν τμήματα Ανδρικών, Γυναικείων, Παιδικών και Αθλητικών ειδών. Σκοπός τους είναι να επιτρέψουν σε ένα κατάστημα να ανήκει σε πολλαπλά clusters ανά τμήμα (πολλαπλή συμμετοχή). Στα δεδομένα τους περιλαμβάνονται γενικές μεταβλητές (π.χ. επιφάνεια πωλήσεων) και ειδικές ανά τμήμα (π.χ. «Men Wall Unit»), καθώς και στοιχεία πωλήσεων, ανταγωνισμού και χαρακτηριστικών καταστημάτων. Ως μεθοδολογία εφαρμόζουν τον αλγόριθμο K-prototype (συνδυασμός K-means για αριθμητικά και K-modes για κατηγορικά δεδομένα) ώστε να διαχειριστούν μίγμα μεταβλητών. Οι συγγραφείς υιοθετούν τετραφασική διαδικασία (ενσωμάτωση δεδομένων και EDA, επιλογή μεταβλητών, clustering και καταγραφή προφίλ), με στόχο τη δημιουργία επαναχρησιμοποιήσιμων προφίλ καταστημάτων ανά τμήμα. Τα αποτελέσματα δείχνουν ότι ο νέος αλγόριθμος εντοπίζει «overlapping» συμμετοχές: για παράδειγμα ένα κατάστημα μπορεί να έχει υψηλή ένταση συμμετοχής σε ανδρικά είδη και μέτρια σε παιδικά, προσφέροντας καλύτερες δυνατότητες σχεδιασμού προσφορών. Η ανάλυση υπογραμμίζει

ότι η «one-size-fits-all» προσέγγιση απέτυχε να εντοπίσει οδηγούς ανά τμήμα, ενώ το προτεινόμενο μοντέλο δημιουργεί διακριτά προφίλ που οδηγούν σε ουσιαστικές επιχειρησιακές αποφάσεις.

Οι Ünal et al. (2022) εφαρμόζουν κλασικές τεχνικές clustering (K-means και Fuzzy C-means) σε δίκτυο 101 καταστημάτων μιας λιανεμπορικής εταιρείας, χρησιμοποιώντας 9 μεταβλητές όπως τοποθεσία καταστήματος, επίπεδο εισοδημάτων περιοχής, αριθμός τιμολογίων και κυκλοφοριακή ταχύτητα αποθέματος. Η μελέτη βρήκε ως βέλτιστο πλήθος clusters 8 για το Fuzzy C-means (με δείκτη Dunn) και 7 για το K-means (με δείκτη Silhouette). Τα clusters εμφανίζουν διαφοροποιημένα χαρακτηριστικά (π.χ. ανάλογα με δημογραφικά της περιοχής ή επίδοση πωλήσεων), υποδεικνύοντας ότι ο συνδυασμός αλγορίθμων μπορεί να βοηθήσει στην εξαγωγή στοχευμένων στρατηγικών marketing ανά ομάδα καταστημάτων.

Τέλος, οι Bilgiç et al. (2015) μελέτησαν τμηματοποίηση 73 καταστημάτων σούπερ μάρκετ στην Κωνσταντινούπολη, με κύριο στόχο τον σχεδιασμό στοχευμένων στρατηγικών μάρκετινγκ. Συγκέντρωσαν χαρακτηριστικά όπως μέγεθος καταστήματος, ύπαρξη χώρου στάθμευσης, αριθμός ανταγωνιστών, γεωγραφικά χαρακτηριστικά (π.χ. κοντά σε πανεπιστήμιο, εργοστάσιο, τουριστικό κέντρο) και δημογραφικά στοιχεία της περιοχής. Μεθοδολογικά, χρησιμοποίησαν ιεραρχικό clustering για τον διαχωρισμό των καταστημάτων σε 5 clusters και στη συνέχεια εφάρμοσαν κανόνες συσχέτισης (Association Rules) σε κάθε cluster. Η ανάλυση κατέληξε σε πέντε ομάδες με διακριτά «καλάθια προϊόντων»: π.χ. ένα cluster περιέχει καταστήματα με υψηλό πλήθος ανταγωνιστών κοντά σε πανεπιστήμιο, ενώ άλλο έχει μεγάλη επιφάνεια πωλήσεων και μέσο πλήθος ανταγωνιστών. Οι κανόνες συσχέτισης αποκάλυψαν διαφορετικά συχνά καλάθια αγορών ανά cluster, δίνοντας πολύτιμες πληροφορίες για στοχευμένες προωθητικές ενέργειες σε κάθε ομάδα καταστημάτων.

Συνολικά, η βιβλιογραφία για την τμηματοποίηση καταστημάτων αναδεικνύει τη σημασία της επιλογής κατάλληλων αλγορίθμων και μεταβλητών ανάλογα με τα δεδομένα και τους επιχειρησιακούς στόχους. Οι πιο σύγχρονες προσεγγίσεις (όπως τα μη-παραμετρικά ή υβριδικά μοντέλα) επιτυγχάνουν πιο σαφείς και ερμηνεύσιμες ομάδες, παρέχοντας στις επιχειρήσεις τη δυνατότητα ανάπτυξης προσαρμοσμένων στρατηγικών. Στην περίπτωση των πλατφορμών διανομής φαγητού, τέτοιες μέθοδοι μπορούν να χρησιμοποιηθούν για την αναγνώριση τύπων καταστημάτων με κοινές ανάγκες και επιδόσεις, υποστηρίζοντας τη στοχευμένη ανάπτυξη, την πρόβλεψη συμπεριφορών και τη βελτιστοποίηση των υπηρεσιών.

3.2 Πρόβλεψη Αποχώρησης (Churn Prediction)

Παρότι σημαντικό μέρος της διαθέσιμης βιβλιογραφίας αφορά την πρόβλεψη αποχώρησης πελατών (customer churn) και όχι ειδικά συνεργαζόμενων καταστημάτων, οι βασικές μεθοδολογικές αρχές παραμένουν κοινές. Η πρόβλεψη churn, ανεξάρτητα από το αν η μονάδα ανάλυσης είναι ο πελάτης ή το κατάστημα, διατυπώνεται ως πρόβλημα δυαδικής ταξινόμησης και περιλαμβάνει κοινές προκλήσεις, όπως ο ορισμός της μεταβλητής-στόχου, η ανισορροπία κλάσεων και η επιλογή κατάλληλων μετρικών αξιολόγησης. Για τον λόγο αυτό, στην παρούσα εργασία αξιοποιείται η ευρύτερη βιβλιογραφία churn prediction ως μεθοδολογικό υπόβαθρο, προσαρμοσμένο στο πλαίσιο αποχώρησης συνεργαζόμενων καταστημάτων της πλατφόρμας.

Η Kumar και Chandrakala (2016) πραγματοποίησαν μια εκτενή συστηματική ανασκόπηση των μεθόδων μηχανικής μάθησης για την πρόβλεψη διαφυγής πελατών σε διαφορετικούς κλάδους, όπως ο τραπεζικός και οι τηλεπικοινωνίες. Οι ερευνητές επισημαίνουν ότι η απώλεια πελατών συνεπάγεται σημαντικά υψηλότερο κόστος απόκτησης νέων, γεγονός που καθιστά κρίσιμη τη χρήση προγνωστικών μοντέλων. Το έργο τους εξετάζει δημοφιλείς αλγορίθμους, όπως νευρωνικά δίκτυα και σύνολα ενισχυμένων ταξινομητών (Boosting Ensembles), καταλήγοντας ότι η πρόβλεψη διαφυγής αποτελεί ουσιαστικά πρόβλημα δυαδικής ταξινόμησης, το οποίο μπορεί να επιλυθεί αποτελεσματικά με επιβεβλημένες μεθόδους μηχανικής μάθησης.

Η μελέτη των Sweidan et al. (2019) επικεντρώνεται σε πραγματικά δεδομένα ηλεκτρονικού εμπορίου, τα οποία περιλαμβάνουν περισσότερους από 200.000 πελάτες και πάνω από 50 μεταβλητές, με βάση την προσέγγιση RFM (Recency, Frequency, Monetary). Σκοπός της έρευνας ήταν να αναγνωριστούν οι παράγοντες που οδηγούν στη διακοπή συνεργασίας και να κατασκευαστεί

ένα μοντέλο πρόβλεψης των “churners”. Τα αποτελέσματα δείχνουν ότι ο δείκτης διαφυγής σχετίζεται άμεσα με το πλήθος προηγούμενων αγορών: πελάτες με μόνο μία αγορά εμφάνισαν ποσοστό διαφυγής άνω του 75%, ενώ εκείνοι με τέσσερις ή περισσότερες αγορές παρουσίασαν μόλις 25%. Οι ερευνητές κατέληξαν ότι το Gradient Boosting ήταν το πιο αποδοτικό μοντέλο, επιτυγχάνοντας υψηλή ακρίβεια και καλή ισορροπία ανάμεσα σε ευαισθησία και ειδικότητα.

Οι Sari et al. (2023) εφάρμοσαν τεχνικές πρόβλεψης διαφυγής σε δεδομένα συνδρομητών τηλεπικοινωνιακής εταιρείας, συγκρίνοντας πέντε διαφορετικούς αλγόριθμους ensemble (AdaBoost, Gradient Boost, XGBoost, CatBoost, LightGBM). Σύμφωνα με τα αποτελέσματα, ο αλγόριθμος XGBoost παρουσίασε την υψηλότερη ακρίβεια, περίπου 81,2%, μετά από βελτιστοποίηση υπερπαραμέτρων μέσω Grid Search. Η μελέτη αυτή αναδεικνύει τη σημασία της σωστής ρύθμισης των μοντέλων ενίσχυσης, καθώς και την υπεροχή των ensemble τεχνικών έναντι παραδοσιακών ταξινομητών στην αντιμετώπιση προβλημάτων διαφυγής.

Μια πιο πρόσφατη συστηματική επισκόπηση από τους Imani et al. (2025) συγκρίνει σύγχρονες μεθόδους μηχανικής και βαθιάς μάθησης για πρόβλεψη διαφυγής, στο διάστημα 2020–2024. Οι συγγραφείς επισημαίνουν ότι οι ensemble μέθοδοι, όπως XGBoost και LightGBM, εξακολουθούν να κυριαρχούν σε εφαρμογές πρόβλεψης διαφυγής, ενώ οι τεχνικές βαθιάς μάθησης (όπως LSTM και CNN) χρησιμοποιούνται όλο και περισσότερο σε δεδομένα μεγάλης πολυπλοκότητας. Ωστόσο, εντοπίζονται σημαντικές προκλήσεις, όπως η ασυμμετρία κλάσεων (μεγάλη διαφορά μεταξύ churners και non-churners) και η περιορισμένη ερμηνευσιμότητα των μοντέλων. Επισημαίνεται επίσης ότι, παρά τη σημαντική πρόοδο της τεχνητής νοημοσύνης στον τομέα, παραμένουν ζητήματα όπως το concept drift και η σύνδεση των προβλέψεων με επιχειρησιακούς δείκτες απόδοσης.

Συνολικά, η βιβλιογραφία δείχνει ότι η πρόβλεψη διαφυγής αποτελεί ένα δυναμικό και εξελισσόμενο πεδίο έρευνας. Η χρήση τεχνικών ensemble και βαθιάς μάθησης έχει βελτιώσει σημαντικά την ακρίβεια των προβλέψεων, ωστόσο οι σύγχρονες μελέτες επισημαίνουν την ανάγκη για μοντέλα που συνδυάζουν προγνωστική δύναμη και ερμηνευσιμότητα, ειδικά σε περιβάλλοντα με ασύμμετρα δεδομένα, όπως αυτά των online πλατφορμών διανομής φαγητού.

4. Μεθοδολογικό Πλαίσιο και Αναλυτική Προσέγγιση

Το παρόν κεφάλαιο παρουσιάζει το μεθοδολογικό υπόβαθρο και το εύρος τεχνικών που χρησιμοποιούνται στη βιβλιογραφία για shop clustering και churn prediction. Στο πρακτικό/πειραματικό μέρος της διατριβής εφαρμόζονται και αξιολογούνται οι κατάλληλες μέθοδοι στο πλαίσιο των διαθέσιμων δεδομένων.

4.1 Προεπεξεργασία και Προετοιμασία Δεδομένων

Η Μηχανική Μάθηση (Machine Learning) αποτελεί θεμελιώδη περιοχή της τεχνητής νοημοσύνης και επικεντρώνεται στη δημιουργία αλγορίθμων που μπορούν να «μάθουν» πρότυπα από δεδομένα και να πραγματοποιούν προβλέψεις ή λήψη αποφάσεων χωρίς να έχουν ρητά προγραμματιστεί για κάθε περίπτωση. Διακρίνεται σε δύο βασικές κατηγορίες: την επιβλεπόμενη μάθηση (supervised learning), όπου ο στόχος είναι να προβλεφθεί μια γνωστή ετικέτα (label) με βάση τα χαρακτηριστικά εισόδου, και τη μη επιβλεπόμενη μάθηση (unsupervised learning), όπου επιδιώκεται η ανακάλυψη κρυφών δομών ή προτύπων χωρίς προκαθορισμένες ετικέτες.

Στο πλαίσιο εφαρμογών όπως η πρόβλεψη αποχώρησης (churn prediction), το πρόβλημα συνήθως διατυπώνεται ως δυαδική ταξινόμηση (binary classification problem), ενώ στην τμηματοποίηση καταστημάτων (shop clustering) αξιοποιούνται τεχνικές μη επιβλεπόμενης μάθησης με στόχο τον εντοπισμό ομοιογενών ομάδων με κοινά λειτουργικά ή εμπορικά χαρακτηριστικά. Ο συνδυασμός των δύο προσεγγίσεων μπορεί να υποστηρίξει τη λήψη αποφάσεων, τη στοχευμένη υποστήριξη συνεργαζόμενων καταστημάτων και τη μείωση του κινδύνου απώλειας συνεργασιών.

Κεντρικό στάδιο στη δημιουργία οποιουδήποτε μοντέλου μηχανικής μάθησης είναι η προεπεξεργασία των δεδομένων (data preprocessing). Η διαδικασία αυτή στοχεύει στην καθαρότητα και την ομοιογένεια των δεδομένων, μειώνοντας σφάλματα και αυξάνοντας τη γενικευσιμότητα των μοντέλων. Ένα από τα πιο κρίσιμα βήματα είναι η κλιμάκωση χαρακτηριστικών (feature scaling), όπως η τυποποίηση (standardization) ή άλλες μορφές κανονικοποίησης, ώστε οι μεταβλητές να βρίσκονται σε συγκρίσιμη κλίμακα και να μην κυριαρχούν εκείνες με μεγάλες τιμές ή διαφορετικές μονάδες μέτρησης. Η τυποποίηση εκφράζεται μαθηματικά ως:

$$z = \frac{x - \mu}{\sigma}$$

όπου μ είναι ο μέσος όρος και σ η τυπική απόκλιση των δεδομένων στο σύνολο εκπαίδευσης. Η διαδικασία αυτή μετατρέπει κάθε χαρακτηριστικό σε νέα μεταβλητή με μέση τιμή μηδέν και διασπορά ίση με ένα.

Σε περιπτώσεις όπου τα δεδομένα περιέχουν ακραίες τιμές ή παρουσιάζουν έντονες αποκλίσεις, χρησιμοποιείται συχνά ανθεκτική κλιμάκωση (robust scaling), η οποία βασίζεται στη διάμεσο και το ενδοτεταρτημοριακό εύρος (IQR):

$$z = \frac{x - \text{median}(x)}{IQR}, IQR = Q3 - Q1$$

Η μέθοδος αυτή μειώνει την επίδραση ακραίων παρατηρήσεων και διατηρεί μεγαλύτερη σταθερότητα στην κλιμάκωση των χαρακτηριστικών.

Ένα ακόμη σημαντικό ζήτημα είναι η αντιμετώπιση ελλειπουσών τιμών (missing values). Συνηθισμένες τεχνικές περιλαμβάνουν την αντικατάσταση με μέσο όρο (mean) ή διάμεσο (median), ενώ πιο εξελιγμένες προσεγγίσεις, όπως K-Nearest Neighbors imputation ή μοντελοστραφής συμπλήρωση, επιτρέπουν εκτίμηση βάσει συσχετισμένων χαρακτηριστικών. Για την αποφυγή διαρροής πληροφορίας (data leakage), οι μετασχηματισμοί και οι παράμετροί τους πρέπει να υπολογίζονται στο training set και να εφαρμόζονται με τον ίδιο τρόπο στα σύνολα αξιολόγησης.

Η διάσπαση των δεδομένων σε train, validation και test sets αποτελεί επίσης κρίσιμο βήμα. Σε προβλήματα με χρονική εξάρτηση, η διάσπαση οφείλει να λαμβάνει υπόψη τη χρονική σειρά (time-series split). Επιπλέον, το feature engineering, δηλαδή η δημιουργία νέων χαρακτηριστικών, αποτελεί βασικό πυλώνα για τη βελτίωση της απόδοσης και της ερμηνείας των μοντέλων. Ενδεικτικά παραδείγματα περιλαμβάνουν δείκτες Recency–Frequency–Monetary (RFM), ποσοστά ακυρώσεων, χρόνους παράδοσης και δείκτες εποχικότητας.

4.2 Τεχνικές Τμηματοποίησης

Η τμηματοποίηση (clustering) στοχεύει στον διαχωρισμό ενός συνόλου δεδομένων σε ομάδες (clusters) που παρουσιάζουν υψηλή ενδοομαδική ομοιογένεια και σαφή διαφοροποίηση μεταξύ τους. Στο πλαίσιο των επιχειρησιακών δεδομένων, όπως τα καταστήματα μιας πλατφόρμας διανομής, η διαδικασία αυτή επιτρέπει την ανάδειξη κατηγοριών καταστημάτων με παρόμοια λειτουργικά πρότυπα, διευκολύνοντας την λήψη στρατηγικών αποφάσεων (όπως προωθητικές ενέργειες, υποστήριξη ή στόχευση καμπανιών).

4.2.1 Αλγόριθμοι τμηματοποίησης

K-Means (Lloyd's Algorithm, k-means++)

Ο αλγόριθμος K-Means αποτελεί μια από τις πιο διαδεδομένες μεθόδους τμηματοποίησης. Ο στόχος του είναι η διαίρεση των δεδομένων σε K συστάδες, όπου κάθε σημείο ανήκει στην ομάδα με το πλησιέστερο κέντρο (centroid). Ο K-Means επιδιώκει την ελαχιστοποίηση του αθροίσματος τετραγωνικών αποστάσεων (Within-Cluster Sum of Squares - WCSS):

$$J = \sum_{i=1}^K \sum_{x \in C_i} |x - \mu_i|^2$$

Η εκπαίδευση του μοντέλου ακολουθεί επαναληπτική διαδικασία δύο βημάτων: (α) εκχώρηση κάθε σημείου στο πλησιέστερο κέντρο και (β) επαναυπολογισμός των κεντροειδών ως ο μέσος όρος των σημείων της συστάδας. Ο αλγόριθμος συγκλίνει σε τοπικό ελάχιστο, ενώ η αρχικοποίηση μέσω της μεθόδου k-means++ βελτιώνει τη σταθερότητα επιλέγοντας κέντρα με πιθανότητα ανάλογη της απόστασης από τα ήδη επιλεγμένα.

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

Ιεραρχική Ομαδοποίηση (Agglomerative – Ward)

Η ιεραρχική ομαδοποίηση (Hierarchical Clustering) κατασκευάζει ένα δενδρόγραμμα συγχωνεύσεων που απεικονίζει τη σχέση εγγύτητας μεταξύ ομάδων. Η μέθοδος Ward ελαχιστοποιεί την αύξηση της ενδοσυσταδικής διασποράς κατά τη συγχώνευση δύο ομάδων A και B, σύμφωνα με τη σχέση:

$$D(A, B) = \frac{|A| |B|}{|A| + |B|} |\mu_A - \mu_B|^2$$

Το κόστιμο του δενδρογράμματος σε κατάλληλο ύψος καθορίζει τον αριθμό των τελικών ομάδων, ενώ η οπτική αναπαράσταση παρέχει πολύτιμες πληροφορίες για την ερμηνεία των σχέσεων μεταξύ καταστημάτων ή άλλων επιχειρησιακών οντοτήτων.

Πυκνοτικές Μέθοδοι (DBSCAN και HDBSCAN)

Ο αλγόριθμος DBSCAN (Density-Based Spatial Clustering of Applications with Noise) ορίζει συστάδες ως περιοχές με υψηλή πυκνότητα σημείων και χαρακτηρίζει τα απομονωμένα σημεία ως θόρυβο. Χρησιμοποιεί δύο βασικές παραμέτρους: ε (ακτίνα γειτονιάς) και minPts (ελάχιστος αριθμός σημείων). Ένα σημείο θεωρείται πυρήνας όταν ισχύει:

$$|N_\epsilon(x)| \geq \text{minPts}$$

Ο DBSCAN μπορεί να εντοπίσει αυθαίρετα σχήματα και είναι ανθεκτικός σε outliers, όμως απαιτεί προσεκτική επιλογή της παραμέτρου ε, π.χ. μέσω του k-distance plot. Ο HDBSCAN αποτελεί επέκταση για δεδομένα με ετερογενή πυκνότητα.

Fuzzy C-Means (FCM)

Ο αλγόριθμος Fuzzy C-Means (FCM) αποτελεί επέκταση του K-Means, επιτρέποντας σε κάθε σημείο να ανήκει ταυτόχρονα σε περισσότερες από μία συστάδες με διαφορετικούς βαθμούς

συμμετοχής. Σε αντίθεση με το K-Means, όπου κάθε δείγμα αντιστοιχίζεται αποκλειστικά σε μία συστάδα, στο FCM κάθε δείγμα x_i συνδέεται με έναν βαθμό συμμετοχής $u_{ik} \in [0, 1]$, που εκφράζει τον βαθμό στον οποίο το δείγμα ανήκει στη συστάδα k .

Η συνάρτηση κόστους που ελαχιστοποιεί ο αλγόριθμος είναι:

$$J_m = \sum_{i=1}^N \sum_{k=1}^c (u_{ik})^m |x_i - c_k|^2, \quad m > 1$$

όπου u_{ik} είναι ο βαθμός συμμετοχής, c_k το κέντρο της συστάδας, και m ο παράγοντας ασάφειας (fuzzifier).

Οι εξισώσεις ενημέρωσης είναι:

$$c_k = \frac{\sum_{i=1}^N (u_{ik})^m x_i}{\sum_{i=1}^N (u_{ik})^m}$$

$$u_{ik} = \left(\sum_{j=1}^c \left(\frac{|x_i - c_k|}{|x_i - c_j|} \right)^{\frac{2}{m-1}} \right)^{-1}$$

Η διαδικασία επαναλαμβάνεται μέχρι η μεταβολή των κέντρων να γίνει μικρότερη από ένα όριο σύγκλισης. Ο παράγοντας m ελέγχει τον βαθμό «ασάφειας»: όταν $m \rightarrow 1$, η μέθοδος τείνει στο K-Means, ενώ για υψηλότερες τιμές οι συμμετοχές γίνονται πιο ομοιόμορφες. Το FCM είναι ιδιαίτερα χρήσιμο όταν οι συστάδες έχουν αλληλεπικαλυπτόμενα όρια, όπως σε περιπτώσεις επιχειρησιακών προφίλ.

Φασματική Τμηματοποίηση (Spectral Clustering)

Η φασματική τμηματοποίηση (Spectral Clustering) αποτελεί μία από τις πιο ισχυρές μεθόδους ομαδοποίησης για δεδομένα που δεν μπορούν να διαχωριστούν γραμμικά. Η βασική ιδέα της μεθόδου είναι η αναπαράσταση των δεδομένων ως γράφου, όπου κάθε σημείο αποτελεί έναν κόμβο και τα βάρη των ακμών εκφράζουν τον βαθμό ομοιότητας μεταξύ ζευγών σημείων. Με την αξιοποίηση των ιδιοτιμών και ιδιοδιανυσμάτων του Λαπλασιανού πίνακα του γράφου, η μέθοδος μπορεί να αποκαλύψει πολύπλοκες, μη γραμμικές δομές στα δεδομένα.

Για έναν γράφο με πίνακα ομοιότητας W και πίνακα βαθμών D , ορίζεται ο Λαπλασιανός:

$$L = D - W$$

και η κανονικοποιημένη μορφή του:

$$L_{\text{sym}} = D^{-1/2} L D^{-1/2}$$

Τα βασικά βήματα της μεθόδου είναι:

1. Κατασκευή πίνακα ομοιότητας W (π.χ. Gaussian kernel).
2. Υπολογισμός κανονικοποιημένου Λαπλασιανού L_{sym} .
3. Εξαγωγή των K μικρότερων ιδιοδιανυσμάτων του L_{sym} .
4. Εφαρμογή K-Means στα ιδιοδιανύσματα.

Η μέθοδος αποδίδει άριστα σε δεδομένα με μη γραμμικά όρια και περίπλοκες γεωμετρίες, ωστόσο απαιτεί προσεκτική επιλογή παραμέτρων και έχει υψηλότερο υπολογιστικό κόστος ($O(N^3)$).

4.2.2 Επιλογή αριθμού συστάδων και αξιολόγηση ποιότητας

Η εύρεση του κατάλληλου πλήθους συστάδων αποτελεί κρίσιμο στάδιο στη διαδικασία της τμηματοποίησης. Υπάρχουν διάφορα κριτήρια επιλογής, με τα πιο διαδεδομένα να είναι τα εξής:

Μέθοδος Elbow (Elbow Method)

Υπολογίζεται η ενδοσυσταδική απόσταση (WCSS) για διαφορετικές τιμές του K και σχεδιάζεται η καμπύλη $J(K)$. Ο βέλτιστος αριθμός συστάδων αντιστοιχεί στο σημείο όπου η μείωση του σφάλματος αρχίζει να γίνεται οριακή, το χαρακτηριστικό “γόνατο” της καμπύλης (elbow). Η μέθοδος αυτή είναι εύχρηστη, αλλά ορισμένες φορές η ύπαρξη “γόνατου” δεν είναι απολύτως ξεκάθαρη.

Gap Statistic (Tibshirani et al., 2001)

Η μέθοδος αυτή συγκρίνει την απόδοση του clustering στα πραγματικά δεδομένα με μια κατανομή αναφοράς όπου δεν υπάρχει δομή (random uniform distribution).

Για κάθε K υπολογίζεται:

$$Gap(K) = E^*[\log(W_K)] - \log(W_K)$$

όπου W_K είναι η ενδοσυσταδική απόσταση των πραγματικών δεδομένων και $E^*[\cdot]$ η αντίστοιχη τιμή στα δεδομένα αναφοράς. Ο βέλτιστος αριθμός συστάδων επιλέγεται ως ο μικρότερος K που ικανοποιεί:

$$Gap(K) \geq Gap(K + 1) - s_{K+1}$$

όπου s_K είναι η τυπική απόκλιση του Gap .

Ο Gap Statistic θεωρείται από τις πιο αξιόπιστες μεθόδους, καθώς στηρίζεται σε στατιστική σύγκριση και όχι μόνο σε οπτική επιθεώρηση. Παράλληλα, αποτελεί και τη μέθοδο που εφαρμόζεται συχνά σε αυτοματοποιημένα περιβάλλοντα όπως το SAS Model Studio για την επιλογή του μέγιστου αριθμού συστάδων.

Δείκτης Silhouette

Ο δείκτης Silhouette μετρά το πόσο καλά ταιριάζει κάθε σημείο στη συστάδα του σε σχέση με τις γειτονικές συστάδες:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

με $a(i)$ την ενδοσυσταδική απόσταση και $b(i)$ την απόσταση του σημείου από την πλησιέστερη άλλη συστάδα. Τιμές κοντά στο 1 υποδεικνύουν ισχυρή ομαδοποίηση, τιμές κοντά στο 0 δείχνουν επικάλυψη συστάδων, ενώ αρνητικές τιμές δείχνουν πιθανή κακή ταξινόμηση. Αν και είναι πιο υπολογιστικά απαιτητικός δείκτης, προσφέρει μια ολοκληρωμένη εικόνα της ποιότητας του clustering.

Στο πλαίσιο της παρούσας εργασίας, το SAS Model Studio υπολόγισε αυτόματα το βέλτιστο K μέσω του Gap Statistic, καταλήγοντας σε $K = 5$, το οποίο χρησιμοποιείται στα επόμενα υποκεφάλαια.

4.3 Μοντέλα Πρόβλεψης Αποχώρησης Καταστημάτων

Η πρόβλεψη αποχώρησης καταστημάτων (merchant churn prediction) αφορά την ανάπτυξη μοντέλων που εκτιμούν την πιθανότητα ένα συνεργαζόμενο κατάστημα να τερματίσει τη συνεργασία του με μια πλατφόρμα. Σε επιχειρησιακά περιβάλλοντα με μεγάλο πλήθος συνεργατών, η έγκαιρη αναγνώριση καταστημάτων υψηλού κινδύνου υποστηρίζει στοχευμένες ενέργειες διατήρησης και βελτιστοποιεί την κατανομή πόρων υποστήριξης. Μεθοδολογικά, το πρόβλημα διατυπώνεται συνήθως ως δυαδική ταξινόμηση (churn / non-churn), όπου τα μοντέλα εκπαιδεύονται σε ιστορικά

δεδομένα με χαρακτηριστικά απόδοσης και συμπεριφοράς καταστημάτων (π.χ. όγκος παραγγελιών, αξία συναλλαγών, αξιολογήσεις, χρόνοι, ακυρώσεις, μεταβολές δραστηριότητας) και στόχο την πρόβλεψη της ετικέτας αποχώρησης.

4.3.1 Μοντέλα ταξινόμησης για churn prediction

Λογιστική Παλινδρόμηση (Logistic Regression)

Η λογιστική παλινδρόμηση αποτελεί έναν από τους πιο διαδεδομένους αλγόριθμους δυαδικής ταξινόμησης και χρησιμοποιείται για την εκτίμηση της πιθανότητας εμφάνισης ενός συμβάντος, όπως η αποχώρηση ενός πελάτη (churn). Σε αυτό το μοντέλο, η πιθανότητα $P(Y = 1 | x)$ ότι ο πελάτης θα αποχωρήσει εκφράζεται μέσω της λογιστικής (sigmoid) συνάρτησης ως εξής:

$$P(Y = 1 | x) = \frac{1}{1 + \exp(-(w^T x + b))}$$

όπου $x = (x_1, x_2, \dots, x_p)$ είναι το διάνυσμα χαρακτηριστικών, w το διάνυσμα βαρών και b ο όρος σταθερού (bias). Ο λογάριθμος του λόγου πιθανοτήτων (log-odds) ικανοποιεί τη γραμμική σχέση:

$$\log\left(\frac{P(Y = 1 | x)}{P(Y = 0 | x)}\right) = w^T x + b$$

Η εκπαίδευση του μοντέλου πραγματοποιείται με μεγιστοποίηση της λογιστικής πιθανοφάνειας ή, ισοδύναμα, με ελαχιστοποίηση της αρνητικής λογαριθμικής απώλειας (cross-entropy):

$$\mathcal{L}(w) = - \sum_{i=1}^N (y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i))$$

Στην πράξη, η βελτιστοποίηση γίνεται μέσω του αλγορίθμου gradient descent και των παραλλαγών του. Για να αποφευχθεί η υπερπροσαρμογή (overfitting) και να επιτευχθεί επιλογή χαρακτηριστικών, συχνά ενσωματώνεται κανονικοποίηση L1 (LASSO) ή L2 (Ridge) στα βάρη του μοντέλου. Η λογιστική παλινδρόμηση προσφέρει υψηλή ερμηνευσιμότητα, καθώς τα βάρη επιτρέπουν την κατανόηση της επίδρασης κάθε χαρακτηριστικού στην πιθανότητα αποχώρησης. Ωστόσο, το μοντέλο βασίζεται στην υπόθεση γραμμικότητας στον χώρο των χαρακτηριστικών και μπορεί να αποδώσει ανεπαρκώς όταν τα δεδομένα παρουσιάζουν σύνθετες μη γραμμικές συσχέτισεις.

Δέντρα Απόφασης (Decision Trees)

Τα Δέντρα Απόφασης (Decision Trees) αποτελούν μία από τις πιο διαδεδομένες και ερμηνεύσιμες τεχνικές επιβλεπόμενης μάθησης, οι οποίες χρησιμοποιούνται για προβλήματα ταξινόμησης και παλινδρόμησης. Στο πλαίσιο της πρόβλεψης αποχώρησης πελατών (churn prediction), τα δέντρα απόφασης επιτρέπουν τον διαχωρισμό των πελατών σε ομάδες βάσει χαρακτηριστικών όπως η συχνότητα αγορών, το χρηματικό ποσό, η δραστηριότητα στην πλατφόρμα ή ο χρόνος από την τελευταία παραγγελία, παράγοντας κανόνες της μορφής IF-THEN που μπορούν να ερμηνευτούν άμεσα από τη διοίκηση. Η βασική ιδέα πίσω από τα δέντρα απόφασης είναι η αναδρομική διάσπαση του συνόλου δεδομένων σε υποσύνολα που είναι όσο το δυνατόν πιο καθαρά (homogeneous) ως προς την ετικέτα-στόχο. Σε κάθε κόμβο του δέντρου επιλέγεται το χαρακτηριστικό που προσφέρει τη μέγιστη μείωση στην αβεβαιότητα της πληροφορίας, δηλαδή τη μέγιστη τιμή του Information Gain. Το

κριτήριο πληροφορίας μπορεί να βασίζεται στην εντροπία (Entropy) ή στη μέτρηση Gini (Gini Impurity). Η εντροπία ορίζεται ως:

$$Entropy(S) = - \sum_{i=1}^c p_i \log_2(p_i)$$

όπου p_i είναι η αναλογία των στοιχείων της κλάσης i στο σύνολο S . Η εντροπία εκφράζει το επίπεδο αβεβαιότητας ή ανακατεμένης πληροφορίας. Όσο πιο καθαρή είναι μια ομάδα (δηλαδή όσο ένα χαρακτηριστικό διαχωρίζει καλύτερα τις κλάσεις), τόσο μικρότερη είναι η εντροπία της. Το Information Gain μετρά τη μείωση της εντροπίας μετά τη διάσπαση του συνόλου S βάσει ενός χαρακτηριστικού A και υπολογίζεται ως:

$$Gain(S, A) = Entropy(S) - \sum_v \left(\frac{|S_v|}{|S|} \cdot Entropy(S_v) \right)$$

Η ιδιότητα αυτή εξασφαλίζει ότι το χαρακτηριστικό με το μεγαλύτερο Information Gain επιλέγεται για τη διάσπαση του κόμβου. Εναλλακτικά, μπορεί να χρησιμοποιηθεί ο δείκτης Gini, ο οποίος ορίζεται ως:

$$Gini(S) = 1 - \sum_{i=1}^c (p_i)^2$$

Η τιμή του δείκτη Gini κυμαίνεται από 0 (απόλυτα καθαρή συστάδα) έως 0.5 (μέγιστη αβεβαιότητα). Κατά την εκπαίδευση, το δέντρο συνεχίζει να αναπτύσσεται μέχρι να ικανοποιηθεί κάποιο κριτήριο τερματισμού, όπως το ελάχιστο μέγεθος φύλλου ή το μέγιστο βάθος. Για να αποφευχθεί το φαινόμενο του υπερπροσαρμογής (overfitting), εφαρμόζονται τεχνικές κλαδέματος (pruning), οι οποίες αφαιρούν κόμβους που δεν προσφέρουν ουσιαστική βελτίωση στην απόδοση. Το κλάδεμα μπορεί να είναι προληπτικό (pre-pruning) ή εκ των υστέρων (post-pruning). Στην πράξη, τα Decision Trees αποτελούν τη βάση για πιο σύνθετους και ισχυρούς αλγορίθμους όπως τα Random Forest και τα Gradient Boosting Machines.

Τυχαίο Δάσος (Random Forest)

Τα Τυχαία Δάση (Random Forest) είναι μια ισχυρή μέθοδος ensemble μάθησης, η οποία αποτελείται από ένα πλήθος δέντρων απόφασης. Κάθε δέντρο εκπαιδεύεται σε τυχαίο υποσύνολο του συνόλου δεδομένων μέσω bootstrap sampling, ενώ σε κάθε διαχωρισμό κόμβου επιλέγεται τυχαίο υποσύνολο χαρακτηριστικών. Με αυτόν τον τρόπο μειώνεται η συσχέτιση μεταξύ των δέντρων και περιορίζεται η διασπορά (variance) του μοντέλου. Η τελική πρόβλεψη κατηγορίας προκύπτει μέσω πλειοψηφικής ψήφου των προβλέψεων όλων των δέντρων:

$$\hat{y}(x) = \text{mode}\{h_t(x) : t = 1, \dots, T\}$$

όπου $h_t(x)$ είναι η πρόβλεψη του t -οστού δέντρου.

Κάθε δέντρο χωρίζει τα δεδομένα χρησιμοποιώντας κριτήρια όπως ο δείκτης Gini ή η εντροπία πληροφορίας (information gain). Το Random Forest μπορεί να μοντελοποιήσει σύνθετες και μη γραμμικές σχέσεις μεταξύ των χαρακτηριστικών, να λάβει υπόψη αλληλεπιδράσεις μεταξύ μεταβλητών και παρουσιάζει υψηλή ανθεκτικότητα στον θόρυβο και στα μη-παραμετρικά δεδομένα.

Επιπλέον, παρέχει μέτρα σημαντικότητας χαρακτηριστικών (feature importance), διευκολύνοντας την αξιολόγηση της συμβολής κάθε μεταβλητής. Μειονεκτήματα αποτελούν η σχετικά περιορισμένη ερμηνευσιμότητα (θεωρείται "μαύρο κουτί") και το αυξημένο υπολογιστικό κόστος για μεγάλα σύνολα δεδομένων.

Gradient Boosting Machines

Οι Gradient Boosting Machines (GBM) αποτελούν μεθόδους ενίσχυσης (boosting) που κατασκευάζουν ένα ισχυρό ταξινομητή συνδυάζοντας μεγάλο αριθμό αδύναμων μοντέλων, συνήθως δέντρων απόφασης μικρού βάθους. Η βασική ιδέα του boosting είναι ότι κάθε νέο δέντρο εκπαιδεύεται με στόχο να διορθώσει τα σφάλματα των προηγούμενων. Το συνολικό μοντέλο εκφράζεται ως:

$$F_m(x) = \sum_{t=1}^m \gamma_t h_t(x)$$

όπου $h_t(x)$ είναι το t -οστό δέντρο και γ_t ο συντελεστής βάρους του. Η εκπαίδευση πραγματοποιείται διαδοχικά: κάθε νέο δέντρο h_m προσαρμόζεται ώστε να μειώσει τα σφάλματα του μοντέλου στο προηγούμενο βήμα F_{m-1} , ελαχιστοποιώντας μια συνάρτηση κόστους $L(y, F(x))$ μέσω βαθμωτής καθόδου.

Η ενημέρωση του μοντέλου δίνεται συνήθως από τον κανόνα:

$$F_m(x) = F_{m-1}(x) + \eta \rho_m h_m(x)$$

όπου η είναι ο ρυθμός μάθησης (learning rate) και ρ_m ο βέλτιστος συντελεστής για το δέντρο h_m .

Παραλλαγές της μεθόδου χρησιμοποιούν διαφορετικές συναρτήσεις κόστους όπως το log-loss (για ταξινόμηση) ή το squared loss (για παλινδρόμηση). Δημοφιλείς υλοποιήσεις του GBM περιλαμβάνουν τα XGBoost, LightGBM και CatBoost, τα οποία προσφέρουν σημαντικές βελτιώσεις στην ταχύτητα και σε μεθόδους κανονικοποίησης για την αποφυγή υπερπροσαρμογής. Τα GBM επιτυγχάνουν εξαιρετική ακρίβεια σε περίπλοκα σύνολα δεδομένων, αλλά απαιτούν σχολαστικό ρυθμισμό υπερπαραμέτρων (π.χ. βάθος δέντρου, learning rate, αριθμός δέντρων). Επιπλέον, λόγω της σειριακής φύσης της εκπαίδευσης, είναι πιο υπολογιστικά απαιτητικά σε σχέση με μεθόδους όπως τα Random Forest.

Υποστήριξη Διανυσμάτων (SVM)

Οι Μηχανές Υποστήριξης Διανυσμάτων (Support Vector Machines, SVM) είναι ισχυροί αλγόριθμοι ταξινόμησης που επιδιώκουν να βρουν το διαχωριστικό υπερεπίπεδο (hyperplane) που μεγιστοποιεί το περιθώριο (margin) μεταξύ δύο κλάσεων. Στη γραμμικά διαχωρίσιμη περίπτωση, το βέλτιστο υπερεπίπεδο υπολογίζεται λύνοντας το εξής πρόβλημα βελτιστοποίησης:

$$\min_{w,b} \frac{1}{2} |w|^2$$

υπό τον περιορισμό:

$$y_i(w^T x_i + b) \geq 1$$

Για τα μη γραμμικά διαχωρίσιμα δεδομένα εισάγονται μεταβλητές χαλάρωσης ξ_i , και το πρόβλημα επεκτείνεται σε:

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_i \xi_i$$

με περιορισμούς:

$$y_i(w^\top x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0.$$

όπου C είναι παράμετρος που ελέγχει το κόστος των σφαλμάτων (trade-off μεταξύ περιθωρίου και λάθους ταξινόμησης). Για να αντιμετωπιστούν μη γραμμικές σχέσεις, χρησιμοποιείται η τεχνική του πυρήνα (kernel trick), όπου ένα kernel $K(x, x')$ όπως ο Gaussian RBF ή πολυωνυμικός πυρήνας επιτρέπει την προβολή των δεδομένων σε υψηλότερη διάσταση χωρίς ρητό μετασχηματισμό, καθιστώντας δυνατή τη γραμμική διάκριση σε αυτόν τον χώρο. Τα SVM προσφέρουν θεωρητικά βέλτιστο περιθώριο και λειτουργούν πολύ καλά σε μεσαίου μεγέθους δεδομένα με καθαρά διαχωρίσιμες κλάσεις. Ωστόσο, η επιλογή του πυρήνα, της παραμέτρου C και του bandwidth του RBF απαιτεί προσεκτικό ρυθμισμό. Επιπλέον, η επεκτασιμότητά τους περιορίζεται σε πολύ μεγάλα σύνολα δεδομένων σε σχέση με πιο απλές μεθόδους.

Νευρωνικά Δίκτυα (Neural Networks)

Τα Νευρωνικά Δίκτυα (Neural Networks – NN) αποτελούν μια ιδιαίτερα ισχυρή κατηγορία μοντέλων μηχανικής μάθησης, τα οποία εμπνέονται από τη λειτουργία των βιολογικών νευρώνων. Ένα από τα βασικότερα μοντέλα είναι το Πολυεπίπεδο Perceptron (Multi-Layer Perceptron – MLP), το οποίο αποτελείται από διαδοχικά επίπεδα τεχνητών νευρώνων. Κάθε νευρώνας λαμβάνει ένα σύνολο εισόδων, τις σταθμίζει μέσω των αντίστοιχων βαρών και παράγει μια έξοδο μέσω μιας μη γραμμικής συνάρτησης ενεργοποίησης. Η έξοδος του νευρώνα j στο επίπεδο l δίνεται από:

$$a_j^{(l)} = \phi \left(\sum_i w_{ji}^{(l)} a_i^{(l-1)} + b_j^{(l)} \right)$$

όπου:

- $\phi(\cdot)$ = συνάρτηση ενεργοποίησης (sigmoid, tanh, ReLU),
- $w_{ji}^{(l)}$ = βάρη που συνδέουν τον νευρώνα i με τον νευρώνα j ,
- $b_j^{(l)}$ = όρος μετατόπισης (bias),
- $a_i^{(l-1)}$ = έξοδοι του προηγούμενου επιπέδου.

Η εκπαίδευση του δικτύου πραγματοποιείται μέσω της μεθόδου οπισθογώνιας διάδοσης (backpropagation), η οποία ελαχιστοποιεί μια συνάρτηση κόστους, συνήθως την cross-entropy, προσαρμόζοντας τα βάρη με βάση το σφάλμα της πρόβλεψης. Η ενημέρωση των βαρών δίνεται από:

$$w_{ij}^{(l)} \leftarrow w_{ij}^{(l)} - \eta \frac{\partial L}{\partial w_{ij}^{(l)}}$$

όπου L είναι η συνάρτηση απώλειας και η ο ρυθμός μάθησης (learning rate).

Το τελικό επίπεδο εξόδου του δικτύου χρησιμοποιεί συνήθως μία από τις παρακάτω συναρτήσεις:

- Sigmoid (για δυαδική ταξινόμηση):

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

- Softmax (για πολυταξική ταξινόμηση):

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$$

Τα Νευρωνικά Δίκτυα μπορούν να προσεγγίσουν εξαιρετικά πολύπλοκες μη γραμμικές συναρτήσεις και να αποτυπώσουν αλληλεπιδράσεις μεταξύ χαρακτηριστικών, αρκεί να υπάρχει επαρκής ποσότητα δεδομένων και κατάλληλος ρυθμός παραμετροποίησης. Ωστόσο, απαιτούν προσεκτική επιλογή υπερπαραμέτρων (learning rate, batch size, αριθμός επιπέδων, dropout) και είναι επιρρεπή στην υπερπροσαρμογή χωρίς τεχνικές κανονικοποίησης (regularization). Παρά τη μειωμένη ερμηνευσιμότητά τους σε σχέση με γραμμικά ή δένδροειδή μοντέλα (όπως η Λογιστική Παλινδρόμηση και τα Δέντρα Απόφασης), τα Νευρωνικά Δίκτυα αποδίδουν ιδιαίτερα καλά σε εφαρμογές πρόβλεψης αποχώρησης πελατών, όπου τα δεδομένα είναι υψηλής διάστασης και εμφανίζουν πολύπλοκα, μη γραμμικά πρότυπα.

4.3.2 Αξιολόγηση, ανισορροπία κλάσεων, επιλογή κατωφλίου και βαθμονόμηση

Η αξιολόγηση των μοντέλων πρόβλεψης αποχώρησης (churn) αποτελεί κρίσιμο στάδιο, καθώς συνδέει την απόδοση του αλγορίθμου με τον πραγματικό επιχειρησιακό στόχο: την έγκαιρη αναγνώριση καταστημάτων υψηλού κινδύνου με ελεγχόμενο κόστος παρεμβάσεων. Στο παρόν υποκεφάλαιο παρουσιάζονται οι βασικές μετρικές, τεχνικές αντιμετώπισης ανισορροπίας, μέθοδοι επιλογής κατωφλίου και διαδικασίες βαθμονόμησης, οι οποίες αποτελούν αναπόσπαστο μέρος της αξιολόγησης μοντέλων μηχανικής μάθησης.

Πίνακας σύγχυσης και βασικές μετρικές

Ο πίνακας σύγχυσης περιλαμβάνει τέσσερις ποσότητες:

- True Positives (TP): σωστά προβλεπόμενοι churners
- False Positives (FP): μη-churners που προβλέφθηκαν ως churners
- True Negatives (TN): σωστά προβλεπόμενοι μη-churners
- False Negatives (FN): churners που δεν ανιχνεύθηκαν

Από τα παραπάνω προκύπτουν οι θεμελιώδεις μετρικές:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

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

Σε έντονα ανισόρροπες περιπτώσεις όπως συνήθως συμβαίνει στα προβλήματα churn όπου οι αποχωρούντες είναι μειοψηφία, η καμπύλη Precision-Recall και η μετρική PR-AUC παρέχουν συχνά πιο κατατοπιστική εικόνα από τη ROC-AUC, η οποία τείνει να είναι αισιόδοξη σε datasets με μεγάλη πλειοψηφική κλάση.

Ανισορροπία κλάσεων

Η ανισορροπία αποτελεί χαρακτηριστικό γνώρισμα των εφαρμογών churn, όπου συνήθως ένα μικρό ποσοστό πελατών εγκαταλείπει την υπηρεσία. Η άνιση κατανομή μπορεί να οδηγήσει σε μοντέλα που “μαθαίνουν” να προβλέπουν μόνο την πλειοψηφία, επιτυγχάνοντας υψηλή ακρίβεια αλλά χαμηλή ανάκτηση των πραγματικών churners.

Συνήθεις τεχνικές αντιμετώπισης είναι:

1. Στάθμιση κλάσεων (class weights): αυξημένη “ποινή” για λάθη στη μειοψηφία.
2. Undersampling: μείωση των παραδειγμάτων της πλειοψηφίας.
3. Oversampling / SMOTE: δημιουργία ή σύνθεση νέων παραδειγμάτων της μειοψηφίας.
4. Προσαρμογή κατωφλίου (threshold tuning): αύξηση του Recall για στρατηγικές υψηλής κάλυψης κινδύνου.

Επιλογή κατωφλίου με κριτήριο κόστους-οφέλους

Το προεπιλεγμένο κατώφλι 0.5 της πιθανότητας churn σπάνια οδηγεί στο βέλτιστο επιχειρησιακό αποτέλεσμα. Έστω κατώφλι τ . Ένα απλοποιημένο κριτήριο οικονομικού οφέλους μπορεί να εκφραστεί ως:

$$Profit(\tau) = TP(\tau) \cdot V - (TP(\tau) + FP(\tau)) \cdot C$$

όπου:

- V : οικονομική αξία διατήρησης ενός πελάτη
- C : κόστος παρέμβασης (επικοινωνία, κίνητρα)

Το βέλτιστο κατώφλι προκύπτει συνήθως μέσω:

- σάρωσης της καμπύλης Precision–Recall,
- βελτιστοποίησης σε επικυρωτικό σύνολο,
- ή χρήσης του δείκτη Youden για ισορροπία ευαισθησίας–ειδικότητας:

$$J = TPR - FPR$$

Βαθμονόμηση πιθανοτήτων

Σε πολλά μοντέλα ταξινόμησης, οι παραγόμενες τιμές δεν αντιστοιχούν απαραίτητα σε καλά βαθμονομημένες πιθανότητες. Η βαθμονόμηση (calibration) εξασφαλίζει ότι η εκτιμώμενη πιθανότητα churn αντανακλά πραγματικές αναλογίες κινδύνου, π.χ. ότι μια πρόβλεψη 0.30 σημαίνει πως περίπου το 30% των αντίστοιχων πελατών πράγματι αποχωρεί. Αυτό είναι κρίσιμο σε εφαρμογές λήψης αποφάσεων, όπου οι πιθανότητες αποτελούν είσοδο σε στρατηγικές κόστους–οφέλους.

Οι δύο πιο διαδεδομένες μέθοδοι βαθμονόμησης είναι οι εξής:

(α) Platt Scaling

Το Platt scaling εφαρμόζει logistic regression πάνω στις μη βαθμονομημένες εξόδους του μοντέλου $f(x)$. Η τελική βαθμονομημένη πιθανότητα δίνεται από:

$$p = \frac{1}{1 + \exp(Af(x) + B)}$$

Οι παράμετροι A και B εκτιμώνται σε ξεχωριστό σύνολο επικύρωσης. Το Platt scaling έχει χαμηλή πολυπλοκότητα και λειτουργεί ιδιαίτερα καλά σε μοντέλα με μονοτονική σχέση score–πιθανότητας.

(β) Ισοτονική Παλινδρόμηση (Isotonic Regression)

Η ισοτονική παλινδρόμηση προσαρμόζει μια μονοτονικά μη φθίνουσα συνάρτηση πάνω στα ζεύγη (score, πραγματική ετικέτα), χωρίς παραμετρικές υποθέσεις. Είναι πιο ευέλικτη από το Platt scaling, αλλά μπορεί να υπερπροσαρμοστεί όταν τα δεδομένα βαθμονόμησης είναι περιορισμένα.

Η ακρίβεια της βαθμονόμησης εξετάζεται μέσω:

- Reliability diagrams, τα οποία συγκρίνουν προβλεπόμενες πιθανότητες με πραγματικές συχνότητες
- Brier Score, το οποίο μετρά το σφάλμα πρόβλεψης σε επίπεδο πιθανότητας:

$$Brier = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2$$

Χαμηλότερο Brier score υποδηλώνει καλύτερη βαθμονόμηση.

Παρακολούθηση μεταβολών (model drift)

Μετά την ανάπτυξη του μοντέλου σε παραγωγή, οι κατανομές χαρακτηριστικών και στόχων ενδέχεται να μεταβληθούν με τον χρόνο (concept drift, data drift). Ένας διαδεδομένος δείκτης παρακολούθησης είναι ο Population Stability Index (PSI):

$$PSI = \sum_b (p_b - q_b) \ln \left(\frac{p_b}{q_b} \right)$$

Τιμές:

- $PSI < 0.1 \rightarrow$ σταθερό
- $0.1 \leq PSI < 0.25 \rightarrow$ μέτρια αλλαγή
- $PSI \geq 0.25 \rightarrow$ σημαντικό drift, ανάγκη retraining

5. Μελέτη Περίπτωσης μιας Ηλεκτρονικής Πλατφόρμας Διανομής Φαγητού

5.1 Περιγραφή και Δομή Συνόλου Δεδομένων

Το σύνολο δεδομένων που χρησιμοποιείται στην παρούσα διατριβή προέρχεται από πλατφόρμα ηλεκτρονικής παραγγελίας φαγητού και αφορά συνεργαζόμενα καταστήματα εστίασης σε διάφορες πόλεις της Ελλάδας. Τα δεδομένα είναι αποθηκευμένα σε περιβάλλον SAS Viya και περιλαμβάνουν συνολικά 17.238 εγγραφές και 22 μεταβλητές, οι οποίες αποτυπώνουν γεωγραφικά, λειτουργικά και εμπορικά χαρακτηριστικά των καταστημάτων.

Οι μεταβλητές διακρίνονται σε κατηγορικές και αριθμητικές. Στις κατηγορικές μεταβλητές περιλαμβάνονται, μεταξύ άλλων, η πόλη στην οποία δραστηριοποιείται το κατάστημα, η περιοχή, η κατηγορία κουζίνας και ο τύπος καταστήματος ή δραστηριότητας. Επιπλέον, υπάρχουν αρκετοί δυαδικοί δείκτες (τιμές True/False) που περιγράφουν τη σχέση του καταστήματος με την πλατφόρμα, όπως αν οι παραγγελίες εκτελούνται από διανομείς της πλατφόρμας, αν το κατάστημα ανήκει σε

αλυσίδα, αν προωθείται ιδιαίτερα στην εφαρμογή, αν συμμετέχει σε προωθητικές ενέργειες, αν προσφέρει υπηρεσία Pinata και αν υπάρχει ενεργός ανταγωνιστής στην ίδια περιοχή.

Οι αριθμητικές μεταβλητές αποτυπώνουν κυρίως δείκτες ζήτησης, ποιότητας υπηρεσίας και οικονομικής επίδοσης. Ενδεικτικά, περιλαμβάνεται η μέση βαθμολογία του καταστήματος στην πλατφόρμα, ο μέσος χρόνος παράδοσης σε λεπτά, ο αριθμός αξιολογημένων παραγγελιών, οι ημέρες παρουσίας του καταστήματος στην πλατφόρμα, ο αριθμός παραγγελιών του τελευταίου τριμήνου και η ελάχιστη αξία παραγγελίας. Επιπλέον, υπάρχουν σύνθετοι δείκτες που αφορούν τη μέση μηνιαία απόδοση των καταστημάτων, όπως ο μέσος μηνιαίος αριθμός παραγγελιών, ο μέσος μηνιαίος αριθμός μοναδικών χρηστών, ο μέσος μηνιαίος τζίρος σε ευρώ και η μέση μηνιαία προμήθεια που αποδίδεται στην πλατφόρμα.

Σε επίπεδο πληρότητας των δεδομένων, η πλειονότητα των μεταβλητών δεν παρουσιάζει ελλείπουσες τιμές. Εξαιρεση αποτελεί η μεταβλητή της μέσης βαθμολογίας, στην οποία εμφανίζεται ένα σύνολο καταστημάτων χωρίς διαθέσιμη αξιολόγηση από τους χρήστες, καθώς και ένας πολύ μικρός αριθμός ελλείψεων στην ελάχιστη αξία παραγγελίας. Οι υπόλοιπες μεταβλητές είναι πλήρεις για όλες τις εγγραφές, γεγονός που διευκολύνει τη μεταγενέστερη χρήση τους τόσο στην ανάλυση συσταδοποίησης όσο και στην πρόβλεψη διαφυγής συνεργαζόμενων καταστημάτων.

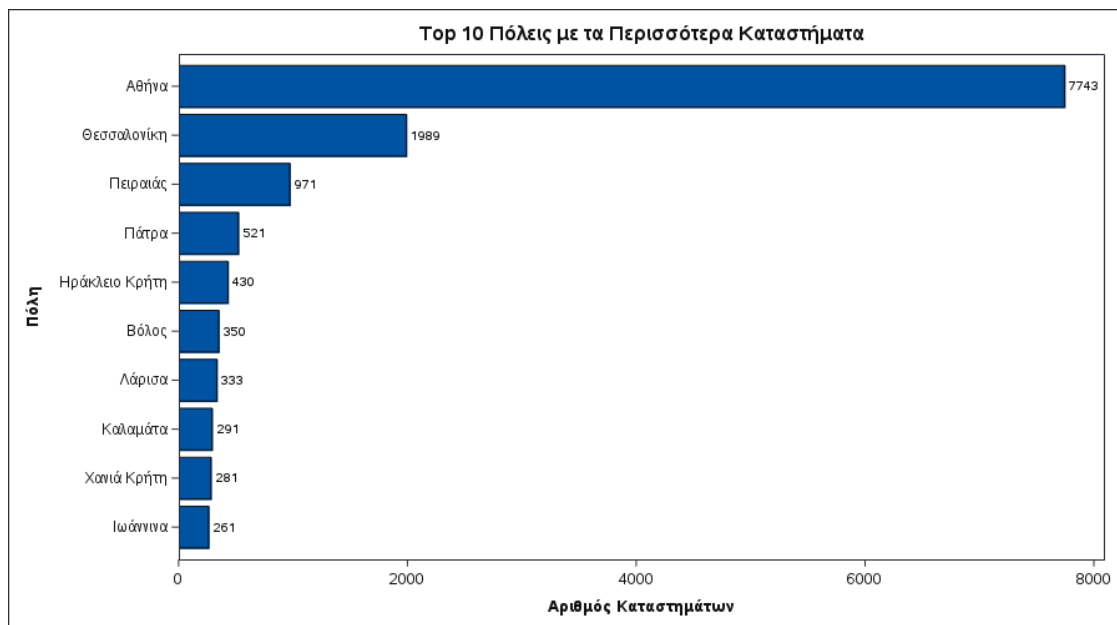
Συνολικά, η ποικιλία των μεταβλητών επιτρέπει την προσέγγιση των καταστημάτων από πολλαπλές οπτικές: γεωγραφική (πόλη, περιοχή), λειτουργική (χρόνος παράδοσης, ημέρες παρουσίας στην πλατφόρμα), ποιοτική (βαθμολογίες, αριθμός αξιολογήσεων) και οικονομική (όγκοι παραγγελιών, τζίρος, προμήθειες). Στα επόμενα υποκεφάλαια παρουσιάζεται αρχικά η διερευνητική ανάλυση των βασικών μεταβλητών και στη συνέχεια η εφαρμογή των τεχνικών συσταδοποίησης και πρόβλεψης διαφυγής που περιγράφηκαν στο θεωρητικό μέρος της εργασίας.

5.2 Εξερευνητική Ανάλυση Δεδομένων (EDA)

Η εξερευνητική ανάλυση δεδομένων (Exploratory Data Analysis) έχει στόχο να προσφέρει μια συνολική εικόνα του δικτύου καταστημάτων της πλατφόρμας, πριν την εφαρμογή των μεθόδων συσταδοποίησης και πρόβλεψης διαφυγής. Στην ενότητα αυτή παρουσιάζονται τα βασικά περιγραφικά στατιστικά στοιχεία και τα σημαντικότερα πρότυπα που προκύπτουν από τα δεδομένα, με έμφαση στη γεωγραφική κατανομή των καταστημάτων, στις κατηγορίες κουζίνας, στις βαθμολογίες και στους χρόνους παράδοσης, καθώς και σε δείκτες οικονομικής απόδοσης.

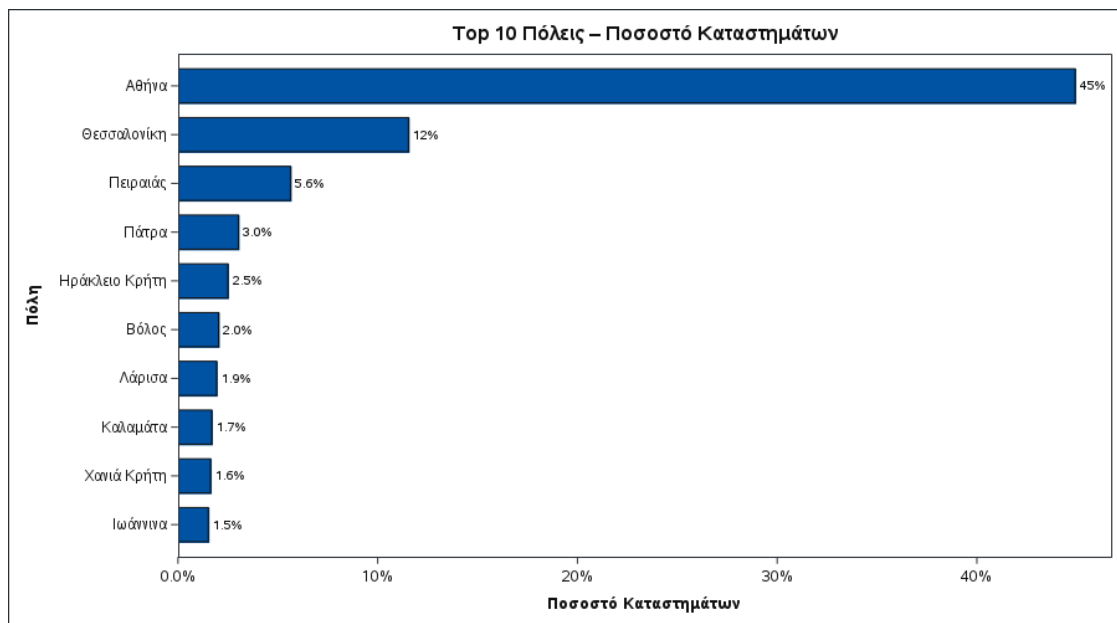
Γεωγραφική κατανομή καταστημάτων

Η γεωγραφική κατανομή των συνεργαζόμενων καταστημάτων στην πλατφόρμα παρουσιάζει ιδιαίτερο ενδιαφέρον, καθώς αναδεικνύει τον βαθμό διεξόδου της υπηρεσίας σε διαφορετικές περιοχές της χώρας και τις διαφορές που ενδέχεται να επηρεάζουν τη λειτουργία των καταστημάτων. Η ανάλυση συχνοτήτων ανά πόλη δείχνει σαφή υπεροχή των μεγάλων αστικών κέντρων, με την Αθήνα και τη Θεσσαλονίκη να συγκεντρώνουν το μεγαλύτερο μέρος του συνολικού δικτύου.



Εικόνα 1 – Κατανομή Καταστημάτων ανά Πόλη

Σε απόλυτους αριθμούς, η Αθήνα διαθέτει το μεγαλύτερο πλήθος συνεργαζόμενων καταστημάτων, σημαντικά υψηλότερο από κάθε άλλη πόλη. Ακολουθεί η Θεσσαλονίκη, ενώ οι υπόλοιπες πόλεις, όπως ο Πειραιάς, η Πάτρα και το Ηράκλειο, εμφανίζουν μεν μικρότερο αριθμό καταστημάτων, αλλά εξακολουθούν να αποτελούν σημαντικές αγορές.



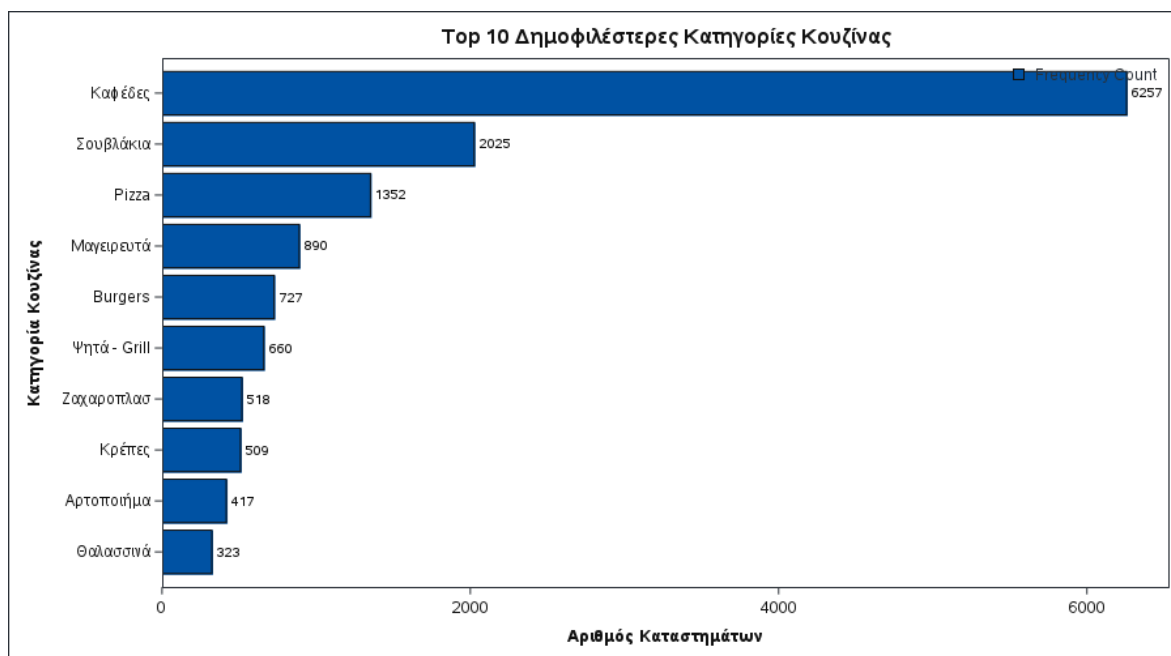
Εικόνα 2 – Ποσοστό Καταστημάτων ανά Πόλη

Σε ποσοστιαία βάση, η Αθήνα αντιπροσωπεύει περίπου το 45% του συνόλου των καταστημάτων της πλατφόρμας, ενώ η Θεσσαλονίκη ακολουθεί με 12%. Οι υπόλοιπες πόλεις παρουσιάζουν μικρότερα ποσοστά, συχνά κάτω από 5%, γεγονός που αναδεικνύει την ισχυρή

συγκέντρωση στα μεγάλα αστικά κέντρα. Η έντονη αυτή συγκέντρωση αντανακλά τόσο την πληθυσμιακή κατανομή της χώρας όσο και την αυξημένη υιοθέτηση ψηφιακών υπηρεσιών στις μεγαλύτερες πόλεις. Τα ευρήματα αυτά είναι σημαντικά για τα επόμενα στάδια της ανάλυσης, καθώς η γεωγραφική μεταβλητή ενδέχεται να επηρεάσει τη συσταδοποίηση και τους παράγοντες διαφυγής των καταστημάτων.

Κατηγορίες κουζίνας

Η ανάλυση των κατηγοριών κουζίνας αποτυπώνει τη διάρθρωση του δικτύου καταστημάτων της πλατφόρμας και αναδεικνύει τις κατηγορίες που κυριαρχούν στην αγορά. Το γράφημα παρουσιάζει τις δέκα πιο συχνές κατηγορίες κουζίνας, επιτρέποντας τη σύγκριση της σχετικής παρουσίας τους στο σύνολο των συνεργαζόμενων καταστημάτων.



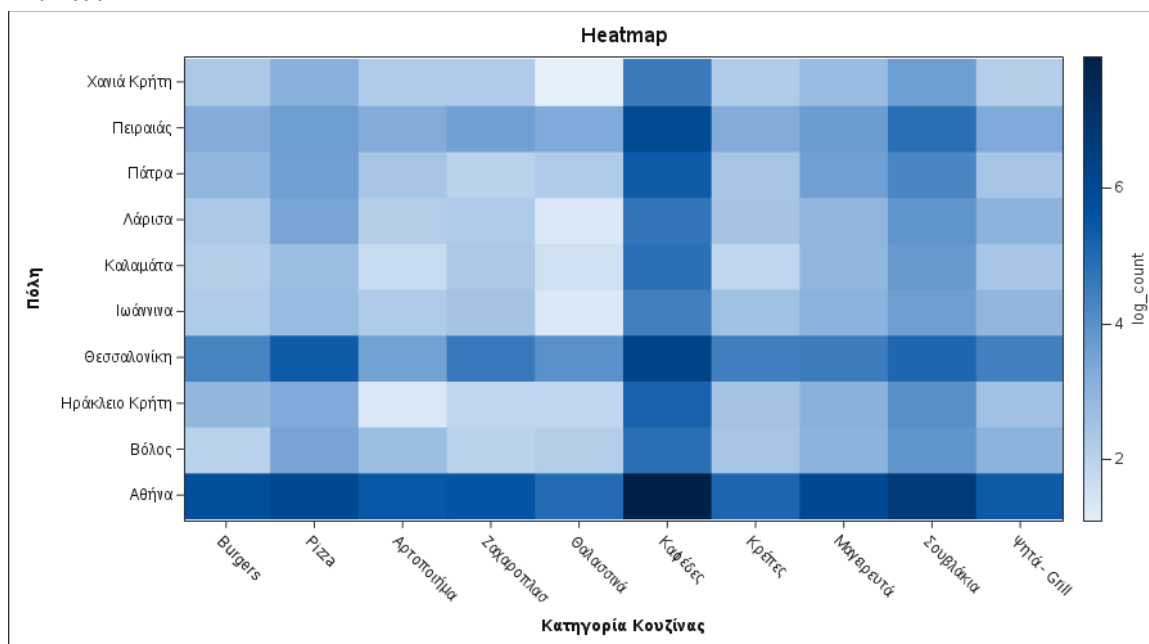
Εικόνα 3 – Κατανομή Καταστημάτων ανά Κατηγορία Κουζίνας

Η κατηγορία «Καφέδες» κυριαρχεί με 6.257 καταστήματα, γεγονός που αντανακλά την ιδιαίτερα υψηλή ζήτηση για ροφήματα και προϊόντα ταχείας εξυπηρέτησης. Ακολουθούν τα «Σουβλάκια» και η «Pizza», με 2.025 και 1.352 καταστήματα αντίστοιχα, δύο κατηγορίες που αποτελούν σταθερές επιλογές γρήγορου φαγητού με ευρεία απήχηση. Οι κατηγορίες «Μαγειρευτά» και «Burgers» εμφανίζουν επίσης αξιοσημείωτη παρουσία, υποδηλώνοντας το αυξανόμενο ενδιαφέρον για πιο ολοκληρωμένα γεύματα ή σύγχρονες προτάσεις εστίασης. Αντίθετα, κατηγορίες όπως τα «Ζαχαροπλαστεία», οι «Κρέπες» και τα «Αρτοποιεία» καταγράφουν μικρότερο αριθμό καταστημάτων, αλλά παραμένουν σημαντικές για τη συνολική ποικιλία της πλατφόρμας. Η κατανομή αυτή αναδεικνύει τις προτιμήσεις των καταναλωτών και την κυριαρχία ειδών υψηλής κατανάλωσης σε σχέση με πιο εξειδικευμένες επιλογές.

Συνδυασμός πόλης και κατηγορίας κουζίνας

Για την καλύτερη κατανόηση της σχέσης ανάμεσα στη γεωγραφική τοποθεσία και τον τύπο κουζίνας, δημιουργήθηκε heatmap που απεικονίζει τον αριθμό των καταστημάτων ανά συνδυασμό πόλης και κατηγορίας. Η οπτικοποίηση αυτή επιτρέπει την αναγνώριση προτύπων που δεν είναι ορατά μέσα

από τις απλές συχνότητες, παρουσιάζοντας με σαφήνεια τις κατηγορίες που κυριαρχούν σε κάθε περιοχή.



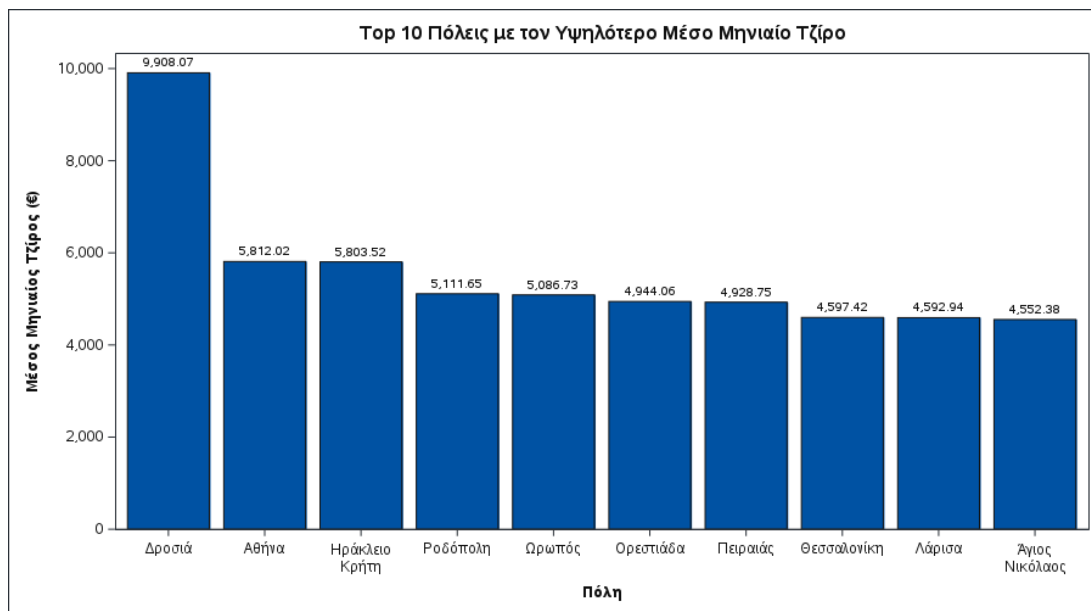
Εικόνα 4 – Κατανομή Κατηγοριών Κουζίνας ανά Πόλη

Η ανάλυση του heatmap δείχνει ότι η Αθήνα παρουσιάζει τη μεγαλύτερη πυκνότητα καταστημάτων σε όλες σχεδόν τις κατηγορίες κουζίνας, γεγονός που αντικατοπτρίζει το μέγεθος και την ποικιλομορφία της τοπικής αγοράς. Ιδιαίτερα υψηλή συγκέντρωση παρατηρείται στις κατηγορίες «Καφέδες», «Σουβλάκια» και «Pizza», οι οποίες αποτελούν τις πιο δημοφιλείς επιλογές των χρηστών και καλύπτουν μεγάλο εύρος περιοχών της πόλης. Η Θεσσαλονίκη εμφανίζει επίσης ισχυρή παρουσία, κυρίως στις κατηγορίες «Καφέδες», «Σουβλάκια» και «Pizza». Το προφίλ της πόλης φαίνεται πιο ισορροπημένο, με αρκετές κατηγορίες να εμφανίζουν παρόμοιες εντάσεις, γεγονός που υποδηλώνει διαφοροποιημένη ζήτηση και σταθερότητα σε πολλούς τύπους εστίασης. Σε μεσαίου μεγέθους πόλεις, όπως ο Πειραιάς, η Πάτρα, το Ηράκλειο και η Λάρισα, παρατηρούνται χαρακτηριστικά “τοπικά” μοτίβα. Ορισμένες κατηγορίες παρουσιάζουν υψηλότερη πυκνότητα σε συγκεκριμένες περιοχές, όπως τα «Μαγειρευτά» ή τα «Ζαχαροπλαστεία», κάτι που μπορεί να αντανακλά τοπικές γαστρονομικές προτιμήσεις ή διαφορές στην τοπική αγορά εστίασης.

Η κατανομή αυτή αναδεικνύει ότι ο συνδυασμός πόλης και κατηγορίας κουζίνας δεν είναι τυχαίος, αλλά διαμορφώνεται από παράγοντες όπως η τοπική ζήτηση, το είδος του πελατολογίου, ο πληθυσμός και ο ανταγωνισμός της κάθε περιοχής. Τα ευρήματα αυτά αποτελούν σημαντικό υπόβαθρο για τα επόμενα στάδια της εργασίας.

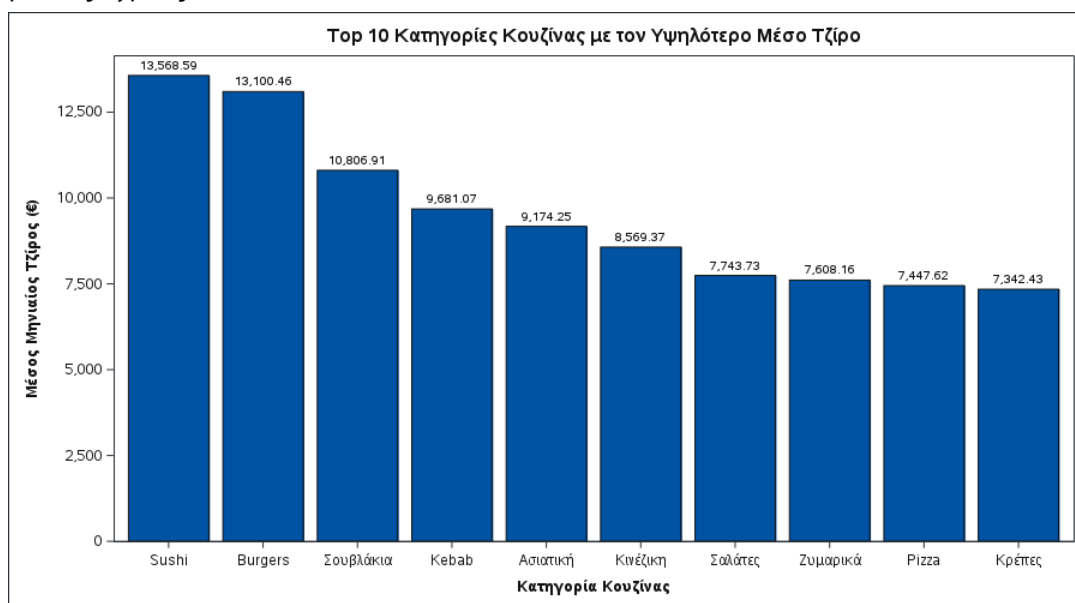
Οικονομική απόδοση καταστημάτων

Η οικονομική απόδοση των καταστημάτων αποτελεί βασικό δείκτη της δραστηριότητάς τους στην πλατφόρμα, καθώς συνδέεται άμεσα με τη ζήτηση, τη μέση αξία παραγγελίας και τη λειτουργική αποδοτικότητα σε κάθε περιοχή. Ο μέσος μηνιαίος τζίρος (GMV) χρησιμοποιείται ως κύριος δείκτης σύγκρισης και επιτρέπει την αξιολόγηση διαφορών τόσο ανάμεσα στις πόλεις όσο και μεταξύ των κατηγοριών κουζίνας.



Εικόνα 5 – Μέσος Μηνιαίος Τζίρος ανά Πόλη

Η ανάλυση ανά πόλη δείχνει έντονες διαφοροποιήσεις, με τη Δροσιά να εμφανίζεται στην πρώτη θέση με μέσο μηνιαίο τζίρο περίπου €9.908. Η υψηλή αυτή τιμή σχετίζεται όχι μόνο με το εισοδηματικό προφίλ της περιοχής, αλλά και με το γεγονός ότι στη Δροσιά λειτουργεί μικρός αριθμός καταστημάτων, τα οποία όμως έχουν υψηλή μέση αξία παραγγελίας. Επειδή ο δείκτης εκφράζει μέσο τζίρο ανά κατάστημα, περιοχές με λίγα αλλά premium καταστήματα εμφανίζουν συχνά σημαντικά αυξημένους μέσους όρους. Στην Αθήνα (€5.812) και στο Ηράκλειο Κρήτης (€5.803) οι τιμές είναι χαμηλότερες, κυρίως λόγω της πολύ μεγαλύτερης επιχειρηματικής πυκνότητας, ενώ άλλες πόλεις, όπως η Ροδόπολη, ο Ωρωπός και η Ορεστιάδα, παρουσιάζουν σταθερούς αλλά διαφοροποιημένους μέσους τζίρους.



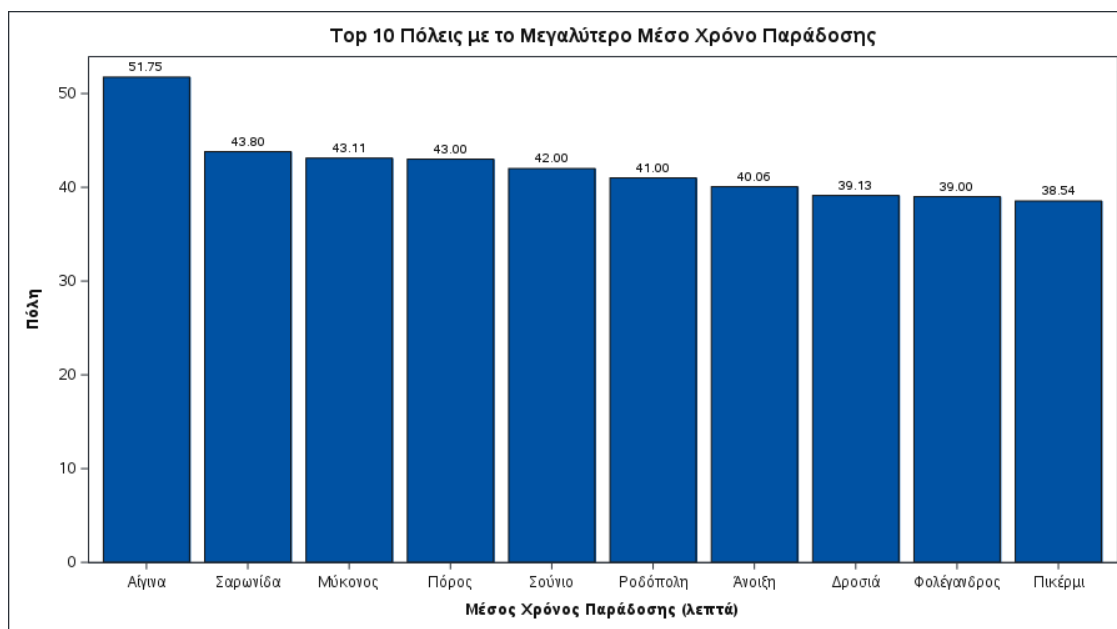
Εικόνα 6 – Μέσος Μηνιαίος Τζίρος ανά Κατηγορία Κουζίνας

Σε επίπεδο κατηγορίας κουζίνας, οι υψηλότερες τιμές παρατηρούνται στο Sushi (€13.568) και στα Burgers (€13.100), κατηγορίες που συνδέονται με premium προϊόντα και υψηλότερη τιμή ανά παραγγελία. Ακολουθούν τα Σουβλάκια, το Kebab και η Ασιατική κουζίνα, οι οποίες συνδυάζουν δημοτικότητα με αυξημένη μέση αξία παραγγελίας. Αντίθετα, κατηγορίες όπως οι Κρέπες και η Pizza παρουσιάζουν χαμηλότερο μέσο τζίρο ανά κατάστημα, παρότι έχουν υψηλό αριθμό συνολικών παραγγελιών, γεγονός που υποδηλώνει ότι ο όγκος δεν οδηγεί πάντα σε υψηλή μέση οικονομική απόδοση.

Συνολικά, η ανάλυση δείχνει ότι ο μέσος μηνιαίος τζίρος επηρεάζεται από δύο παράγοντες: την αγοραστική δύναμη της περιοχής και τον αριθμό των καταστημάτων που δραστηριοποιούνται σε αυτή. Περιοχές με λίγα, αλλά υψηλής αξίας καταστήματα εμφανίζουν πολύ υψηλότερο μέσο GMV, ενώ πιο κορεσμένες αγορές έχουν χαμηλότερους μέσους όρους παρά τον αυξημένο συνολικό όγκο παραγγελιών. Τα ευρήματα αυτά αποτελούν κρίσιμο υπόβαθρο για την ανάλυση συσταδοποίησης που ακολουθεί.

Χρόνος παράδοσης καταστημάτων

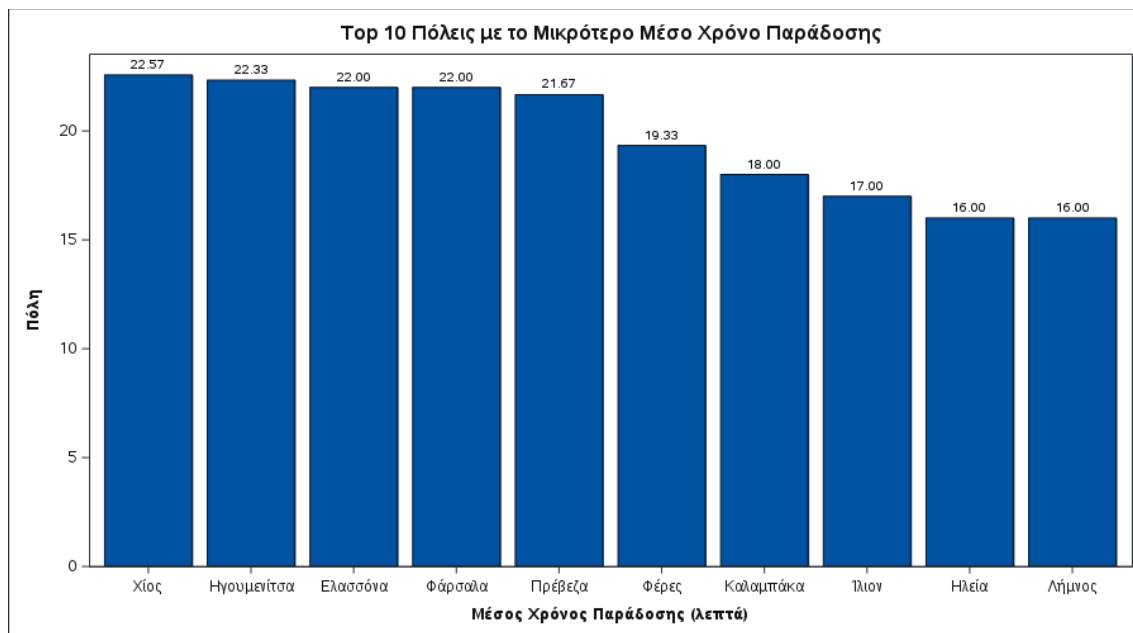
Ο χρόνος παράδοσης αποτελεί κρίσιμο επιχειρησιακό δείκτη, καθώς επηρεάζει άμεσα την εμπειρία των χρηστών, την πιθανότητα επαναληπτικών παραγγελιών και συνολικά τη λειτουργική αποτελεσματικότητα των συνεργαζόμενων καταστημάτων. Η ανάλυση των μέσων χρόνων παράδοσης ανά πόλη επιτρέπει την κατανόηση των διαφορών στην ταχύτητα εξυπηρέτησης και τον εντοπισμό περιοχών όπου ενδέχεται να υπάρχουν καθυστερήσεις ή ιδιαίτερες επιχειρησιακές απαιτήσεις.



Εικόνα 7 – Πόλεις με το Μεγαλύτερο Μέσο Χρόνο Παράδοσης

Στις περιοχές με το μεγαλύτερο μέσο χρόνο παράδοσης παρατηρούνται τιμές που κυμαίνονται από περίπου 38 έως 52 λεπτά. Πρώτη εμφανίζεται η Αίγινα (51.75 λεπτά), γεγονός που πιθανότατα συνδέεται με γεωγραφικούς περιορισμούς, αυξημένες αποστάσεις εντός του νησιού και ενδεχομένως μειωμένη διαθεσιμότητα διανομών. Ακολουθούν η Σαρωνίδα, η Μύκονος, ο Πόρος και το Σούνιο, περιοχές στις οποίες η εποχικότητα, η τουριστική δραστηριότητα ή η αραιή χωροταξική δομή μπορεί να αυξάνουν τον χρόνο παράδοσης. Οι τιμές αυτές υποδηλώνουν ότι ορισμένες τοποθεσίες

αντιμετωπίζουν εγγενείς λειτουργικές δυσκολίες που επηρεάζουν αρνητικά την ταχύτητα εξυπηρέτησης.



Εικόνα 8 – Πόλεις με το Μικρότερο Μέσο Χρόνο Παράδοσης

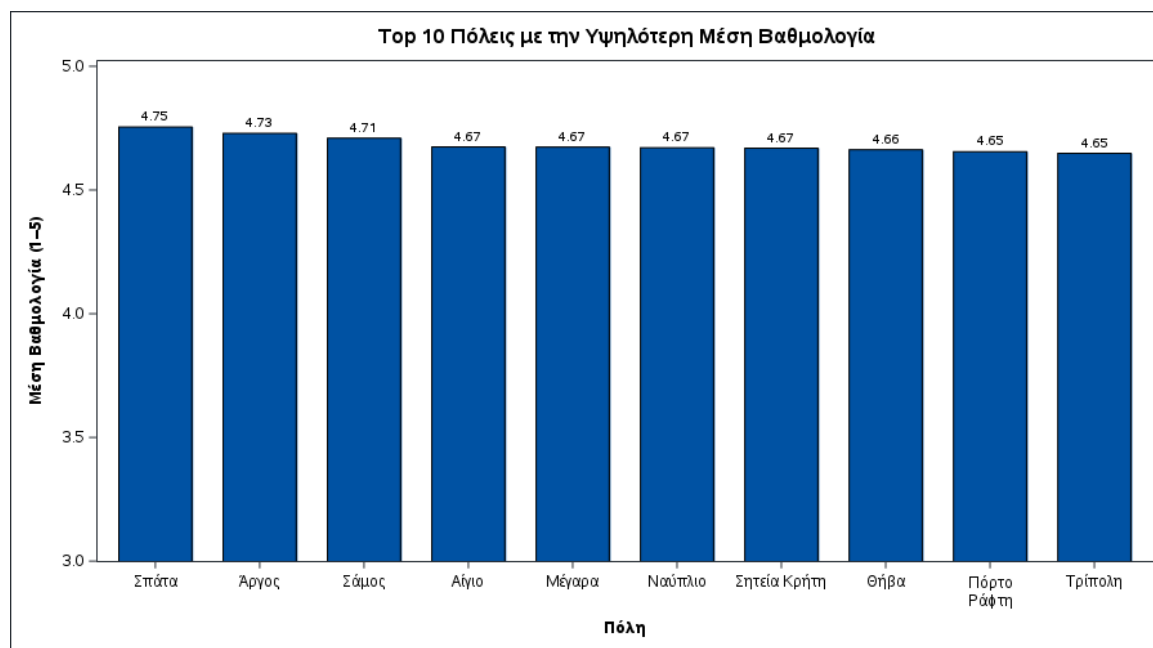
Οι περιοχές με το μικρότερο μέσο χρόνο παράδοσης εμφανίζουν ιδιαίτερα υψηλή λειτουργική αποτελεσματικότητα. Η Ηλεία και η Λήμνος, με χρόνο παράδοσης 16 λεπτά, αποτελούν τις πιο αποδοτικές περιοχές του δείγματος, γεγονός που σχετίζεται με το μικρότερο εύρος αποστάσεων και τη σταθερή ζήτηση. Ακολουθούν το Ίλιον, η Καλαμπάκα, οι Φέρες και η Πρέβεζα, όπου οι χρόνοι κυμαίνονται μεταξύ 17 και 22 λεπτών. Οι τιμές αυτές υποδηλώνουν ομαλό συντονισμό μεταξύ καταστημάτων και διανομέων και πιο προβλέψιμες ροές παραγγελιών, χαρακτηριστικά που συμβάλλουν στη συνεπή και γρήγορη εξυπηρέτηση.

Η σύγκριση των πόλεων με τους μεγαλύτερους και μικρότερους χρόνους παράδοσης δείχνει ότι η απόδοση επηρεάζεται σημαντικά από τη γεωγραφική διάταξη, την τουριστική δραστηριότητα και τον βαθμό αστικού συγκεντρωτισμού. Οι νησιωτικές και παραθαλάσσιες περιοχές εμφανίζουν αυξημένους χρόνους λόγω μεγαλύτερων αποστάσεων ή εποχικών διακυμάνσεων στη ζήτηση, ενώ οι πιο συμπαγείς ηπειρωτικές πόλεις καταγράφουν πολύ ταχύτερη και πιο σταθερή εξυπηρέτηση. Τα ευρήματα αυτά είναι ιδιαίτερα χρήσιμα για τα επόμενα στάδια της εργασίας, καθώς ο χρόνος παράδοσης ενδέχεται να σχετίζεται με την ποιότητα υπηρεσίας και την πιθανότητα διαφυγής ενός καταστήματος από την πλατφόρμα.

Αξιολόγηση καταστημάτων

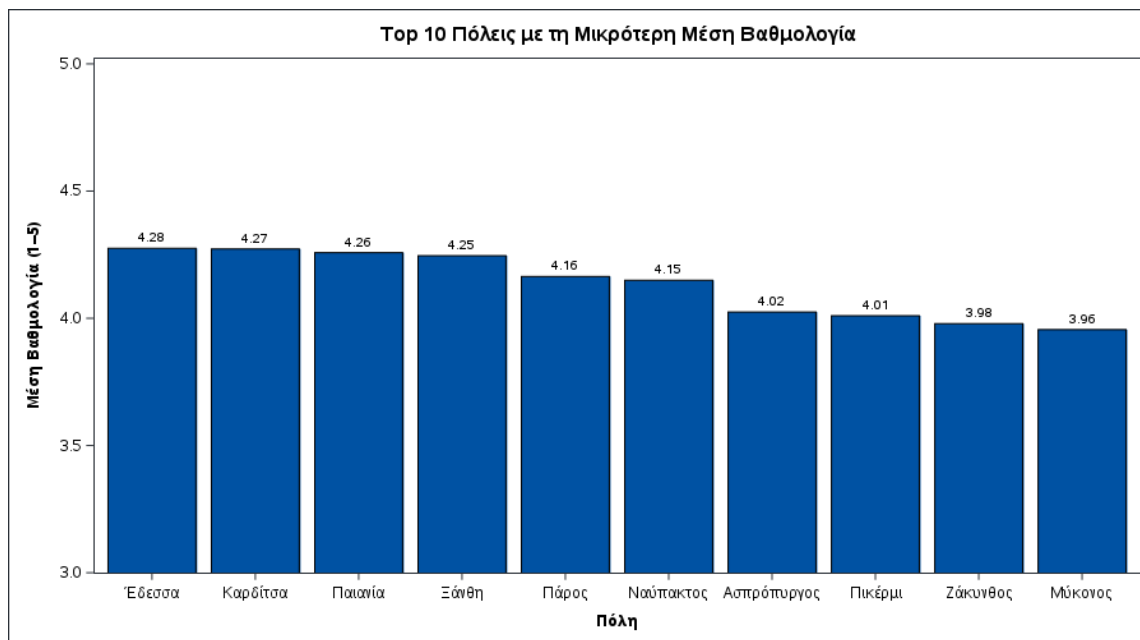
Η μέση βαθμολογία των καταστημάτων αποτελεί έναν από τους σημαντικότερους δείκτες ποιότητας και ικανοποίησης των χρηστών, καθώς αποτυπώνει τη συνολική εμπειρία που προσφέρει κάθε επιχείρηση σε όρους εξυπηρέτησης, αξιοπιστίας και ποιότητας των προϊόντων. Η ανάλυση των μέσων βαθμολογιών ανά πόλη επιτρέπει την αναγνώριση περιοχών που διακρίνονται για τη θετική εμπειρία πελατών, αλλά και εκείνων όπου παρατηρούνται χαμηλότερες επιδόσεις, πιθανώς λόγω επιχειρησιακών περιορισμών ή εποχικών επιβαρύνσεων. Η συσχέτιση της αξιολόγησης με τον χρόνο παράδοσης προσφέρει επιπλέον εικόνα για το πώς λειτουργικοί παράγοντες μπορεί να επηρεάζουν

τη συνολική ικανοποίηση, καθιστώντας τις βαθμολογίες κρίσιμο εργαλείο αξιολόγησης της απόδοσης των συνεργαζόμενων καταστημάτων.



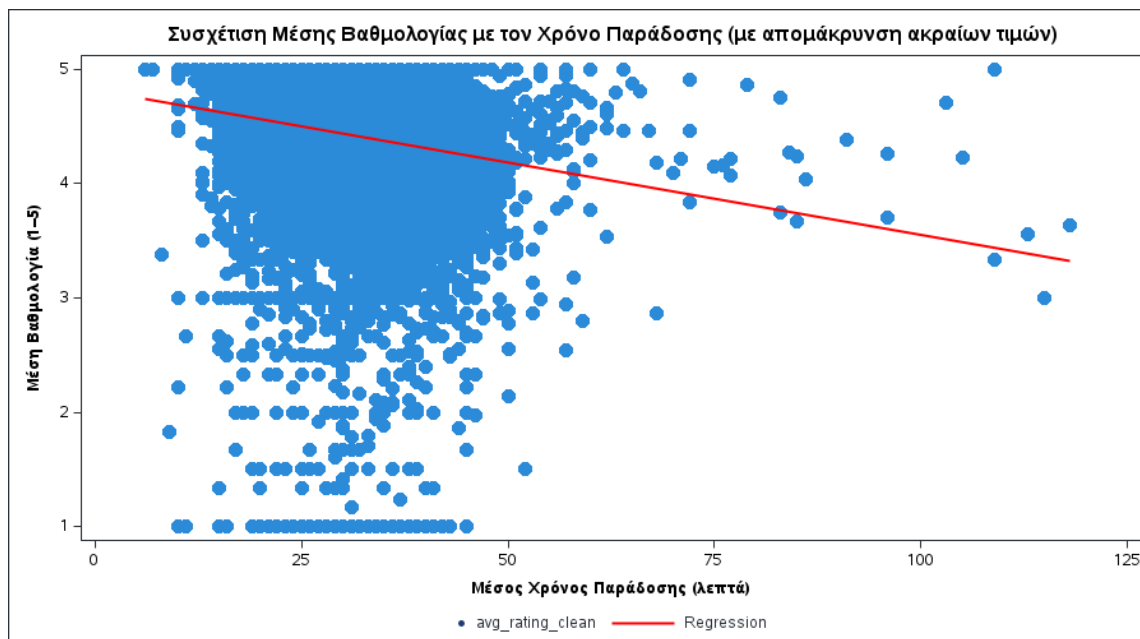
Εικόνα 9 – Top 10 Πόλεις με την Υψηλότερη Μέση Βαθμολογία

Οι περιοχές με τις υψηλότερες μέσες βαθμολογίες παρουσιάζουν ιδιαίτερα θετικές επιδόσεις, με τιμές που κυμαίνονται από 4.65 έως 4.75. Το εύρος αυτό αντικατοπτρίζει υψηλή και σταθερή ικανοποίηση των χρηστών, γεγονός που συνδέεται με ποιοτική εξυπηρέτηση, συνέπεια και ομαλή λειτουργία των καταστημάτων. Περιοχές όπως οι Σπέτσες, το Άργος και η Σάμος ξεχωρίζουν για τις πολύ υψηλές αξιολογήσεις τους, υποδηλώνοντας ευνοϊκές συνθήκες λειτουργίας και θετική φήμη στην τοπική αγορά.



Εικόνα 10 – Top 10 Πόλεις με τη Μικρότερη Μέση Βαθμολογία

Οι περιοχές με τις χαμηλότερες μέσες βαθμολογίες παρουσιάζουν τιμές μεταξύ 3.96 και 4.28, παραμένοντας ωστόσο σε σχετικά ικανοποιητικά επίπεδα. Η Μύκονος και η Ζάκυνθος καταγράφουν τις χαμηλότερες επιδόσεις, εύρημα που πιθανόν σχετίζεται με εποχική ζήτηση, αυξημένο φόρτο εργασίας ή ανομοιογενή ποιότητα εξυπηρέτησης.



Εικόνα 11 – Συσχέτιση Μέσης Βαθμολογίας με τον Χρόνο Παράδοσης

Η συσχέτιση μεταξύ μέσης βαθμολογίας και χρόνου παράδοσης εμφανίζεται ιδιαίτερα ασθενής, όπως υποδεικνύονται τόσο από τη μεγάλη διασπορά των σημείων όσο και από την ήττα αρνητική

κλίση της γραμμής παλινδρόμησης. Παρότι παρατηρείται μια τάση όπου μεγαλύτεροι χρόνοι παράδοσης συνδέονται με ελαφρώς χαμηλότερες αξιολογήσεις, η σχέση αυτή δεν είναι γραμμικά ισχυρή, γεγονός που φανερώνει ότι η τελική βαθμολογία των χρηστών διαμορφώνεται από ένα σύνολο παραγόντων πέρα από την ταχύτητα εξυπηρέτησης. Παράγοντες όπως η ποιότητα των προϊόντων, η ακρίβεια της παραγγελίας, η συνέπεια της εμπειρίας και ο συνολικός επαγγελματισμός του καταστήματος φαίνεται να επηρεάζουν πιο καθοριστικά την αξιολόγηση, εξηγώντας γιατί υψηλές βαθμολογίες εμφανίζονται ακόμη και σε περιπτώσεις με σχετικά αυξημένους χρόνους παράδοσης. Το γεγονός ότι η απομάκρυνση ακραίων τιμών δεν ενίσχυσε τη συσχέτιση ενισχύει την άποψη ότι η σχέση είναι εγγενώς ασθενής, επιβεβαιώνοντας τον πολυπαραγοντικό χαρακτήρα της ικανοποίησης των χρηστών.

6. Τμηματοποίηση Καταστημάτων (Shop Clustering)

6.1 Πειραματική Διαδικασία

Η παρούσα ενότητα περιγράφει τη διαδικασία που ακολουθήθηκε για την τμηματοποίηση των συνεργαζόμενων καταστημάτων της πλατφόρμας, με σκοπό τον εντοπισμό διακριτών και ερμηνεύσιμων ομάδων καταστημάτων με κοινά λειτουργικά και εμπορικά χαρακτηριστικά. Η ανάλυση πραγματοποιήθηκε σε επίπεδο καταστήματος (shop-level) και βασίστηκε σε συγκεντρωτικούς δείκτες που συνοψίζουν τη δραστηριότητα και την απόδοση κάθε καταστήματος στην πλατφόρμα (π.χ. μέσοι όροι και μηνιαία μεγέθη).

Στόχος της τμηματοποίησης είναι η ανάδειξη προφίλ καταστημάτων με διαφορετικά επίπεδα δραστηριότητας, ζήτησης και επιχειρησιακής επίδοσης, ώστε να υποστηριχθεί η λήψη στοχευμένων αποφάσεων και η διαμόρφωση διαφοροποιημένων στρατηγικών συνεργασίας και υποστήριξης ανά κατηγορία καταστημάτων.

Μονάδα ανάλυσης αποτελεί το κατάστημα, όπου κάθε παρατήρηση αντιστοιχεί σε ένα μοναδικό κατάστημα και περιγράφεται μέσω χαρακτηριστικών που αποτυπώνουν συνολικά τη συμπεριφορά του. Η επιλογή shop-level κρίθηκε καταλληλότερη από εναλλακτικές προσεγγίσεις (όπως ανάλυση σε επίπεδο παραγγελίας), καθώς επιτρέπει συγκρίσεις μεταξύ καταστημάτων με διαφορετικό όγκο δραστηριότητας και οδηγεί σε πιο σταθερές και επιχειρησιακά αξιοποιήσιμες συστάδες.

Η διαδικασία δεν περιορίστηκε στην απευθείας εφαρμογή ενός αλγορίθμου clustering, αλλά περιέλαβε διερευνητικά στάδια και πειραματισμό με εναλλακτικά σύνολα μεταβλητών, με στόχο την επίτευξη ποιοτικών και ερμηνεύσιμων αποτελεσμάτων. Ιδιαίτερη έμφαση δόθηκε στην τεκμηριωμένη επιλογή χαρακτηριστικών μέσω ανάλυσης συσχετίσεων (correlation analysis) και στην απομάκρυνση πλεοναστικών ή ακατάλληλων μεταβλητών που επηρέαζαν αρνητικά την ποιότητα της ομαδοποίησης. Παράλληλα, εφαρμόστηκαν βασικά βήματα προεπεξεργασίας, όπως χειρισμός ακραίων τιμών και τυποποίηση (standardization), τα οποία συνέβαλαν στη βελτίωση της ποιότητας των clusters και στη σταθερότητα των αποτελεσμάτων.

Με βάση τα παραπάνω, στη συνέχεια τεκμηριώνεται η επιλογή των χαρακτηριστικών που χρησιμοποιήθηκαν στην ομαδοποίηση, καθώς και τα βασικά βήματα προετοιμασίας των δεδομένων που προηγήθηκαν της τελικής εφαρμογής του clustering.

6.1.1 Υποψήφιες μεταβλητές και κριτήρια επιλογής

Η επιλογή των μεταβλητών αποτελεί ένα από τα πιο κρίσιμα στάδια στη διαδικασία τμηματοποίησης, καθώς καθορίζει τον τρόπο κατανομής των δεδομένων και την πρακτική ερμηνευσιμότητα των παραγόμενων clusters. Ειδικά σε μεθόδους που βασίζονται σε αποστάσεις, όπως το K-Means, η ποιότητα της τμηματοποίησης επηρεάζεται σημαντικά από το αν οι μεταβλητές αποτυπώνουν

διαφορετικές διαστάσεις συμπεριφοράς ή αν, αντίθετα, εκφράζουν το ίδιο υποκείμενο φαινόμενο (πλεονασμός πληροφορίας).

Στο πλαίσιο της παρούσας ανάλυσης, εξετάστηκε αρχικά ένα ευρύ σύνολο υποψηφίων μεταβλητών σε επίπεδο καταστήματος, οι οποίες προκύπτουν από συγκεντρωτικά στατιστικά δραστηριότητας και αποτυπώνουν κρίσιμες επιχειρησιακές πτυχές της λειτουργίας ενός καταστήματος στην πλατφόρμα. Ενδεικτικά, στο αρχικό σύνολο υποψηφίων χαρακτηριστικών συμπεριλήφθηκαν μεταβλητές που σχετίζονται με: (α) τον όγκο ζήτησης (π.χ. μέσος μηνιαίος αριθμός παραγγελιών), (β) την οικονομική απόδοση (π.χ. μέσο μηνιαίο GMV, μέση μηνιαία προμήθεια), (γ) τη λειτουργική αποτελεσματικότητα/ποιότητα εξυπηρέτησης (π.χ. μέσος χρόνος παράδοσης) και (δ) την “ωριμότητα” παρουσίας στην πλατφόρμα.

Τα κριτήρια επιλογής των μεταβλητών διαμορφώθηκαν με τρόπο ώστε η τελική τμηματοποίηση να είναι ταυτόχρονα:

- (i) επιχειρησιακά αξιοποιήσιμη, δηλαδή να οδηγεί σε segments που μπορούν να περιγραφούν με σαφήνεια και να συνδεθούν με συγκεκριμένες στρατηγικές, και
- (ii) μεθοδολογικά ορθή, δηλαδή να βασίζεται σε μεταβλητές που δεν είναι υπερβολικά συσχετισμένες μεταξύ τους και δεν “διπλομετρούν” την ίδια πληροφορία.

Συνεπώς, πέρα από την επιχειρησιακή σημασία των χαρακτηριστικών, δόθηκε ιδιαίτερη έμφαση στη στατιστική διερεύνηση της συσχέτισης μεταξύ των υποψηφίων μεταβλητών. Μεταβλητές με πολύ υψηλή συσχέτιση τείνουν να καθοδηγούν δυσανάλογα το clustering προς μία συγκεκριμένη διάσταση (π.χ. “μέγεθος” καταστήματος), με αποτέλεσμα οι παραγόμενες ομάδες να χάνουν σε ερμηνευσιμότητα και να μην αποτυπώνουν ουσιαστικές διαφορές σε άλλες επιχειρησιακές πτυχές (όπως η ταχύτητα εξυπηρέτησης ή η παλαιότητα συνεργασίας). Για τον λόγο αυτό, πριν οριστικοποιηθεί το τελικό σύνολο χαρακτηριστικών, πραγματοποιήθηκε correlation analysis και δοκιμάστηκαν εναλλακτικοί συνδυασμοί μεταβλητών, ώστε να επιλεγεί ένα σύνολο που παρέχει συμπληρωματική πληροφορία και οδηγεί σε πιο σταθερά και διακριτά clusters.

Στην επόμενη ενότητα (6.1.2) παρουσιάζονται τα αποτελέσματα της ανάλυσης συσχέτισης, καθώς και οι αποτυχημένες αρχικές απόπειρες τμηματοποίησης που ανέδειξαν την ανάγκη απόρριψης πλεοναστικών μεταβλητών. Η διαδικασία αυτή οδήγησε στη σταδιακή κατάληξη στις τελικές μεταβλητές που χρησιμοποιήθηκαν στο clustering, με στόχο την παραγωγή ομάδων καταστημάτων με σαφή επιχειρησιακή ταυτότητα.

6.1.2 Correlation analysis & απόρριψη “πλεοναστικών” μεταβλητών

Στο στάδιο της επιλογής μεταβλητών πραγματοποιήθηκαν διαδοχικές πειραματικές δοκιμές (trial-and-error), με στόχο να προκύψουν συστάδες που είναι ταυτόχρονα στατιστικά συνεκτικές και επιχειρησιακά ερμηνεύσιμες (actionable). Αρχικά, δοκιμάστηκε clustering με μεταβλητές που αποτυπώνουν τον «όγκο δραστηριότητας» ενός καταστήματος, ωστόσο τα αποτελέσματα δεν παρήγαγαν σαφή προφίλ με πρακτική αξιοποίηση. Για τον λόγο αυτό, πριν καταλήξουμε στην τελική τριάδα μεταβλητών, εφαρμόστηκε ανάλυση συσχέτισης Pearson ώστε να εντοπιστούν και να απορριφθούν μεταβλητές που μεταφέρουν πλεοναστική πληροφορία.

Αποτυχημένη αρχική δοκιμή (clustering με GMV & Orders)

Η πρώτη προσέγγιση στηρίχθηκε στις μεταβλητές avg_delivery_time_minutes, avg_shop_monthly_gmv και avg_shop_monthly_orders. Ο αλγόριθμος παρήγαγε 4 συστάδες, όμως το αποτέλεσμα δεν κρίθηκε επιχειρησιακά αξιοποιήσιμο: οι ομάδες που προέκυψαν λειτουργούσαν κυρίως ως «βαθμίδες» του ίδιου φαινομένου (όγκος/ζήτηση), χωρίς καθαρή διαφοροποίηση σε προφίλ καταστημάτων που να οδηγούν σε διαφορετικές στρατηγικές/ενέργειες.

Cluster_ID	Συχνότητα	Ποσοστό συχνότητας	avg_delivery_time_minutes	avg_shop_monthly_gmv	avg_shop_monthly_orders
1	3,915	26.06%	33.18	€627.37	60.49
2	5,782	38.49%	36.12	€4,569.04	466.60
3	2,360	15.71%	24.62	€128.98	15.01
4	2,966	19.74%	22.70	€2,277.23	324.75
Άθροισμα	15,023	100.00%	30.90	€2,391.87	261.82

Εικόνα 12 - Αρχικό Clustering με avg_delivery_time_minutes, avg_shop_monthly_gmv και avg_shop_monthly_orders (4 συστάδες)

Κρίσιμος λόγος ήταν ότι οι μεταβλητές GMV και Orders αποτυπώνουν σε μεγάλο βαθμό την ίδια διάσταση (όγκος πωλήσεων και αντίστοιχα πλήθος παραγγελιών), άρα η ταυτόχρονη χρήση τους «διπλομετρά» τον ίδιο άξονα στο clustering. Το γεγονός αυτό επιβεβαιώθηκε στη συνέχεια με Pearson correlation, όπου παρατηρήθηκε πολύ υψηλή συσχέτιση μεταξύ avg_shop_monthly_gmv και avg_shop_monthly_orders ($r \approx 0.94$). Ως αποτέλεσμα, το συγκεκριμένο σενάριο απορρίφθηκε και ακολούθησε συστηματικό screening συσχετίσεων.

Code

Log

Results

Edit Code

1 With Variables:

avg_shop_monthly_orders

1 Variables:

avg_shop_monthly_gmv

Pearson Correlation Coefficients, N = 17238

avg_shop_monthly_gmv

avg_shop_monthly_orders

0.94453

Εικόνα 13 - Pearson Correlation μεταξύ avg_shop_monthly_orders και avg_shop_monthly_gmv

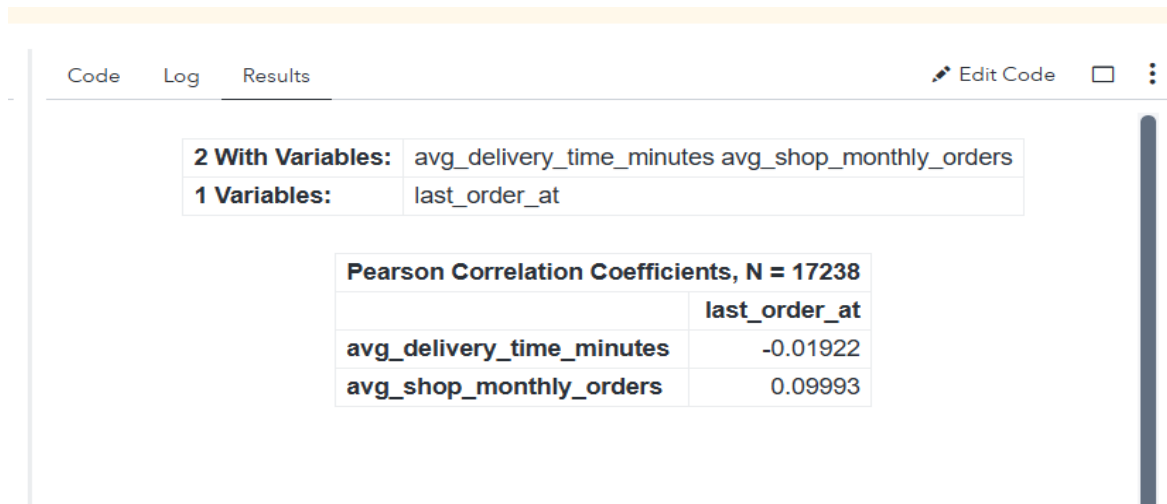
Συστηματικός έλεγχος συσχέτισης (Pearson) και “αποτυχημένες” δοκιμές μεταβλητών

Μετά την αρχική απόρριψη του σεναρίου GMV + Orders λόγω πολύ υψηλής συσχέτισης, ακολούθησε συστηματικός έλεγχος υποψήφιων μεταβλητών με Pearson correlation coefficient. Στόχος ήταν να εντοπιστούν μεταβλητές που (α) προσθέτουν ανεξάρτητη πληροφορία στο clustering και (β) δεν αναπαριστούν την ίδια επιχειρησιακή διάσταση (ώστε να μην “μετριέται διπλά” ο ίδιος παράγοντας). Ως πρακτικός κανόνας, ζεύγη μεταβλητών με $|r| \geq 0.80-0.90$ θεωρήθηκαν υπερβολικά συσχετισμένα και απορρίφθηκαν σε κοινή χρήση.

Δοκιμή A: Έλεγχος της μεταβλητής last_order_at (αναμενόμενη διάσταση “recency”)

Αρχικά εξετάστηκε η μεταβλητή last_order_at ως πιθανός δείκτης “πρόσφατης δραστηριότητας” (recency), σε συνδυασμό με avg_shop_monthly_orders και avg_delivery_time_minutes. Τα αποτελέσματα Pearson έδειξαν πολύ χαμηλή γραμμική συσχέτιση με τις βασικές μεταβλητές δραστηριότητας ($r \approx 0.10$ με orders και $r \approx -0.02$ με delivery time). Αυτό υποδηλώνει ότι, στην παρούσα μορφή της, η last_order_at δεν λειτουργεί ως καθαρή συνεχής κλίμακα συμπεριφοράς για clustering.

Επιπλέον, επιχειρησιακά η “ημερομηνία τελευταίας παραγγελίας” είναι πιο χρήσιμη όταν μετασχηματίζεται σε recency σε ημέρες (π.χ. days_since_last_order ως διαφορά από μια κοινή ημερομηνία αναφοράς). Η χρήση της “ως έχει” (timestamp/ημερομηνία) μπορεί να οδηγήσει σε τεχνητές ομαδοποιήσεις που σχετίζονται με το ημερολογιακό πλαίσιο και όχι απαραίτητα με το πραγματικό πρόφιλ λειτουργίας. Για τον λόγο αυτό η μεταβλητή δεν επιλέχθηκε στη συγκεκριμένη φάση, και προτιμήθηκαν πιο σταθερές/ερμηνεύσιμες μεταβλητές.



Εικόνα 14 – Pearson Correlation των avg_delivery_time_minutes και avg_shop_monthly_orders με τη last_order_at

Δοκιμή B: Έλεγχος της μεταβλητής avg_shop_monthly_users (πλεονασμός με “όγκο”)

Στη συνέχεια δοκιμάστηκε η μεταβλητή avg_shop_monthly_users, με στόχο να αποτυπωθεί η “βάση πελατών/χρηστών” ενός καταστήματος. Ωστόσο, τα αποτελέσματα έδειξαν σχεδόν τέλεια συσχέτιση μεταξύ avg_shop_monthly_users και avg_shop_monthly_orders ($r \approx 0.99999$), δηλαδή πρακτικά οι δύο μεταβλητές μεταφέρουν την ίδια πληροφορία (όγκος δραστηριότητας). Αυτό σημαίνει ότι αν συνυπάρξουν στο clustering, ο αλγόριθμος θα “βλέπει” δύο φορές την ίδια διάσταση και θα ευνοεί υπερβολικά τις διαφορές όγκου έναντι άλλων χαρακτηριστικών. Για να αποφευχθεί αυτός ο πλεονασμός, η avg_shop_monthly_users απορρίφθηκε και κρατήθηκε ως βασικός δείκτης δραστηριότητας η avg_shop_monthly_orders, η οποία είναι πιο άμεσα ερμηνεύσιμη επιχειρησιακά (μετρά καθαρά τη ζήτηση/παραγγελιοληψία).

Code	Log	Results	Edit Code		
3 With Variables: avg_delivery_time_minutes avg_shop_monthly_orders days_on_platform					
1 Variables: avg_shop_monthly_users					
Pearson Correlation Coefficients, N = 17238					
	avg_shop_monthly_users				
avg_delivery_time_minutes	0.07903				
avg_shop_monthly_orders	0.99999				
days_on_platform	0.20793				

Εικόνα 15 – Pearson Correlation της avg_shop_monthly_users σε Σχέση με avg_shop_monthly_orders, avg_delivery_time_minutes και days_on_platform

Τελική επιλογή μεταβλητών μετά το correlation analysis

Μετά τις αποτυχημένες/απορριφθείσες δοκιμές μεταβλητών (συνδυασμοί με GMV–Orders, καθώς και επιλογές που αποδείχθηκαν πλεοναστικές όπως Users–Orders, αλλά και η χρήση του last_order_at στην αρχική του μορφή), η διαδικασία επιλογής χαρακτηριστικών οργανώθηκε σε correlation screening με τον συντελεστή Pearson. Στόχος ήταν να διαμορφωθεί ένα τελικό σύνολο μεταβλητών που: (α) δεν παρουσιάζει ισχυρή γραμμική συσχέτιση μεταξύ των μεταβλητών (ώστε να αποφεύγεται η “διπλή μέτρηση” της ίδιας επιχειρησιακής διάστασης στις αποστάσεις του clustering), και (β) καλύπτει διαφορετικές και συμπληρωματικές πτυχές της λειτουργίας ενός καταστήματος, ώστε οι συστάδες που θα προκύψουν να είναι ερμηνεύσιμες και επιχειρησιακά αξιοποιήσιμες.

Βάσει των παραπάνω κριτηρίων, επιλέχθηκαν οι ακόλουθες τρεις τελικές μεταβλητές για το shop clustering:

- avg_shop_monthly_orders: δείκτης ζήτησης/όγκου δραστηριότητας, δηλαδή πόσο “κινείται” το κατάστημα σε μηνιαία βάση. Επιλέχθηκε ως ο βασικός άξονας μεγέθους/δραστηριότητας, καθώς είναι άμεσα ερμηνεύσιμος επιχειρησιακά.
- avg_delivery_time_minutes: δείκτης λειτουργικής απόδοσης/ταχύτητας εξυπηρέτησης. Αποτυπώνει την επιχειρησιακή αποτελεσματικότητα και είναι κρίσιμος, επειδή διαφοροποιεί καταστήματα που μπορεί να έχουν παρόμοιο όγκο αλλά διαφορετική εμπειρία εξυπηρέτησης.
- days_on_platform: δείκτης ωριμότητας/παρουσίας στην πλατφόρμα. Συνδέεται με τη “φάση ζωής” ενός καταστήματος (νέο/αναπτυσσόμενο/καθιερωμένο) και μπορεί να σχετίζεται με την εμπειρία συνεργασίας, τη σταθερότητα λειτουργίας και τον βαθμό ενσωμάτωσης σε διαδικασίες της πλατφόρμας.

Έλεγχος συσχέτισης και αιτιολόγηση της τελικής επιλογής

Ο έλεγχος Pearson επιβεβαίωσε ότι οι τρεις μεταβλητές δεν είναι έντονα συσχετισμένες και επομένως λειτουργούν ως διακριτές διαστάσεις στο clustering. Συγκεκριμένα, η συσχέτιση μεταξύ avg_shop_monthly_orders και avg_delivery_time_minutes είναι εξαιρετικά χαμηλή ($r \approx 0.078$), υποδηλώνοντας ότι ο όγκος παραγγελιών και ο χρόνος παράδοσης αποτελούν σχεδόν ανεξάρτητες μεταβλητές. Αυτό είναι ιδιαίτερα σημαντικό, διότι επιτρέπει στον αλγόριθμο να σχηματίσει συστάδες που διαφοροποιούνται όχι μόνο ως προς το “μέγεθος” ενός καταστήματος, αλλά και ως προς τη λειτουργική του συμπεριφορά. Με άλλα λόγια, ένα κατάστημα μπορεί να είναι υψηλής ζήτησης χωρίς απαραίτητα να είναι αργό στην παράδοση, ή να έχει χαμηλότερη ζήτηση αλλά να εξυπηρετεί ταχύτερα· άρα, η διάσταση της απόδοσης δεν “καταρρέει” πάνω στη διάσταση του όγκου.

CodeLogResultsEdit Code

1 With Variables:	avg_shop_monthly_orders
1 Variables:	avg_delivery_time_minutes

Pearson Correlation Coefficients, N = 17238	
	avg_delivery_time_minutes
avg_shop_monthly_orders	0.07846

Εικόνα 16 - Pearson Correlation μεταξύ avg_shop_monthly_orders και avg_delivery_time_minutes

Παράλληλα, η μεταβλητή `days_on_platform` παρουσίασε χαμηλή έως μέτρια συσχέτιση με τις άλλες δύο ($r \approx 0.21$ με `avg_shop_monthly_orders` και $r \approx 0.17$ με `avg_delivery_time_minutes`). Το αποτέλεσμα αυτό είναι επίσης επιθυμητό καθώς δείχνει ότι η “ωριμότητα/παισιότητα” ενός καταστήματος στην πλατφόρμα δεν ταυτίζεται ούτε με τον όγκο παραγγελιών ούτε με τον χρόνο παράδοσης, άρα προσθέτει ξεχωριστή και συμπληρωματική πληροφορία. Έτσι, το clustering μπορεί να εντοπίσει προφίλ καταστημάτων που διαφέρουν όχι μόνο σε επίπεδο δραστηριότητας και απόδοσης, αλλά και σε επίπεδο “tenure” (π.χ. νέα καταστήματα με χαμηλή δραστηριότητα, ώριμα καταστήματα με υψηλή δραστηριότητα, κ.λπ.), κάτι που αυξάνει την επιχειρησιακή αξία της τμηματοποίησης.

Code Log Results  Edit Code 

2 With Variables:	avg_delivery_time_minutes avg_shop_monthly_orders
1 Variables:	days_on_platform

Pearson Correlation Coefficients, N = 17238	
	days_on_platform
avg_delivery_time_minutes	0.16530
avg_shop_monthly_orders	0.20760

Εικόνα 17 – Pearson Correlation της days_on_platform σε Σχέση με avg_shop_monthly_orders και avg_delivery_time_minutes

Συνολικά, το τελικό τρίπτυχο μεταβλητών καλύπτει τρεις διαφορετικές επιχειρησιακές ερωτήσεις (“πόσο”, “πώς”, “πόσο καιρό”) και παρέχει ένα ισορροπημένο σύνολο εισόδου για clustering:

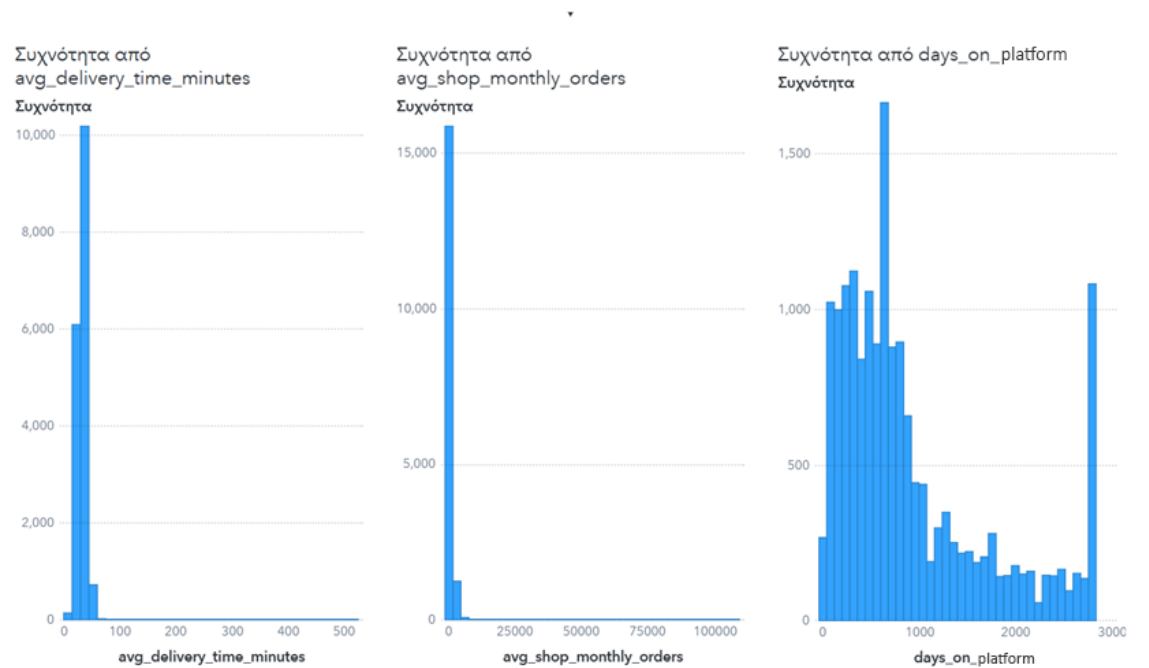
- `avg_shop_monthly_orders` → “πόσο” (ένταση ζήτησης/volume)
- `avg_delivery_time_minutes` → “πώς” (λειτουργική απόδοση/εξυπηρέτηση)
- `days_on_platform` → “πόσο καιρό” (ωριμότητα/στάδιο παρουσίας)

Με βάση τα παραπάνω, η επιλογή των συγκεκριμένων μεταβλητών θεωρήθηκε κατάλληλη για την τελική εφαρμογή του shop clustering, καθώς αποφεύγει πλεονασμό πληροφορίας και αυξάνει την πιθανότητα παραγωγής συστάδων με σαφή επιχειρησιακή ερμηνεία.

6.1.3 Καθαρισμός δεδομένων & χειρισμός ακραίων τιμών (outliers)

Πριν εφαρμοστεί ο αλγόριθμος clustering, πραγματοποιήθηκε στοχευμένος έλεγχος ποιότητας και κατανομής των τριών τελικών μεταβλητών. Το βήμα αυτό κρίθηκε απαραίτητο, επειδή οι ακραίες τιμές μπορούν να επηρεάσουν δυσανάλογα τις αποστάσεις μεταξύ παρατηρήσεων και, κατ' επέκταση, τους υπολογισμούς των κέντρων των συστάδων (centroids). Στην πράξη, λίγα “ακραία” καταστήματα με υπερβολικά μεγάλους χρόνους, παραγγελίες ή πολύ μεγάλη παλαιότητα στην πλατφόρμα είναι δυνατόν να “τραβήξουν” τον αλγόριθμο, οδηγώντας σε συστάδες που δεν αντανakλούν το κυρίαρχο επιχειρησιακό μοτίβο της πλειονότητας των καταστημάτων.

Η διερεύνηση των κατανομών με histograms και boxplots ανέδειξε ότι οι μεταβλητές εμφανίζουν έντονη δεξιά ουρά (right-skewness), δηλαδή συγκέντρωση τιμών σε χαμηλά–μεσαία επίπεδα και ταυτόχρονα παρουσία λίγων παρατηρήσεων με πολύ υψηλές τιμές. Το φαινόμενο αυτό είναι αναμενόμενο σε επιχειρησιακά δεδομένα: π.χ. λίγα καταστήματα μπορεί να έχουν εξαιρετικά υψηλό όγκο παραγγελιών, ενώ σε ειδικές περιπτώσεις μπορεί να παρατηρούνται και πολύ μεγάλοι χρόνοι παράδοσης (λόγω τοποθεσίας, λειτουργικού μοντέλου ή εποχικών αιχμών). Ωστόσο, για σκοπούς τμηματοποίησης (και όχι ανίχνευσης ανωμαλιών), στόχος είναι οι συστάδες να περιγράφουν “σταθερά” και επαναλαμβανόμενα προφίλ καταστημάτων, όχι να καθορίζονται από σπάνιες/ακραίες συμπεριφορές. Οι κατανομές των μεταβλητών παρουσιάζονται στην Εικόνα 18.

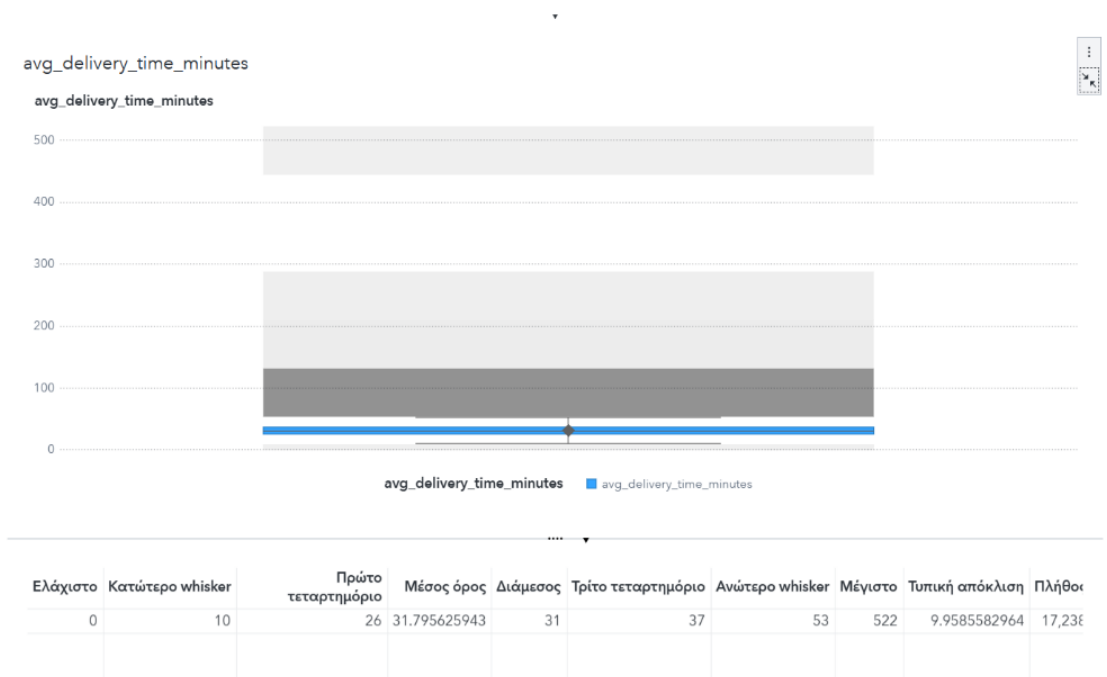


Εικόνα 18 - Histograms των Τελικών Μεταβλητών

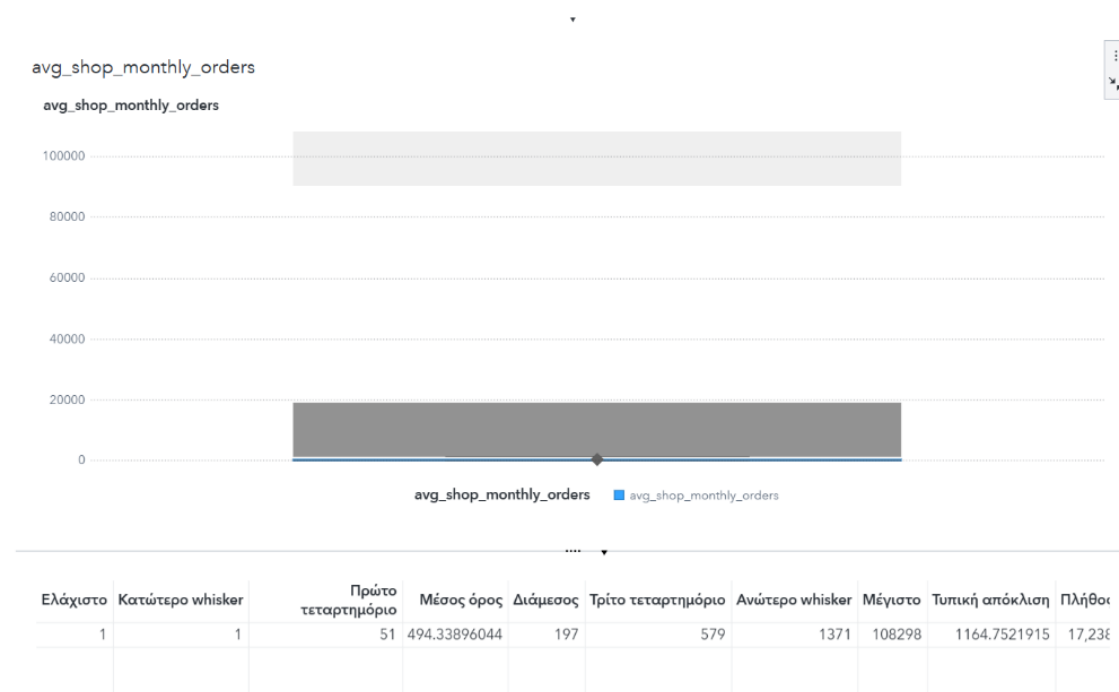
Για τον περιορισμό αυτής της επίδρασης, εφαρμόστηκε στο SAS Model Studio χειρισμός ακραίων τιμών μέσω upper limits. Η επιλογή αυτή επιτρέπει να διατηρηθούν οι παρατηρήσεις στο dataset (ώστε να μην χάνεται πληροφορία ή να μην αλλοιώνεται η αντιπροσωπευτικότητα του δείγματος), αλλά ταυτόχρονα αποτρέπει το να κυριαρχούν οι πολύ μεγάλες τιμές στη διαμόρφωση των συστάδων. Συγκεκριμένα, ορίστηκαν τα εξής ανώτατα όρια στις τρεις μεταβλητές:

- $\text{avg_delivery_time_minutes} \leq 53$
- $\text{avg_shop_monthly_orders} \leq 1371$
- $\text{days_on_platform} \leq 2683$

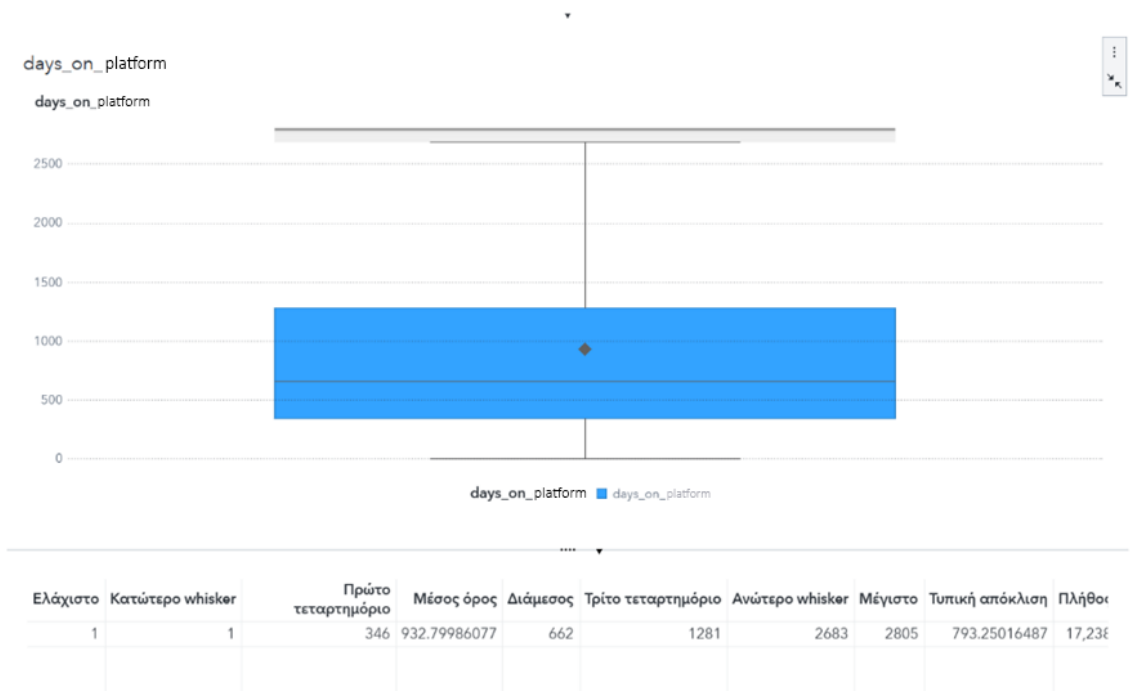
Η επιλογή των ανώτατων ορίων τεκμηριώνεται οπτικά στα boxplots των Εικόνων 19,20,21.



Εικόνα 19 - Boxplot της Μεταβλητής avg_delivery_time_minutes



Εικόνα 20 - Boxplot της Μεταβλητής avg_shop_monthly_orders



Εικόνα 21 - Boxplot της Μεταβλητής days_on_platform

Με την εφαρμογή των παραπάνω ορίων, οι παρατηρήσεις που ξεπερνούν τα συγκεκριμένα thresholds δεν επιτρέπεται να επηρεάζουν δυσανάλογα τις αποστάσεις και τους centroid υπολογισμούς. Έτσι, το clustering “διαβάζει” καλύτερα τη δομή του κύριου πληθυσμού και οδηγεί σε ομάδες που είναι πιο συνεκτικές και επιχειρησιακά ερμηνεύσιμες (π.χ. καταστήματα με παρόμοια λειτουργική συμπεριφορά και επίπεδο ζήτησης). Παράλληλα, επειδή οι τρεις μεταβλητές βρίσκονται σε διαφορετικές μονάδες και κλίμακες (λεπτά, πλήθος παραγγελιών, ημέρες), εφαρμόστηκε standardization πριν την εκπαίδευση του μοντέλου. Αυτό είναι κρίσιμο ειδικά σε distance-based μεθόδους όπως το K-Means, καθώς χωρίς κανονικοποίηση μία μεταβλητή με μεγαλύτερη αριθμητική κλίμακα (π.χ. orders) θα μπορούσε να “υπερκαθορίσει” τις αποστάσεις και να μειώσει τον ρόλο των υπόλοιπων μεταβλητών. Μετά τον χειρισμό outliers και την τυποποίηση, το dataset θεωρήθηκε κατάλληλο για την τελική διαδικασία clustering και για αξιόπιστο profiling των clusters.

6.2 Αποτελέσματα Τμηματοποίησης Καταστημάτων

Στην παρούσα ενότητα παρουσιάζεται η τελική τμηματοποίηση των συνεργαζόμενων καταστημάτων και τα βασικά αποτελέσματα που προέκυψαν από την εφαρμογή του μοντέλου στο SAS Model Studio. Το σημείο εκκίνησης για την τελική λύση ήταν η προηγούμενη πειραματική διαδικασία, όπου τεκμηριώθηκε η επιλογή των καταλληλότερων μεταβλητών και εφαρμόστηκαν τα απαραίτητα βήματα προεπεξεργασίας, ώστε τα δεδομένα να είναι κατάλληλα για distance-based ομαδοποίηση. Με βάση αυτό το “καθαρό” και ελεγχόμενο σύνολο εισόδου, η τμηματοποίηση σχεδιάστηκε με στόχο να παραχθούν ομάδες καταστημάτων που να είναι όχι μόνο στατιστικά συνεκτικές, αλλά και ερμηνεύσιμες στο επιχειρησιακό πλαίσιο της πλατφόρμας.

Η παρουσίαση των αποτελεσμάτων ακολουθεί μια λογική ροή: αρχικά περιγράφεται συνοπτικά ο τρόπος υλοποίησης της διαδικασίας (pipeline και βασικές ρυθμίσεις), στη συνέχεια τεκμηριώνεται η επιλογή του αριθμού συστάδων μέσω του κριτηρίου που παρέχει το εργαλείο, και ακολούθως παρουσιάζεται η κατανομή των clusters και η συγκριτική τους απεικόνιση. Έπειτα, πραγματοποιείται profiling των ομάδων με χρήση πινάκων cross-tab, ώστε να αποτυπωθούν οι διαφορές τους στους κύριους άξονες που περιγράφουν το προφίλ ενός καταστήματος. Τέλος, τα clusters αποκτούν επιχειρησιακή ερμηνεία μέσω ονοματοδοσίας και σχολιασμού και συνδέονται με πρακτικές

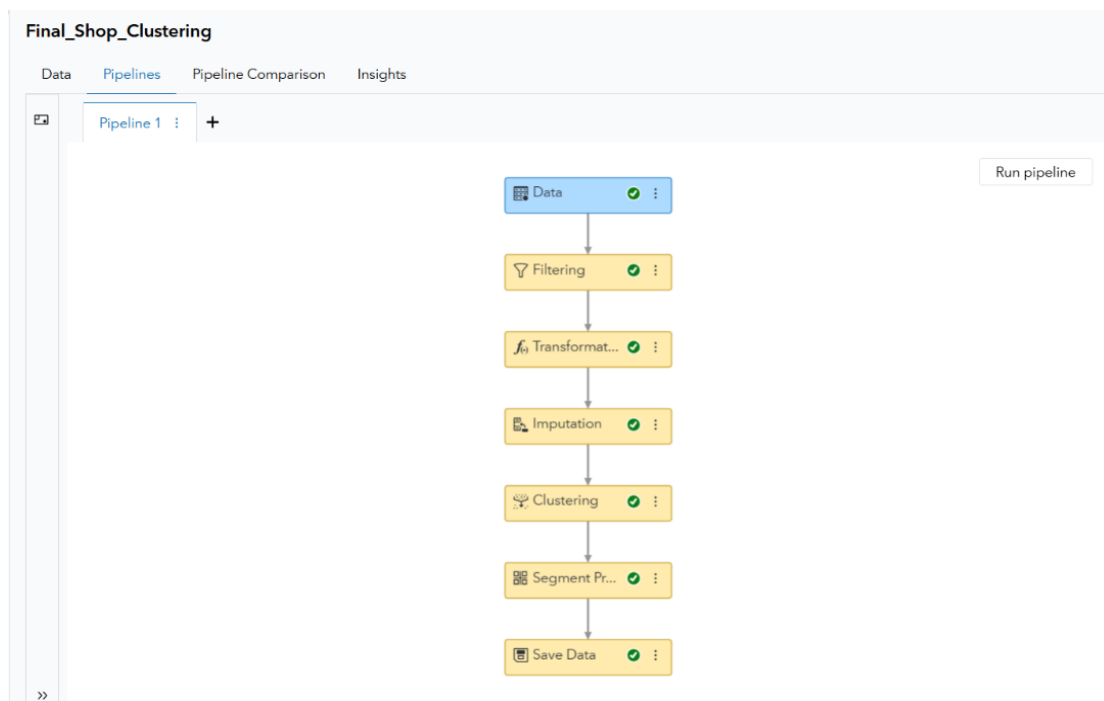
προτάσεις, δείχνοντας πώς η τμηματοποίηση μπορεί να αξιοποιηθεί ως εργαλείο υποστήριξης αποφάσεων και στοχευμένων ενεργειών ανά κατηγορία καταστημάτων.

6.2.1 Διαδικασία υλοποίησης στο SAS Model Studio

Η τελική τμηματοποίηση των καταστημάτων υλοποιήθηκε στο περιβάλλον SAS Model Studio μέσω ενός ολοκληρωμένου pipeline, ώστε η διαδικασία να είναι πλήρως αναπαραγώγιμη και να αποτυπώνει καθαρά όλα τα βήματα από την είσοδο των δεδομένων έως την παραγωγή του τελικού segmentation. Το pipeline που ακολουθήθηκε παρουσιάζεται στην Εικόνα 22 και οργανώνεται σε διακριτικούς κόμβους (nodes), οι οποίοι καλύπτουν την προετοιμασία των δεδομένων, την εκπαίδευση του clustering αλγόριθμου και την εξαγωγή των αποτελεσμάτων.

Συγκεκριμένα, το pipeline ξεκινά από τον κόμβο Data, όπου ορίζεται το τελικό shop-level dataset που προέκυψε από τη φάση επιλογής μεταβλητών και προεπεξεργασίας. Στη συνέχεια, μέσω του Filtering εφαρμόζονται οι απαραίτητοι έλεγχοι/περιορισμοί (ώστε η είσοδος του clustering να περιλαμβάνει έγκυρες και συγκρίσιμες παρατηρήσεις), ενώ στον κόμβο Transformations πραγματοποιούνται οι μετασχηματισμοί που απαιτούνται πριν από distance-based τμηματοποίηση (π.χ. κανονικοποίηση/τυποποίηση και οι χειρισμοί ακραίων τιμών που περιγράφηκαν στο προηγούμενο υποκεφάλαιο). Ακολουθεί ο κόμβος Imputation, ο οποίος διασφαλίζει ότι τυχόν ελλείπουσες τιμές δεν οδηγούν σε απώλεια καταστημάτων από την ανάλυση και ότι το dataset μπορεί να χρησιμοποιηθεί ενιαία από τον αλγόριθμο.

Ο πυρήνας της διαδικασίας βρίσκεται στον κόμβο Clustering, όπου εφαρμόστηκε ο επιλεγμένος αλγόριθμος τμηματοποίησης για την παραγωγή των τελικών ομάδων καταστημάτων. Έπειτα, ο κόμβος Segment Profiling χρησιμοποιήθηκε για τη συνοπτική αποτύπωση/προφίλ των παραγόμενων clusters (μέσοι όροι, μεγέθη, συγκρίσεις ανά μεταβλητή), διευκολύνοντας την επιχειρησιακή ερμηνεία και την ονοματοδοσία τους. Τέλος, μέσω του Save Data αποθηκεύτηκε το τελικό dataset με την ετικέτα cluster/segment ανά κατάσταση, ώστε να μπορεί να αξιοποιηθεί τόσο σε πίνακες/διαγράμματα όσο και σε επόμενα στάδια ανάλυσης.



Εικόνα 22 – Τελικό Pipeline Shop Clustering στο SAS Model Studio

Επειδή η συγκεκριμένη εργασία αφορά μη-επιβλεπόμενη μάθηση (unsupervised learning), δεν απαιτήθηκε διάσπαση σε training/validation/test όπως συμβαίνει σε μοντέλα πρόβλεψης. Για τον λόγο αυτό, αξιοποιήθηκε το σύνολο των διαθέσιμων παρατηρήσεων για την εκπαίδευση της τμηματοποίησης (Training = 100%, Validation/Test = 0%), όπως φαίνεται στην Εικόνα 23. Στη συνέχεια τεκμηριώνεται η επιλογή του αριθμού συστάδων και ακολούθως παρουσιάζονται τα τελικά segments που προέκυψαν.

Project Settings

Partition Data
Event-Based Sampling
Node Configuration
Rules
Output Library
Logging
Compute Context

Partition Data

☒ Create partition variable

Note: These settings are active only when a partition variable is not set within the data. Using a data source with a pre-defined partition variable or manually selecting a partition variable will override these settings.

Method:
Stratify

Training:

100	100.00%
-----	---------

Validation:

0	0.00%
---	-------

Test:

0	0.00%
---	-------

Εικόνα 23 – Project Settings

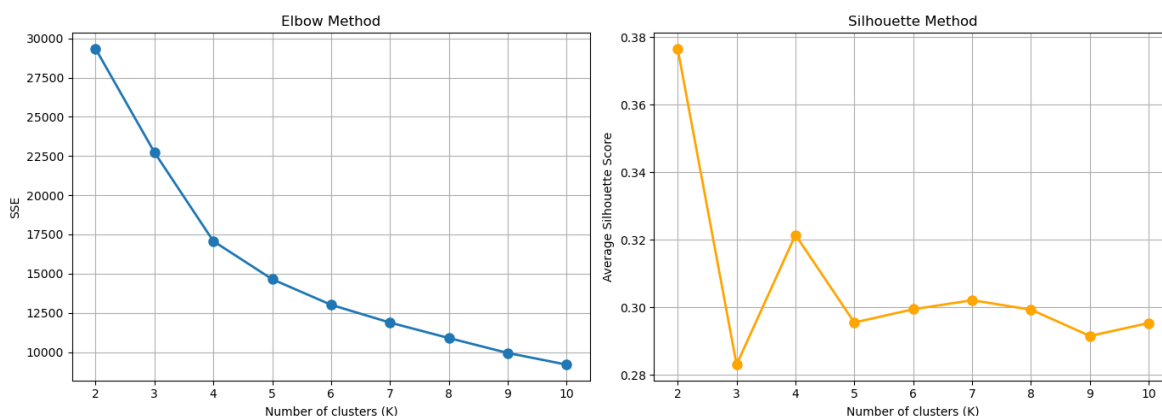
Κρίσιμο βήμα πριν την ερμηνεία των αποτελεσμάτων είναι η επιλογή του πλήθους των clusters (K). Στο SAS Model Studio, η επιλογή του K πραγματοποιήθηκε αυτόματα μέσω του Aligned Box Criterion (ABC). Η λογική του ABC είναι ότι συγκρίνει λύσεις clustering για διαφορετικά K μέσα σε προκαθορισμένο εύρος (π.χ. 2 έως 6) και επιλέγει εκείνη που προσφέρει τον καλύτερο συμβιβασμό μεταξύ συμπαγών συστάδων και ουσιαστικού διαχωρισμού, αποφεύγοντας υπερβολικά μικρό ή υπερβολικά μεγάλο αριθμό ομάδων. Στα αποτελέσματα του ABC εμφανίζεται και ένας δείκτης “Gap” στο πλαίσιο του συγκεκριμένου κριτηρίου, ως μέτρο σύγκρισης της δομής των πραγματικών δεδομένων με μια αναφορά (reference). Με βάση το κριτήριο GlobalPeak, το προτεινόμενο πλήθος clusters ήταν K = 5.

Για λόγους πληρότητας και για να διερευνηθεί περαιτέρω η επιλογή του πλήθους συστάδων (K), εφαρμόστηκαν επιπρόσθετα οι μέθοδοι Elbow και Silhouette σε Python, πάνω στο ίδιο σύνολο εισόδου που χρησιμοποιήθηκε για τη συσταδοποίηση (οι τελικές μεταβλητές: avg_shor_monthly_orders, avg_delivery_time_minutes, days_on_platform), μετά την προεπεξεργασία/κανονικοποίηση. Στο διάγραμμα Elbow (SSE/Inertia ως προς K), παρατηρείται απότομη μείωση του σφάλματος για μικρές τιμές του K και στη συνέχεια επιβράδυνση του ρυθμού μείωσης, κάτι που υποδεικνύει σημείο καμπής γύρω από K=4–5. Η λογική του Elbow είναι ότι, μετά

Επιχειρησιακή Αναλυτική στην Πράξη: Τμηματοποίηση και Πρόβλεψη Αποχώρησης Πελατών στον Κλάδο Ηλεκτρονικής Διανομής Φαγητού 48

από αυτό το σημείο, η προσθήκη επιπλέον συστάδων οδηγεί σε μικρό κέρδος (μικρή επιπλέον μείωση του SSE) σε σχέση με την αύξηση της πολυπλοκότητας του μοντέλου.

Παράλληλα, το Silhouette score αξιολογεί τη συνοχή και τον διαχωρισμό των συστάδων, συγκρίνοντας πόσο «κοντά» είναι κάθε παρατήρηση στο cluster της έναντι των γειτονικών clusters. Στα αποτελέσματα παρατηρείται ότι το μέγιστο silhouette επιτυγχάνεται για μικρότερο K (π.χ. K=2), ενώ εμφανίζεται και τοπικά αυξημένη επίδοση για K=4. Αυτό είναι αναμενόμενο σε δεδομένα πραγματικών επιχειρησιακών εφαρμογών, καθώς ο δείκτης silhouette τείνει συχνά να ευνοεί πιο «χονδρικούς» διαχωρισμούς (λίγες συστάδες με καθαρότερη γεωμετρική διάκριση), οι οποίοι όμως δεν οδηγούν απαραίτητα σε επαρκώς λεπτομερή τμηματοποίηση για επιχειρησιακή αξιοποίηση.



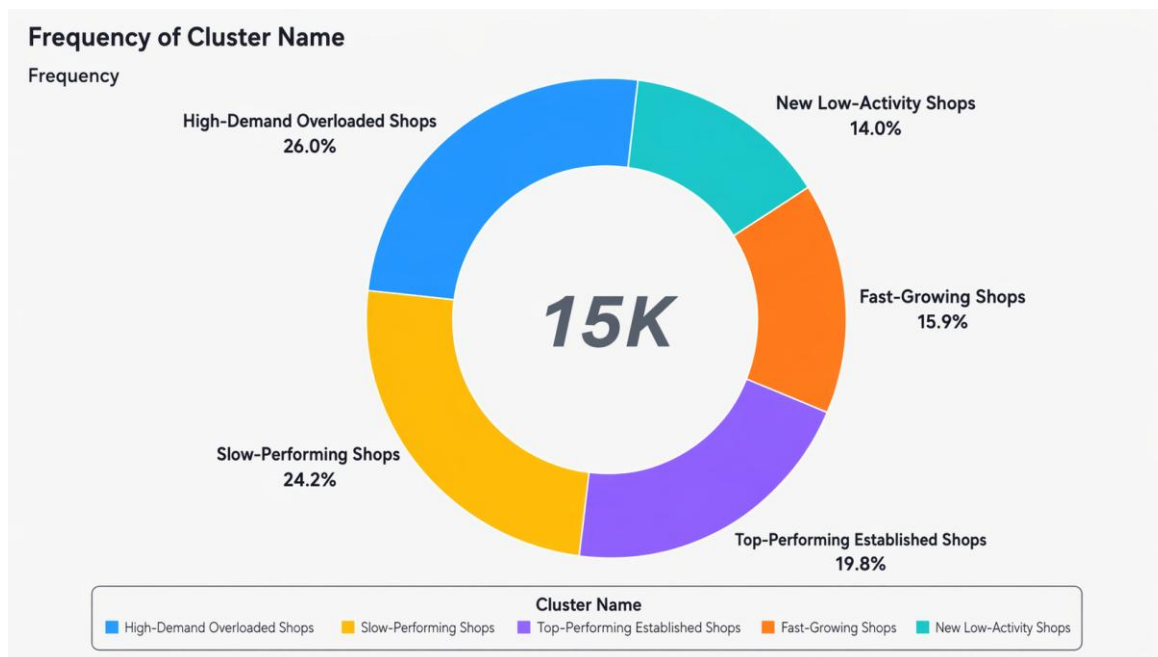
Εικόνα 24 – Elbow και Silhouette Methods για την Επιλογή Αριθμού Συστάδων (K)

Συνολικά, οι δύο εναλλακτικές μέθοδοι λειτουργούν ως συμπληρωματικοί έλεγχοι και υποδεικνύουν ένα εύλογο εύρος τιμών του K (περίπου 4–5). Ωστόσο, η τελική επιλογή διατηρείται ως K=5 σύμφωνα με το Gap Statistic του SAS Model Studio, καθώς παρέχει πιο ισορροπημένη και επιχειρησιακά ερμηνεύσιμη τμηματοποίηση, επιτρέποντας τη διάκριση περισσότερων προφίλ καταστημάτων χωρίς υπερβολική αύξηση της πολυπλοκότητας. Με βάση την τελική επιλογή K=5, στην επόμενη ενότητα παρουσιάζονται η κατανομή και το προφίλ των συστάδων (cluster profiling) και η ερμηνεία τους.

6.2.2 Ανάλυση και ερμηνεία των τελικών clusters

Στην παρούσα ενότητα παρουσιάζονται η κατανομή και το profiling των πέντε (5) clusters που προέκυψαν από την τελική επιλογή K=5 (βλ. ενότητα 6.2.1).

Η εφαρμογή του clustering οδήγησε σε πέντε (5) διακριτές συστάδες για 14.579 καταστήματα. Η κατανομή των καταστημάτων ανά cluster (σε ποσοστά και πλήθος) παρουσιάζεται στην Εικόνα 25 και δείχνει ότι οι ομάδες δεν είναι ακραία ανισομεγέθεις, κάτι που βοηθά ώστε το αποτέλεσμα να είναι πιο σταθερό και αξιοποιήσιμο επιχειρησιακά.

**Εικόνα 25 - Κατανομή Καταστημάτων ανά Cluster (K=5)**

Στη συνέχεια, πραγματοποιήθηκε profiling των συστάδων, δηλαδή ανάλυση των μέσων τιμών των μεταβλητών ανά cluster, ώστε κάθε ομάδα να αποκτήσει ερμηνεύσιμη επιχειρησιακή ταυτότητα. Ο συνοπτικός πίνακας profiling παρουσιάζεται στην Εικόνα 26, και από αυτόν προκύπτουν τα παρακάτω προφίλ:

Cluster Name	Frequency	Frequency Percent	avg_delivery_time_minutes	avg_shop_monthly_orders	days_on_platform
Fast-Growing Shops	2,324	15.94%	19.82	119.08	520.65
High-Demand Overloaded Shops	3,790	26.00%	37.73	569.59	936.36
New Low-Activity Shops	2,038	13.98%	30.63	31.89	204.13
Slow-Performing Shops	3,534	24.24%	33.97	95.72	536.16
Top-Performing Established Shops	2,893	19.84%	27.73	406.59	1,331.56
Άθροισμα	14,579	100.00%	30.99	275.40	749.15

Εικόνα 26 – Cross-Tabulation Table των 5 Clusters

Το πρώτο cluster περιγράφει καταστήματα με πολύ καλή λειτουργική απόδοση (χαμηλοί χρόνοι παράδοσης), μεσαίο επίπεδο ζήτησης και μέτρια παρουσία στην πλατφόρμα. Πρόκειται για καταστήματα που δείχνουν σημάδια ανάπτυξης: λειτουργούν αποδοτικά και έχουν “χτίσει” ένα σταθερό επίπεδο δραστηριότητας, χωρίς όμως να έχουν φτάσει ακόμη σε πολύ υψηλούς όγκους παραγγελιών. Αυτή η κατηγορία μπορεί να θεωρηθεί “αναπτυσσόμενη”, καθώς έχει προϋποθέσεις να μετακινηθεί σε πιο υψηλή απόδοση/ζήτηση εφόσον υποστηριχθεί σωστά.

Το δεύτερο cluster περιλαμβάνει καταστήματα με πολύ υψηλή ζήτηση (μεγάλος όγκος παραγγελιών), αλλά ταυτόχρονα αυξημένους χρόνους παράδοσης, ενώ συνήθως εμφανίζουν και υψηλή “ωριμότητα”/παλαιότητα στην πλατφόρμα. Το προφίλ αυτό δείχνει καταστήματα που “τραβάνε”

μεγάλο όγκο, όμως πιθανόν πιέζονται λειτουργικά (capacity constraints, οργάνωση, προετοιμασία, διαχείριση αιχμών), με αποτέλεσμα να υποχωρεί η ταχύτητα εξυπηρέτησης. Αποτελούν κρίσιμη κατηγορία για την πλατφόρμα, διότι έχουν σημαντική συνεισφορά στη συνολική δραστηριότητα, αλλά ταυτόχρονα είναι και πιο εκτεθειμένα σε ζητήματα ποιότητας υπηρεσίας.

Το τρίτο cluster χαρακτηρίζεται από καταστήματα με χαμηλή δραστηριότητα (πολύ μικρός όγκος παραγγελιών) και μικρή παρουσία στην πλατφόρμα, δηλαδή τυπικά “νέα” ή σχετικά πρόσφατα ενταγμένα καταστήματα. Οι χρόνοι παράδοσης κινούνται σε μέτρια επίπεδα, όμως το βασικό χαρακτηριστικό εδώ είναι ότι η ζήτηση δεν έχει ακόμη αναπτυχθεί. Αυτό το cluster είναι σημαντικό γιατί περιλαμβάνει καταστήματα που βρίσκονται σε “φάση εκκίνησης” και χρειάζονται ενεργοποίηση.

Το τέταρτο cluster αντιστοιχεί σε καταστήματα με χαμηλό έως μεσαίο όγκο παραγγελιών, αλλά σχετικά υψηλούς χρόνους παράδοσης, ενώ η παρουσία τους στην πλατφόρμα είναι συνήθως ούτε πολύ νέα ούτε πλήρως “established”. Το προφίλ αυτό παραπέμπει σε καταστήματα που δεν έχουν υψηλή ζήτηση, αλλά ταυτόχρονα δεν ξεχωρίζουν και λειτουργικά, κάτι που τα κάνει λιγότερο ανταγωνιστικά. Η συγκεκριμένη ομάδα είναι σημαντική επιχειρησιακά, γιατί συχνά αντιπροσωπεύει περιπτώσεις όπου απαιτείται “διάγνωση” των αιτιών: μπορεί να υπάρχουν ζητήματα ποιότητας εξυπηρέτησης, ελλιπής οργάνωση, χαμηλή προβολή ή χαμηλή ελκυστικότητα προϊόντων.

Το πέμπτο cluster περιλαμβάνει καταστήματα με υψηλή ζήτηση, καλούς χρόνους παράδοσης και πολύ μεγάλη παρουσία στην πλατφόρμα. Πρόκειται για ώριμα, σταθερά και αποδοτικά καταστήματα που συνδυάζουν όγκο και λειτουργική συνέπεια, αποτελώντας πρακτικά την “κορυφαία” κατηγορία. Αυτά τα καταστήματα είναι ιδιαίτερα σημαντικά για την πλατφόρμα, καθώς λειτουργούν ως benchmark και έχουν υψηλή επιχειρησιακή αξία.

Συνολικά, τα πέντε clusters αποτυπώνουν διακριτά επιχειρησιακά προφίλ καταστημάτων, συνδυάζοντας τη διάσταση της ζήτησης, της λειτουργικής απόδοσης και της ωριμότητας/παρουσίας στην πλατφόρμα. Ο διαχωρισμός αυτός δεν είναι μόνο στατιστικά χρήσιμος, αλλά επιτρέπει και τη μετάφραση των αποτελεσμάτων σε πρακτικές ενέργειες ανά κατηγορία. Στην επόμενη ενότητα (6.2.3) προτείνονται στοχευμένες επιχειρησιακές παρεμβάσεις για κάθε cluster, με στόχο τη βελτίωση απόδοσης, την ανάπτυξη και τη διατήρηση συνεργασιών.

6.2.3 Επιχειρησιακές προτάσεις ανά cluster καταστημάτων

Σύμφωνα με την τμηματοποίηση που προηγήθηκε, τα καταστήματα της πλατφόρμας δεν αποτελούν έναν ομοιογενή πληθυσμό: εμφανίζουν διαφορετικά επίπεδα ζήτησης, διαφορετική λειτουργική απόδοση (χρόνοι παράδοσης) και διαφορετικό στάδιο ωριμότητας στην πλατφόρμα. Επομένως, η πλατφόρμα e-food δεν είναι αποτελεσματικό να εφαρμόζει ενιαία πολιτική υποστήριξης και εμπορικής συνεργασίας για όλα τα καταστήματα. Αντίθετα, προτείνονται στοχευμένες και διαφοροποιημένες ενέργειες ανά cluster, ώστε να μεγιστοποιείται η απόδοση των συνεργασιών, να βελτιώνεται η εμπειρία πελάτη και να μειώνονται σημεία τριβής που επηρεάζουν τη ζήτηση και την ποιότητα υπηρεσίας.

Cluster 1 – Fast-Growing Shops (αναπτυσσόμενα, αποδοτικά)

Η ομάδα αυτή χαρακτηρίζεται από καλή λειτουργική επίδοση (χαμηλοί χρόνοι παράδοσης) και ένα σταθερό, μεσαίο επίπεδο παραγγελιών. Πρόκειται για καταστήματα που δείχνουν ότι “λειτουργούν σωστά” και έχουν προοπτική να ανέβουν κλίμακα, αρκεί να ενισχυθεί η προβολή και να αρθούν πιθανά εμπόδια ανάπτυξης (π.χ. περιορισμένη προβολή, μη βέλτιστη παρουσίαση μενού, ελλιπείς ώρες λειτουργίας).

Προτεινόμενες ενέργειες:

- Στοχευμένη ενίσχυση προβολής (ranking boosts σε κατάλληλες ώρες/ζώνες, συμμετοχή σε curated λίστες “recommended”, μικρά promo windows) με στόχο να αυξηθεί η ζήτηση χωρίς να “σπάσει” η λειτουργική τους ικανότητα.
- Βελτιστοποίηση καταλόγου/μενού: προτάσεις για καλύτερη φωτογράφιση, ορθότερη κατηγοριοποίηση, προσθήκη best-sellers και ενίσχυση value bundles (ο πελάτης αυξάνει basket, το κατάστημα αυξάνει σταθερά τον όγκο).
- Light incentives για ανάπτυξη (π.χ. μικρές επιδοτήσεις προσφορών ή συμμετοχή σε καμπάνιες νέων περιοχών) ώστε να μετακινηθούν σταδιακά προς υψηλότερο cluster.
- Παρακολούθηση capacity: επειδή αυτά τα καταστήματα έχουν δυναμική, η πλατφόρμα μπορεί να εφαρμόσει “guardrails” (π.χ. monitoring χρόνων/αποδοχής) ώστε η αύξηση όγκου να μην οδηγήσει σε αύξηση delivery time.

Cluster 2 – High-Demand Overloaded Shops (υψηλής ζήτησης αλλά επιβαρυνμένα)

Αυτά τα καταστήματα έχουν πολύ μεγάλο όγκο παραγγελιών, αλλά εμφανίζουν αυξημένους χρόνους παράδοσης. Είναι από τα πιο κρίσιμα για την πλατφόρμα, γιατί συνεισφέρουν σημαντικά στον συνολικό όγκο και τζίρο, όμως ταυτόχρονα είναι πιο εκτεθειμένα σε αρνητικές εμπειρίες (καθυστερήσεις, δυσανεμία, ακυρώσεις, χαμηλότερες αξιολογήσεις). Εδώ η βασική προτεραιότητα δεν είναι να αυξηθεί η ζήτηση (ήδη είναι υψηλή), αλλά να βελτιωθεί η επιχειρησιακή απόδοση.

Προτεινόμενες ενέργειες:

- Operational support / capacity planning: ανάλυση ωρών αιχμής, προτάσεις για προετοιμασία, staffing και ροή παραγωγής. Σε συνεργασία με το κατάστημα, θέσπιση στόχων για μείωση χρόνων σε ώρες αιχμής.
- Διαχείριση ζήτησης με “έξυπνο τρόπο”: προσωρινή ρύθμιση του exposure σε ώρες αιχμής, ή προώθηση χρονικών παραθύρων (scheduled orders) ώστε να εξομαλύνεται ο φόρτος και να μειώνεται το congestion.
- Προτεραιοποίηση διανομής (όπου εφαρμόζεται): σε περιοχές όπου η διανομή πραγματοποιείται από την πλατφόρμα ή συνδυαστικά, μπορεί να υπάρξει καλύτερη κατανομή οδηγών/συνοδηγών ή κανόνες dispatching ώστε να αποφεύγονται μεγάλα queues.
- Βελτίωση χρόνου προετοιμασίας με παρεμβάσεις προϊόντος: “busy mode” με μικρές αλλαγές στο μενού (π.χ. προσωρινή απόκρυψη items που καθυστερούν), καλύτερες επιλογές packaging/standardization.
- Πλαίσιο ποιότητας: έλεγχος ακυρώσεων, complaints και ratings, με feedback προς το κατάστημα, ώστε το υψηλό demand να μην οδηγεί σε μακροπρόθεσμη φθορά.

Cluster 3 – New Low-Activity Shops (νέα/χαμηλής δραστηριότητας, χρειάζονται ενεργοποίηση)

Το cluster αυτό αφορά κυρίως καταστήματα με μικρή παρουσία στην πλατφόρμα και πολύ χαμηλό όγκο παραγγελιών. Αυτό δεν σημαίνει απαραίτητα ότι υστερούν ποιοτικά — συνήθως σημαίνει ότι δεν έχουν ακόμη αποκτήσει προβολή, “ιστορικό”, αξιολογήσεις και σταθερή ζήτηση. Εδώ ο στόχος είναι η ενεργοποίηση και το σωστό onboarding.

Προτεινόμενες ενέργειες:

- Onboarding πρόγραμμα 30–60 ημερών: οδηγός βέλτιστων πρακτικών (μενού, ώρες, φωτογραφίες, minimum order, delivery radius), ώστε να βελτιωθεί η μετατροπή (conversion) από view σε order.
- Starter promotions: μικρές εκπτώσεις ή κουπόνια πρώτης/δεύτερης παραγγελίας, ώστε να “σπάσει” η αδράνεια και να αποκτήσουν οι πρώτες παραγγελίες momentum.
- Boost αξιολογήσεων: ενθάρρυνση για συλλογή ratings (π.χ. σωστή συσκευασία, συνέπεια, μικρές κινήσεις εξυπηρέτησης) γιατί οι αξιολογήσεις επηρεάζουν δραστικά το future demand.
- Τοπική στόχευση: καλύτερη τοποθέτηση σε αναζητήσεις/κατηγορίες στη γειτονιά τους, και προτάσεις για “value” bundles ώστε να αυξηθεί το attractiveness για νέους πελάτες.
- Έλεγχος βασικών KPIs ενεργοποίησης: αν σε συγκεκριμένο διάστημα δεν αυξάνεται η ζήτηση, χρειάζεται διάγνωση (λάθος κατηγοριοποίηση, υψηλό minimum order, κακή φωτογράφιση, πολύ μικρό ωράριο κ.λπ.).

Cluster 4 – Slow-Performing Shops (χαμηλή–μεσαία ζήτηση και αδύναμη λειτουργική επίδοση)

Η ομάδα αυτή συνδυάζει χαμηλότερο όγκο παραγγελιών με υψηλότερους χρόνους παράδοσης και δεν είναι “καινούρια”. Άρα δεν μιλάμε απλά για onboarding, αλλά για καταστήματα που πιθανόν έχουν σταθεροποιηθεί σε μέτρια/χαμηλή απόδοση. Αυτό το cluster είναι κρίσιμο γιατί συχνά “χάνει” ζήτηση λόγω εμπειρίας (delivery time) ή λόγω χαμηλής ελκυστικότητας/προβολής.

Προτεινόμενες ενέργειες:

- Διάγνωση αιτιών (diagnostic checklist): έλεγχος λειτουργικών δεικτών (αποδοχή παραγγελιών, ακυρώσεις, καθυστερήσεις), αλλά και εμπορικών παραγόντων (τιμές, φωτογραφίες, κατηγορίες, ανταγωνισμός, ώρες αιχμής).
- Targeted improvement plan: μικρό σχέδιο βελτίωσης με 2–3 συγκεκριμένους στόχους (π.χ. μείωση χρόνων προετοιμασίας, βελτίωση menu structure, αύξηση availability).
- Quality & visibility interventions: ενίσχυση ποιότητας υπηρεσίας (συσκευασία, accuracy) και παράλληλα βελτίωση ορατότητας στην πλατφόρμα, μόνο εφόσον υπάρχει πρόοδος στους χρόνους/ακυρώσεις.
- Κίνητρα υπό προϋποθέσεις: όχι “τυφλά” promos. Π.χ. συμμετοχή σε καμπάνιες μόνο αν πετυχαίνουν ένα όριο αξιοπιστίας/χρόνου, ώστε να μην ξοδεύεται budget σε καταστήματα που δεν μπορούν να εξυπηρετήσουν σωστά.

- Επανατοποθέτηση (re-positioning): σε περιπτώσεις όπου η ζήτηση είναι χαμηλή λόγω προσφοράς, ενδέχεται να χρειαστεί αλλαγή στρατηγικής (π.χ. focus σε συγκεκριμένες κατηγορίες, value menus, ώρες λειτουργίας).

Cluster 5 – Top-Performing Established Shops (ώριμα και κορυφαία)

Τα καταστήματα αυτά συνδυάζουν υψηλή ζήτηση, καλή λειτουργική απόδοση και μεγάλη παρουσία στην πλατφόρμα. Αποτελούν ουσιαστικά την πιο “ποιοτική και σταθερή” βάση συνεργατών. Εδώ ο στόχος είναι η διατήρηση, η ενίσχυση της σχέσης και η αποφυγή φθοράς (π.χ. να μην μετακινηθούν σε overloaded profile ή να μην μειώσουν την απόδοση).

Προτεινόμενες ενέργειες:

- Retention συνεργασίας / premium partnership: προνομιακή υποστήριξη, ταχύτερη εξυπηρέτηση σε θέματα τεχνικά/εμπορικά, early access σε features.
- Co-marketing: κοινές καμπάνιες, curated placements, συμμετοχή σε flagship προωθήσεις της πλατφόρμας, γιατί έχουν ήδη αποδεδειγμένη απόδοση.
- Benchmarking & knowledge sharing: αξιοποίηση αυτών των καταστημάτων ως “best practice” για άλλες ομάδες (π.χ. οδηγίες για peak preparation, menu optimization).
- Προληπτικός έλεγχος ποιότητας: monitoring για σημάδια κόπωσης (αύξηση χρόνων, complaints) ώστε να προλαμβάνεται η πτώση.
- Ανάπτυξη μέσω νέων υπηρεσιών: όπου είναι εφικτό, δυνατότητα συμμετοχής σε νέες υπηρεσίες/μοντέλα (π.χ. bundles, subscriptions, corporate/catering συνεργασίες) που ενισχύουν τη σχέση και αυξάνουν GMV χωρίς να διακινδυνεύουν την ποιότητα.

Συνολικά, οι παραπάνω προτάσεις αναδεικνύουν πώς τα αποτελέσματα του shop clustering μπορούν να μεταφραστούν σε διαφοροποιημένες επιχειρησιακές στρατηγικές, λαμβάνοντας υπόψη τόσο τη ζήτηση όσο και τη λειτουργική απόδοση των καταστημάτων. Με τον τρόπο αυτό, η πλατφόρμα μπορεί να καταναίμει πόρους και παρεμβάσεις στοχευμένα, μεγιστοποιώντας την αποτελεσματικότητα των ενεργειών της.

7. Πρόβλεψη Αποχώρησης Καταστημάτων (Shop Churn Prediction)

7.1 Πειραματική Διαδικασία

Η παρούσα ενότητα παρουσιάζει τη διαδικασία που ακολουθήθηκε για την ανάπτυξη μοντέλων πρόβλεψης αποχώρησης καταστημάτων από την πλατφόρμα, με σκοπό τον έγκαιρο εντοπισμό καταστημάτων που εμφανίζουν αυξημένη πιθανότητα να αποχωρήσουν. Η ανάλυση πραγματοποιήθηκε σε επίπεδο καταστήματος, όπου κάθε παρατήρηση αντιστοιχεί σε ένα μοναδικό συνεργαζόμενο κατάστημα και περιγράφεται μέσω συγκεντρωτικών δεικτών που αποτυπώνουν τη λειτουργική και εμπορική του συμπεριφορά.

Στο πλαίσιο αυτό, αναπτύχθηκε ένα μοντέλο ταξινόμησης που διαχωρίζει τα καταστήματα με ενδείξεις αποχώρησης από εκείνα που παραμένουν ενεργά. Η μεθοδολογία δεν περιορίστηκε στην εφαρμογή αλγορίθμων, αλλά οργανώθηκε σε διακριτά στάδια: αρχικά ορίστηκε η μεταβλητή-στόχος, στη συνέχεια επιλέχθηκε ο κατάλληλος πληθυσμός ανάλυσης, διαμορφώθηκε το τελικό σύνολο

δεδομένων για μοντελοποίηση και τέλος πραγματοποιήθηκε εκπαίδευση και συγκριτική αξιολόγηση εναλλακτικών μοντέλων.

Ιδιαίτερη έμφαση δόθηκε στον ορισμό ενός επιχειρησιακά εύλογου και μεθοδολογικά τεκμηριωμένου proxy για το churn, καθώς στο διαθέσιμο σύνολο δεδομένων δεν υπήρχε ρητή ένδειξη αποχώρησης. Παράλληλα, αξιολογήθηκε η καταλληλότητα των διαθέσιμων μεταβλητών ως είσοδοι στα μοντέλα, με στόχο να αποφευχθεί η διαρροή πληροφορίας (data leakage) και να ενισχυθεί η αξιοπιστία των προβλέψεων.

Στις επόμενες υποενότητες παρουσιάζονται ο ορισμός της μεταβλητής-στόχου, η δημιουργία του τελικού συνόλου μοντελοποίησης, καθώς και η διαδικασία υλοποίησης και αξιολόγησης των μοντέλων στο περιβάλλον του SAS Model Studio.

7.1.1 Ορισμός της μεταβλητής-στόχου και διαμόρφωση τελικού συνόλου μοντελοποίησης

Η ανάπτυξη ενός μοντέλου πρόβλεψης αποχώρησης καταστημάτων προϋποθέτει, ως πρώτο και βασικό βήμα, τον ορισμό της μεταβλητής-στόχου, δηλαδή της μεταβλητής που θα χρησιμοποιηθεί για τον διαχωρισμό των καταστημάτων σε churned και non-churned. Στην παρούσα εργασία, ο ορισμός αυτός δεν μπορούσε να προκύψει άμεσα από το διαθέσιμο σύνολο δεδομένων, καθώς δεν υπήρχε ρητή μεταβλητή που να δηλώνει τον τερματισμό της συνεργασίας ενός καταστήματος με την πλατφόρμα. Για τον λόγο αυτό, κρίθηκε απαραίτητο το churn να προσεγγιστεί μέσω ενός συμπεριφορικού proxy, βασισμένου στη δραστηριότητα των καταστημάτων και ειδικότερα στον χρόνο αδράνειας τους.

Η επιλογή της αδράνειας ως proxy churn θεωρήθηκε εύλογη, καθώς ένα κατάστημα που παραμένει μεγάλο χρονικό διάστημα χωρίς νέα παραγγελία είναι πιθανότερο να έχει περιορίσει σημαντικά ή να έχει ουσιαστικά διακόψει τη δραστηριότητά του στην πλατφόρμα. Με βάση αυτή τη λογική, ως κύρια μεταβλητή αναφοράς χρησιμοποιήθηκε η last_order_at, η οποία αποτυπώνει την ημερομηνία τελευταίας παραγγελίας κάθε καταστήματος. Ως κοινή ημερομηνία αναφοράς για όλα τα καταστήματα χρησιμοποιήθηκε η μέγιστη τιμή της μεταβλητής αυτής στο σύνολο δεδομένων, δηλαδή η πιο πρόσφατη ημερομηνία παραγγελίας που ήταν διαθέσιμη. Στη συνέχεια, για κάθε κατάστημα υπολογίστηκε η μεταβλητή days_inactive, η οποία εκφράζει το πλήθος των ημερών που έχουν μεσολαβήσει από την τελευταία παραγγελία μέχρι την ημερομηνία αναφοράς.

Πριν όμως εφαρμοστεί οποιοδήποτε όριο αδράνειας, κρίθηκε αναγκαίο να διαμορφωθεί πρώτα ο κατάλληλος πληθυσμός ανάλυσης. Το αρχικό σύνολο δεδομένων περιλάμβανε 17.238 συνεργαζόμενα καταστήματα, όμως δεν θεωρήθηκε μεθοδολογικά ορθό να αξιολογηθούν όλα με τα ίδια κριτήρια. Ο βασικός λόγος ήταν ότι ένα μέρος αυτών αντιστοιχούσε σε νεοεισερχόμενα καταστήματα με πολύ μικρή χρονική παρουσία στην πλατφόρμα, τα οποία δεν διέθεταν επαρκές ιστορικό δραστηριότητας. Εάν τα καταστήματα αυτά περιλαμβάνονταν κανονικά στον ορισμό του churn, θα υπήρχε αυξημένος κίνδυνος λανθασμένης ταξινόμησής τους, είτε επειδή δεν είχαν προλάβει να αναπτύξουν φυσιολογική δραστηριότητα είτε επειδή η περιορισμένη χρονική τους παρουσία θα επηρέαζε τεχνητά τον υπολογισμό της αδράνειας.

Για τον λόγο αυτό εφαρμόστηκε κριτήριο επιλεξιμότητας, σύμφωνα με το οποίο διατηρήθηκαν μόνο τα καταστήματα με τουλάχιστον 90 ημέρες παρουσίας στην πλατφόρμα. Με την εφαρμογή του φίλτρου αυτού, από το αρχικό σύνολο των 17.238 καταστημάτων προέκυψε eligible πληθυσμός 16.319 καταστημάτων, ενώ 919 καταστήματα εξαιρέθηκαν από την περαιτέρω διαδικασία. Η επιλογή αυτή κρίθηκε απαραίτητη ώστε ο ορισμός του churn να εφαρμόζεται μόνο σε καταστήματα με επαρκές ιστορικό παρατήρησης και να αποφεύγονται μεροληπτικές ταξινομήσεις που σχετίζονται με τη μικρή διάρκεια παρουσίας στην πλατφόρμα.

Initial Shops	Eligible Shops	Excluded Shops	Excluded %	Eligible %
17238	16319	919	5.33%	94.67%

Εικόνα 27 - Summary of Eligible Sample

Αφού ορίστηκε ο επιλέξιμος πληθυσμός, το επόμενο βήμα ήταν να καθοριστεί το όριο πάνω από το οποίο ένα κατάστημα θα χαρακτηριζόταν ως churned. Αντί να επιλεγεί εξ αρχής ένας αυθαίρετος αριθμός ημερών, ακολουθήθηκε data-driven προσέγγιση, βασισμένη στην πραγματική κατανομή της μεταβλητής `days_inactive` στο `eligible` δείγμα. Στο πλαίσιο αυτό πραγματοποιήθηκε ανάλυση ευαισθησίας (sensitivity analysis), κατά την οποία εξετάστηκαν τρία εναλλακτικά σενάρια: το ανώτερο 15%, το ανώτερο 10% και το ανώτερο 5% της κατανομής αδράνειας. Τα σενάρια αυτά αντιστοιχούσαν σε ποσοστά churn 15,11%, 10,03% και 5,01% αντίστοιχα. Η σύγκριση των τριών σεναρίων έδειξε ότι το ανώτερο 10% της κατανομής του `days_inactive` προσφέρει την πιο ισορροπημένη προσέγγιση μεταξύ αυστηρότητας και επιχειρησιακής ερμηνευσιμότητας. Το σενάριο του ανώτερου 15% κρίθηκε σχετικά πιο χαλαρό, καθώς ενδέχεται να περιλαμβάνει και καταστήματα με παροδική ή συγκυριακή μείωση δραστηριότητας, ενώ το σενάριο του ανώτερου 5% θεωρήθηκε αρκετά αυστηρό, περιορίζοντας υπερβολικά το πλήθος των καταστημάτων που θα χαρακτηρίζονταν ως churned. Αντίθετα, το ανώτερο 10% φάνηκε να αποτελεί έναν πιο εύλογο συμβιβασμό μεταξύ ρεαλιστικής αποτύπωσης της αποχώρησης και καταλληλότητας για ανάπτυξη μοντέλων πρόβλεψης.

Με βάση το συγκεκριμένο κύριο σενάριο, ως churners ορίστηκαν τα καταστήματα που είχαν να εμφανίσουν παραγγελία για χρονικό διάστημα ίσο ή μεγαλύτερο των 107 ημερών από την κοινή ημερομηνία αναφοράς. Αντίστοιχα, τα καταστήματα με αδράνεια μικρότερη των 107 ημερών θεωρήθηκαν non-churned. Με αυτόν τον τρόπο, η μεταβλητή-στόχος y ορίστηκε ως δυαδική μεταβλητή, όπου $y = 1$ αντιστοιχεί σε churned κατάστημα και $y = 0$ σε non-churned κατάστημα. Στον τελικό `eligible` πληθυσμό προέκυψαν 1.636 churners και 14.683 non-churners.

Scenario	Threshold (days_inactive)	Churners	Non-churners	Eligible Shops	Churn Rate
Top 15%	52	2466	13853	16319	15.11%
Top 10%	107	1636	14683	16319	10.03%
Top 5%	220	817	15502	16319	5.01%

Εικόνα 28 - Sensitivity Analysis for Churn Definition

Μετά τον ορισμό της μεταβλητής-στόχου, εξετάστηκε και η καταλληλότητα των διαθέσιμων μεταβλητών ως εισόδων στα μοντέλα πρόβλεψης. Ιδιαίτερη προσοχή δόθηκε στον εντοπισμό μεταβλητών που θα μπορούσαν να οδηγήσουν σε διαρροή πληροφορίας (data leakage), δηλαδή μεταβλητών που συνδέονται άμεσα ή έμμεσα με τον τρόπο ορισμού του `target`. Με αυτή τη λογική, μεταβλητές όπως η `last_order_at`, καθώς και δείκτες που σχετίζονται άμεσα με την πρόσφατη δραστηριότητα των καταστημάτων, δεν χρησιμοποιήθηκαν ως predictive inputs. Με τον τρόπο αυτό διαμορφώθηκε το τελικό σύνολο μοντελοποίησης, το οποίο αποτέλεσε τη βάση για την ανάπτυξη και συγκριτική αξιολόγηση των μοντέλων πρόβλεψης churn.

7.1.2 Διαδικασία υλοποίησης στο SAS Model Studio

Η ανάπτυξη των μοντέλων πρόβλεψης churn πραγματοποιήθηκε στο περιβάλλον του SAS Model Studio, το οποίο χρησιμοποιήθηκε ως το βασικό υπολογιστικό πλαίσιο για την εκπαίδευση, αξιολόγηση και σύγκριση των εναλλακτικών αλγορίθμων. Η χρήση ενός ενιαίου περιβάλλοντος υλοποίησης κρίθηκε σημαντική, καθώς επέτρεψε την εφαρμογή όλων των μοντέλων πάνω στο ίδιο

σύνολο δεδομένων και κάτω από κοινές συνθήκες ανάλυσης, διευκολύνοντας έτσι τη συστηματική και δίκαιη σύγκρισή τους.

Σε πρώτο στάδιο, στο πλαίσιο του project εφαρμόστηκε στρωματοποιημένος διαχωρισμός (stratified partitioning) των δεδομένων σε υποσύνολα εκπαίδευσης και επικύρωσης. Πιο συγκεκριμένα, το 70% των παρατηρήσεων χρησιμοποιήθηκε ως training partition και το υπόλοιπο 30% ως validation partition, ενώ στην παρούσα φάση δεν δημιουργήθηκε ανεξάρτητο test partition. Η επιλογή της στρωματοποιημένης διάσπασης ήταν ιδιαίτερα σημαντική, επειδή η μεταβλητή-στόχος εμφανίζει ανισόρροπη κατανομή κλάσεων, με τους churners να αποτελούν μειοψηφία στο τελικό δείγμα. Με τον τρόπο αυτό διατηρήθηκε περίπου η ίδια αναλογία churned και non-churned καταστημάτων και στα δύο υποσύνολα, γεγονός που συνέβαλε σε πιο αξιόπιστη αξιολόγηση των μοντέλων.

Project Settings

Partition Data	Partition Data
Event-Based Sampling	☒ Create partition variable
Node Configuration	Note: These settings are active only when a partition variable is not set within the data. Using a data source with a pre-defined partition variable or manually selecting a partition variable will override these settings.
Rules	Method:
Output Library	Stratify
Logging	Training:
Compute Context	70 70.00%
	Validation:
	30 30.00%
	Test:
	0 0.00%

Εικόνα 29 – Project Settings

Σε επίπεδο ρυθμίσεων του Model Comparison node, ως βασικό κριτήριο επιλογής χρησιμοποιήθηκε το Area Under the ROC Curve (AUC ή C-statistic), ενώ το προεπιλεγμένο cut-off ταξινόμησης διατηρήθηκε στο 0,50. Παράλληλα, στο SAS ορίστηκε selection depth ίσο με 10, γεγονός που επιτρέπει την αξιολόγηση των μοντέλων και σε όρους top-decile performance, δηλαδή ως προς το πόσο καλά συγκεντρώνουν churners στο ανώτερο 10% της κατάταξης κινδύνου. Το σημείο αυτό έχει ιδιαίτερη επιχειρησιακή σημασία, καθώς στην πράξη η πλατφόρμα σπάνια θα επέλεγε να παρέμβει σε όλα τα καταστήματα ταυτόχρονα· αντίθετα, ενδιαφέρεται περισσότερο να ιεραρχήσει τους συνεργάτες και να εστιάσει πρώτα στις περιπτώσεις με τη μεγαλύτερη εκτιμώμενη πιθανότητα αποχώρησης. Η επιλογή του AUC ως κύριου criterion metric θεωρείται κατάλληλη για δύο βασικούς λόγους. Πρώτον, το churn prediction διατυπώνεται ως κλασικό πρόβλημα δυαδικής ταξινόμησης, όπου ζητούμενο είναι ο διαχωρισμός των καταστημάτων σε churned και non-churned. Δεύτερον, το ποσοστό της θετικής κλάσης στον κύριο ορισμό του target είναι περίπου 10%, άρα η accuracy από μόνη της δεν επαρκεί για να κρίνει τη χρησιμότητα ενός μοντέλου. Το AUC επιτρέπει τη σύγκριση των

μοντέλων ανεξάρτητα από ένα συγκεκριμένο threshold, αξιολογώντας τη συνολική ικανότητά τους να ιεραρχούν σωστά τις παρατηρήσεις υψηλού και χαμηλού κινδύνου. Παράλληλα, για το τελικό champion model χρησιμοποιήθηκαν και threshold-dependent μετρικές, όπως precision, recall και F1, ώστε η τελική αξιολόγηση να μην περιοριστεί μόνο στη συμπεριφορά της ROC καμπύλης αλλά να συνδεθεί και με πιο πρακτικά επιχειρησιακά σενάρια.

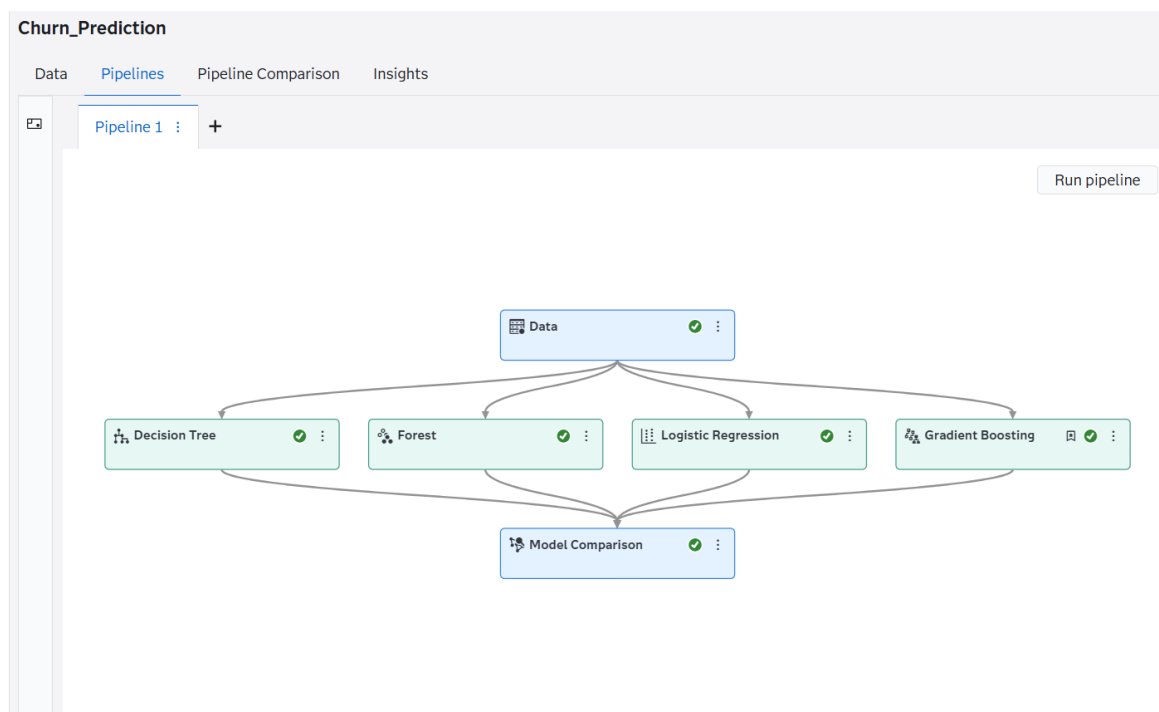
Project Settings

Partition Data	Rules
Event-Based Sampling	Model Comparison
Node Configuration	Class selection statistic:
Rules	Area under curve (C statistic) ▾
Output Library	Interval selection statistic:
Logging	Average squared error ▾
Compute Context	Selection partition:
	Default ▾
	The default selection is Test, then Validate, then Train, based on availability.
	Selection depth:
	10 ▾
	ROC-based cutoff:
	0.50 ▾
	Model
	☐ Override the default classification cutoff

Εικόνα 30 – Επιλογή του AUC ως Κριτηρίου Σύγκρισης Μοντέλων

Μετά τον καθορισμό των παραπάνω ρυθμίσεων, αναπτύχθηκαν στο ίδιο project τέσσερις διαφορετικοί αλγόριθμοι ταξινόμησης: το Logistic Regression, το Decision Tree, το Forest και το Gradient Boosting. Η επιλογή των συγκεκριμένων μοντέλων έγινε με στόχο να εξεταστούν τόσο πιο απλές και ερμηνεύσιμες προσεγγίσεις όσο και πιο σύνθετες ensemble τεχνικές. Η Logistic Regression χρησιμοποιήθηκε ως ένα βασικό και ευρέως χρησιμοποιούμενο μοντέλο αναφοράς για προβλήματα δυαδικής ταξινόμησης, το Decision Tree λόγω της ερμηνευσιμότητάς του, ενώ το Forest και το Gradient Boosting επιλέχθηκαν ως πιο ισχυρές μη γραμμικές προσεγγίσεις με αυξημένη προβλεπτική ικανότητα.

Όλα τα παραπάνω μοντέλα ενσωματώθηκαν σε κοινή pipeline στο SAS Model Studio, ξεκινώντας από τον κόμβο Data και καταλήγοντας στον κόμβο Model Comparison, όπου συγκεντρώθηκαν τα αποτελέσματά τους. Με τον τρόπο αυτό όλα τα μοντέλα αξιολογήθηκαν πάνω στο ίδιο σύνολο δεδομένων, με κοινές ρυθμίσεις partition και κοινό selection criterion, ώστε η τελική σύγκριση να είναι όσο το δυνατόν πιο συνεπής και αξιόπιστη. Η συνολική διάταξη του pipeline παρουσιάζεται στην Εικόνα 31.



Εικόνα 31 – Τελικό Pipeline των Μοντέλων Churn Prediction στο SAS Model Studio

7.2 Αποτελέσματα Πρόβλεψης Αποχώρησης Καταστημάτων

Στην παρούσα ενότητα παρουσιάζονται τα αποτελέσματα της μοντελοποίησης για την πρόβλεψη αποχώρησης καταστημάτων και τεκμηριώνεται η επιλογή του τελικού champion model. Η αξιολόγηση δεν βασίστηκε αποκλειστικά σε μία μόνο μετρική, αλλά σε συνδυασμό δεικτών που αποτυπώνουν τόσο τη συνολική διακριτική ικανότητα των μοντέλων όσο και τη συμπεριφορά τους σε συγκεκριμένα cut-offs. Με άλλα λόγια, η επιλογή του τελικού μοντέλου δεν στηρίχθηκε μόνο στο ποιο μοντέλο εμφάνισε το υψηλότερο validation AUC, αλλά και στο ποιο παρουσίασε την καλύτερη ισορροπία ανάμεσα σε σταθερότητα, γενίκευση και επιχειρησιακή χρησιμότητα.

Η παρουσίαση των αποτελεσμάτων ακολουθεί τέσσερα διακριτά στάδια. Αρχικά, συγκρίνονται τα τέσσερα μοντέλα ταξινόμησης και τεκμηριώνεται η επιλογή του champion model. Στη συνέχεια, εξετάζεται αναλυτικότερα η απόδοση του επιλεγμένου μοντέλου μέσα από confusion matrices και βασικές μετρικές ταξινόμησης. Έπειτα, αναλύονται οι σημαντικότερες μεταβλητές που τροφοδοτούν τις προβλέψεις του μοντέλου, ώστε να προκύψει και η αντίστοιχη ερμηνευτική και επιχειρησιακή αξία. Τέλος, συζητείται ο τρόπος με τον οποίο τα αποτελέσματα μπορούν να αξιοποιηθούν πρακτικά από την πλατφόρμα για τον εντοπισμό και τη διαχείριση συνεργατών αυξημένου κινδύνου.

7.2.1 Σύγκριση μοντέλων και επιλογή champion model

Η αρχική σύγκριση των τεσσάρων εξεταζόμενων μοντέλων πραγματοποιήθηκε στο validation partition, ώστε η αξιολόγηση να βασιστεί σε δεδομένα που δεν χρησιμοποιήθηκαν κατά την εκπαίδευση. Στον Πίνακα της Εικόνας 32 παρουσιάζονται οι βασικότερες validation μετρικές για τα μοντέλα Gradient Boosting, Forest, Logistic Regression και Decision Tree. Εκτός από το validation AUC, συμπεριλαμβάνονται επίσης η accuracy, το F1 score, η τιμή του δείκτη KS, καθώς και δύο top-decile metrics, το cumulative lift και το cumulative response, οι οποίες δείχνουν πόσο αποτελεσματικά κάθε μοντέλο συγκεντρώνει churners στο ανώτερο 10% της κατάταξης κινδύνου.

Μοντέλο	Validation AUC	Accuracy	F1 Score	KS	Cumulative Lift (Top 10%)	Cumulative Response (Top 10%)
Gradient Boosting	0.9078	0.9203	0.4730	0.6594	5.5397	55.3971%
Forest	0.8929	0.9171	0.4499	0.6317	5.5193	55.1935%
Logistic Regression	0.7906	0.8997	0.0000	0.5051	1.7923	17.9226%
Decision Tree	0.7100	0.9169	0.4292	0.4124	4.6192	46.1919%

Εικόνα 32 – Συγκριτική Αξιολόγηση των τεσσάρων Μοντέλων Ταξινόμησης στο Validation Partition

Όπως φαίνεται στην Εικόνα 32, το Gradient Boosting κατέγραψε την καλύτερη συνολική επίδοση, παρουσιάζοντας το υψηλότερο validation AUC (0.9078) μεταξύ των τεσσάρων μοντέλων, γεγονός που το καθιστά το ισχυρότερο μοντέλο ως προς τη συνολική ικανότητα διάκρισης μεταξύ churned και non-churned καταστημάτων. Παράλληλα, εμφάνισε και το υψηλότερο F1 score (0.4730), κάτι που δείχνει ότι στο προεπιλεγμένο threshold πέτυχε την καλύτερη ισορροπία ανάμεσα σε precision και recall. Το Forest κινήθηκε πολύ κοντά στο Gradient Boosting τόσο σε AUC όσο και στις υπόλοιπες μετρικές, γεγονός που επιβεβαιώνει ότι και οι δύο ensemble προσεγγίσεις προσαρμόζονται αποτελεσματικά στη μη γραμμική δομή του προβλήματος. Ωστόσο, η υπεροχή του Gradient Boosting είναι έστω και οριακά σαφής στο validation set, κάτι που αποτέλεσε την πρώτη ένδειξη για την τελική επιλογή του ως champion model.

Η περίπτωση της Logistic Regression έχει ιδιαίτερο μεθοδολογικό ενδιαφέρον. Το validation AUC της (0.7906) ήταν αξιοπρεπές για baseline γραμμικό μοντέλο, ενώ η accuracy άγγιξε το 0.8997. Παρόλα αυτά, το F1 score μηδενίστηκε στο default threshold, γεγονός που σημαίνει ότι το μοντέλο ουσιαστικά δεν παρήγαγε χρήσιμες θετικές προβλέψεις churn στο συγκεκριμένο cut-off. Το σημείο αυτό αναδεικνύει καθαρά γιατί η accuracy δεν μπορεί να χρησιμοποιηθεί ως μοναδικό κριτήριο σε προβλήματα με ανισορροπία κλάσεων: ένα μοντέλο μπορεί να φαίνεται «καλό» επειδή ταξινομεί σωστά τη μεγάλη πλειοψηφία των non-churned καταστημάτων, χωρίς όμως να εντοπίζει αποτελεσματικά τη μικρότερη αλλά επιχειρησιακά κρίσιμη θετική κλάση.

Το Decision Tree, από την άλλη πλευρά, λειτούργησε περισσότερο ως ερμηνεύσιμο baseline παρά ως ισχυρός τελικός υποψήφιος. Παρότι η accuracy του ήταν παρόμοια με εκείνη των υπόλοιπων μοντέλων, το validation AUC περιορίστηκε στο 0.7100 και το KS στο 0.4124, δείχνοντας αισθητά χαμηλότερη διακριτική ικανότητα. Αυτό ήταν αναμενόμενο, καθώς ένα μεμονωμένο δέντρο απόφασης είναι συνήθως λιγότερο ικανό να αποτυπώσει σύνθετες μη γραμμικές αλληλεπιδράσεις σε σχέση με ensemble μεθόδους όπως το Forest και το Gradient Boosting.

Η επιλογή του AUC ως βασικού criterion metric ενισχύει περαιτέρω την αξιοπιστία της σύγκρισης. Σε προβλήματα δυαδικής ταξινόμησης με μη ισορροπημένες κλάσεις, το AUC είναι ιδιαίτερα χρήσιμο επειδή αξιολογεί τη συνολική ικανότητα κατάταξης των παρατηρήσεων σε όλο το εύρος πιθανών thresholds. Έτσι, αποφεύγεται το ενδεχόμενο να ευνοηθεί ένα μοντέλο μόνο και μόνο επειδή αποδίδει καλά σε ένα συγκεκριμένο cut-off. Στην παρούσα εργασία, η χρήση του AUC συμπληρώθηκε από threshold-dependent μετρικές όπως το F1 και το KS, έτσι ώστε η τελική επιλογή να συνδυάζει αφενός γενική προβλεπτική ισχύ και αφετέρου πρακτική απόδοση σε επίπεδο επιχειρησιακής ταξινόμησης. Πέρα από τη σύγκριση στο validation partition, κρίθηκε απαραίτητο να εξεταστεί και η διαφορά ανάμεσα σε train και validation AUC, ώστε να αποτιμηθεί η σταθερότητα και ο βαθμός υπερπροσαρμογής (overfitting) των μοντέλων.

Μοντέλο	Train AUC	Validation AUC	Διαφορά (Train-Validation)
Gradient Boosting	0.9276	0.9078	0.0198
Forest	0.9741	0.8929	0.0812
Logistic Regression	0.8016	0.7906	0.0110
Decision Tree	0.7327	0.7100	0.0227

Εικόνα 33 – Σύγκριση Train και Validation AUC για τον Έλεγχο Γενίκευσης και Υπερπροσαρμογής των Μοντέλων

Η Εικόνα 33 δείχνει ότι το Forest εμφάνισε τη μεγαλύτερη απόσταση μεταξύ εκπαίδευσης και επικύρωσης, με train AUC 0.9741, validation AUC 0.8929 και συνολική διαφορά 0.0812. Η διαφορά αυτή αποτελεί σαφή ένδειξη αυξημένης τάσης υπερπροσαρμογής (overfitting), καθώς το μοντέλο φαίνεται να έχει μάθει πολύ καλά τα πρότυπα του training set, χωρίς όμως να διατηρεί στον ίδιο βαθμό την επίδοσή του σε δεδομένα που δεν έχει “δει”. Παρότι η τελική του επίδοση στο validation partition παραμένει υψηλή, το εύρημα αυτό μειώνει την αξιοπιστία του ως τελικής επιλογής, ειδικά σε σύγκριση με ένα μοντέλο που εμφανίζει καλύτερη ισορροπία μεταξύ απόδοσης και γενίκευσης.

Αντίθετα, το Gradient Boosting παρουσίασε σαφώς πιο ισορροπημένη εικόνα. Συγκεκριμένα, κατέγραψε train AUC 0.9276, validation AUC 0.9078 και διαφορά μόλις 0.0198. Το μικρό αυτό χάσμα δείχνει ότι το μοντέλο όχι μόνο πέτυχε την υψηλότερη επίδοση στο validation set, αλλά ταυτόχρονα διατήρησε και πιο σταθερή συμπεριφορά μεταξύ training και validation partition. Η συγκεκριμένη παρατήρηση είναι ιδιαίτερα σημαντική, διότι υποδηλώνει καλύτερη δυνατότητα γενίκευσης και μικρότερο κίνδυνο υπερπροσαρμογής σε σχέση με το Forest.

Η Logistic Regression εμφάνισε ακόμη μικρότερη διαφορά μεταξύ training και validation (0.0110), γεγονός που υποδηλώνει σταθερή συμπεριφορά. Ωστόσο, η συνολική της επίδοση παρέμεινε σαφώς χαμηλότερη από εκείνη των ensemble μοντέλων. Αντίστοιχα, το Decision Tree παρουσίασε διαφορά 0.0227, αλλά η συνολική του απόδοση ήταν επίσης περιορισμένη. Επομένως, η τελική επιλογή δεν θα μπορούσε να βασιστεί μόνο στη σταθερότητα, αλλά έπρεπε να λάβει υπόψη ταυτόχρονα και την απόλυτη προβλεπτική ικανότητα κάθε μοντέλου.

Συνοψίζοντας, η συγκριτική αξιολόγηση έδειξε ότι το Gradient Boosting αποτελεί το καταλληλότερο μοντέλο για το πρόβλημα της πρόβλεψης churn στο συγκεκριμένο dataset. Το μοντέλο αυτό παρουσίασε το υψηλότερο validation AUC, το υψηλότερο F1 score, πολύ καλή επίδοση στους δείκτες KS και lift, και ταυτόχρονα μικρότερη απόσταση μεταξύ train και validation επίδοσης σε σχέση με το Forest. Με άλλα λόγια, συνδύασε υψηλή προβλεπτική ικανότητα με καλύτερη σταθερότητα και γενίκευση. Για τους λόγους αυτούς επιλέχθηκε ως το τελικό champion model της ανάλυσης.

7.2.2 Αξιολόγηση του champion model

Μετά την επιλογή του Gradient Boosting ως champion model, το επόμενο βήμα ήταν η αναλυτικότερη αξιολόγησή του. Η εξέταση αυτή δεν περιορίστηκε στη συνολική τιμή του AUC, αλλά επεκτάθηκε στους πίνακες ταξινόμησης (confusion matrices) και σε βασικές μετρικές ταξινόμησης, ώστε να αποτυπωθεί με μεγαλύτερη σαφήνεια η συμπεριφορά του μοντέλου σε πρακτικό επίπεδο. Το σημείο αυτό είναι ιδιαίτερα σημαντικό, καθώς σε εφαρμογές churn prediction η αξία ενός μοντέλου δεν εξαρτάται μόνο από το πόσο καλά ιεραρχεί θεωρητικά τις παρατηρήσεις, αλλά και από το πώς συμπεριφέρεται όταν μετατρέπεται σε συγκεκριμένες αποφάσεις τύπου «παρέμβαση / μη παρέμβαση».

Σε επίπεδο συνολικής διακριτικής ικανότητας, το μοντέλο παρουσίασε πολύ υψηλή επίδοση, με validation AUC ίσο με 0.9078. Η τιμή αυτή δείχνει πολύ καλή ικανότητα διαχωρισμού μεταξύ

churned και non-churned καταστημάτων και επιβεβαιώνει ότι το μοντέλο λειτουργεί αποτελεσματικά ως εργαλείο ranking. Στο πλαίσιο της δυαδικής ταξινόμησης, το AUC παραμένει η πιο κατάλληλη κύρια μετρική σύγκρισης, καθώς δεν εξαρτάται από ένα μόνο cut-off και επιτρέπει τη σταθερή αντιπαραβολή διαφορετικών αλγορίθμων. Παρ' όλα αυτά, για την επιχειρησιακή αξιοποίηση του μοντέλου είναι απαραίτητο να εξεταστούν και μετρικές όπως η precision, η recall και το F1 score, οι οποίες συνδέονται άμεσα με το threshold.

Παρακάτω παρουσιάζεται ο πίνακας ταξινόμησης του champion model στο validation partition για το προεπιλεγμένο cut-off 0.50.

Πραγματική κλάση	Πρόβλεψη: Churn	Πρόβλεψη: Non-churn	Σύνολο
Churn (y=1)	175	316	491
Non-churn (y=0)	74	4331	4405
Σύνολο	249	4647	4896

Εικόνα 34 – Confusion Matrix του Champion Model στο Validation Partition για Cut-Off 0.50

Ο πίνακας δείχνει ότι στο default threshold το μοντέλο ταξινόμησε σωστά 175 churners και 4.331 non-churned καταστήματα, ενώ άφησε αταξινομήτους ως churn 316 πραγματικούς churners και παρήγαγε 74 false positives. Με βάση τα παραπάνω, η accuracy του μοντέλου διαμορφώνεται στο 92.03%, η precision στο 70.28%, η recall στο 35.64% και το F1 score στο 47.30%. Παράλληλα, η specificity φθάνει το 98.32%, γεγονός που σημαίνει ότι το μοντέλο στο συγκεκριμένο threshold είναι ιδιαίτερα αποτελεσματικό στο να αναγνωρίζει σωστά τα μη αποχωρήσαντα καταστήματα.

Η εικόνα αυτή δείχνει ότι το champion model, όταν χρησιμοποιείται με cut-off 0.50, είναι σχετικά συντηρητικό. Δηλαδή, όταν προβλέπει ότι ένα κατάστημα είναι churner, έχει αρκετά μεγάλη πιθανότητα να είναι σωστό, όμως ταυτόχρονα αφήνει εκτός σημαντικό μέρος των πραγματικών churners. Η συμπεριφορά αυτή είναι απολύτως αναμενόμενη σε προβλήματα όπου η θετική κλάση είναι μειοψηφική. Ένα πιο αυστηρό threshold τείνει να προστατεύει από την υπερβολική παραγωγή false positives, αλλά πληρώνει το τίμημα της χαμηλότερης recall. Από ακαδημαϊκή σκοπιά, το αποτέλεσμα αυτό είναι αποδεκτό και ερμηνεύσιμο. Από επιχειρησιακή σκοπιά, όμως, χρειάζεται να εξεταστεί αν η πλατφόρμα προτιμά μεγαλύτερη ακρίβεια στις παρεμβάσεις ή μεγαλύτερη κάλυψη των churners.

Για να αποτιμηθεί η επίδραση της επιλογής threshold, εξετάστηκε και το εναλλακτικό cut-off που προέκυψε βάσει του δείκτη KS, το οποίο στο validation set ήταν περίπου 0.08. Ο παρακάτω πίνακας συγκρίνει συνοπτικά τις δύο επιλογές threshold.

Threshold	TP	FP	Precision	Recall	Specificity	F1	Accuracy
0.50 (Default)	175	74	70.28%	35.64%	98.32%	47.30%	92.03%
0.08 (KS)	428	935	31.40%	87.17%	78.77%	46.17%	79.62%

Εικόνα 35 – Σύγκριση του Champion Model για Διαφορετικά Thresholds στο Validation Partition

Η σύγκριση των δύο thresholds δείχνει καθαρά το trade-off ανάμεσα στην precision και την recall. Με το cut-off 0.08, το μοντέλο εντοπίζει πολύ μεγαλύτερο μέρος των πραγματικών churners, ανεβάζοντας την recall στο 87.17%. Με άλλα λόγια, σχεδόν εννέα στους δέκα churners του validation set εντοπίζονται. Ωστόσο, αυτή η βελτίωση συνοδεύεται από σημαντική αύξηση των false positives, γεγονός που οδηγεί σε αισθητή πτώση της precision. Άρα, στο χαμηλότερο threshold το μοντέλο γίνεται σαφώς πιο επιθετικό στον εντοπισμό churners, αλλά με μεγαλύτερο κόστος σε λανθασμένες παρεμβάσεις.

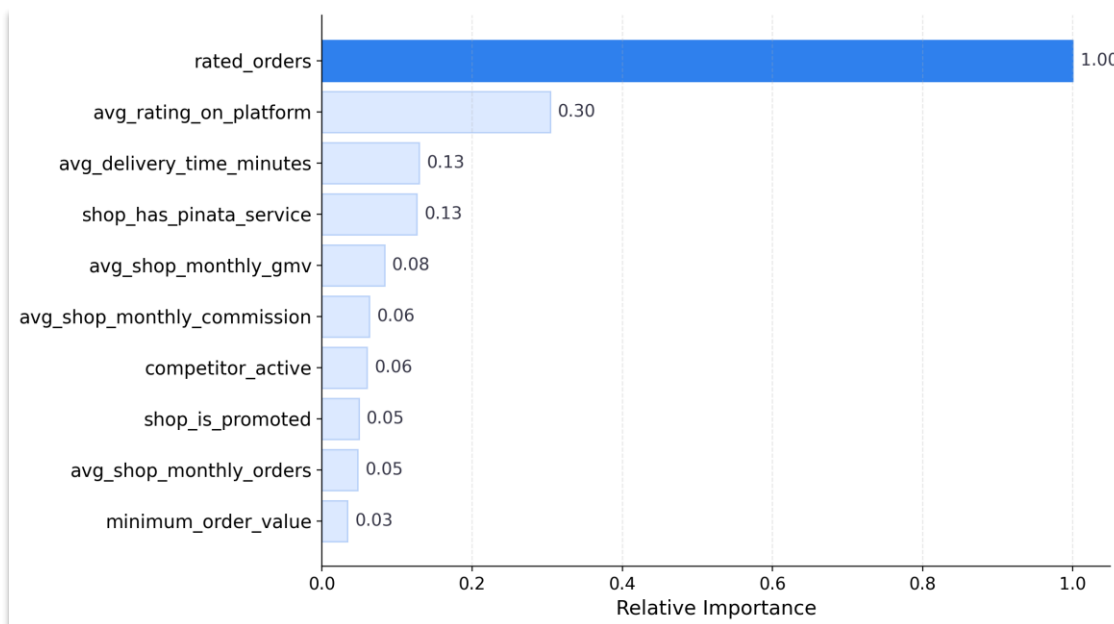
Από την άλλη πλευρά, στο default threshold 0.50 το μοντέλο είναι πιο επιλεκτικό. Εντοπίζει μικρότερο ποσοστό churners, αλλά όταν προβλέπει ότι ένα κατάστημα ανήκει στην κλάση churn, η πρόβλεψη αυτή είναι πολύ πιο αξιόπιστη. Αυτό φαίνεται τόσο από την υψηλότερη precision όσο και από τη σημαντικά υψηλότερη specificity. Το γεγονός ότι το F1 score παραμένει πολύ κοντά ανάμεσα στα δύο thresholds δείχνει ότι η αλλαγή του cut-off δεν βελτιώνει συνολικά την ισορροπία precision–recall, αλλά μετατοπίζει τη συμπεριφορά του μοντέλου κυρίως προς την πλευρά της recall.

Επομένως, η επιλογή threshold δεν είναι απλώς τεχνικό ζήτημα, αλλά συνδέεται άμεσα με τον τρόπο που η πλατφόρμα θέλει να αξιοποιήσει επιχειρησιακά το μοντέλο. Αν ο στόχος είναι η όσο το δυνατόν μεγαλύτερη κάλυψη πιθανών churners, τότε το KS-based cut-off 0.08 φαίνεται πιο κατάλληλο. Αντίθετα, αν η πλατφόρμα προτιμά πιο στοχευμένες και οικονομικά αποδοτικές παρεμβάσεις, το default threshold 0.50 αποτελεί πιο ασφαλή επιλογή, καθώς περιορίζει σημαντικά τις ψευδώς θετικές προβλέψεις.

Στην πράξη, οι δύο επιλογές μπορούν να αντιστοιχηθούν σε διαφορετικά σενάρια χρήσης. Όταν οι ενέργειες retention είναι χαμηλού κόστους, όπως αυτοματοποιημένα μηνύματα ή ήπιες προωθητικές ενέργειες, ένα χαμηλότερο threshold έχει περισσότερο νόημα, επειδή προτεραιότητα είναι να μη χαθούν πραγματικοί churners. Αντίθετα, όταν οι ενέργειες διατήρησης είναι πιο δαπανηρές, όπως προσωπική επικοινωνία από account managers ή εξατομικευμένες εμπορικές παρεμβάσεις, ένα πιο συντηρητικό threshold είναι συνήθως προτιμότερο, γιατί επιτρέπει καλύτερη κατανομή των διαθέσιμων πόρων.

7.2.3 Σημαντικότερες μεταβλητές πρόβλεψης churn

Η επιλογή του Gradient Boosting ως champion model αποκτά μεγαλύτερη αξία όταν συνοδεύεται και από ερμηνεία των βασικών παραγόντων που τροφοδοτούν τις προβλέψεις του. Για τον σκοπό αυτό εξετάστηκε το variable importance του μοντέλου, δηλαδή η σχετική σημαντικότητα των μεταβλητών ως προς τη συμβολή τους στην τελική πρόβλεψη churn. Η ανάλυση αυτή δεν μετατρέπει το μοντέλο σε πλήρως «διαφανές», προσφέρει όμως μια αρκετά χρήσιμη εικόνα για το ποιες επιχειρησιακές διαστάσεις φαίνεται να σχετίζονται περισσότερο με τον κίνδυνο αποχώρησης. Η συνολική κατάσταση των μεταβλητών αποτυπώνεται στην Εικόνα 36.



Εικόνα 36 – Σχετική Σημαντικότητα των Μεταβλητών στο Champion Model

Η πρώτη και πιο έντονη παρατήρηση από την Εικόνα 36 είναι η σαφής κυριαρχία της μεταβλητής `rated_orders`. Η μεταβλητή αυτή υπερέχει αισθητά όλων των υπολοίπων, γεγονός που δείχνει ότι ο αριθμός αξιολογημένων παραγγελιών ενός καταστήματος λειτουργεί ως πολύ ισχυρός δείκτης της συνολικής του παρουσίας και δραστηριότητας στην πλατφόρμα. Επιχειρησιακά, το εύρημα αυτό είναι λογικό, καθώς ένα κατάστημα με περιορισμένη ένταση παραγγελιών και χαμηλότερη εμπλοκή στο οικοσύστημα της πλατφόρμας είναι πιθανότερο να βρίσκεται σε πορεία αποδυνάμωσης ή αποχώρησης. Η μεταβλητή `rated_orders`, επομένως, λειτουργεί ως ένας ισχυρός δείκτης «ζωντανίας» της συνεργασίας.

Αμέσως μετά ακολουθεί η μεταβλητή `avg_rating_on_platform`, γεγονός που αναδεικνύει τον ρόλο της ποιότητας εξυπηρέτησης και της εμπειρίας πελάτη. Η υψηλή σχετική σημαντικότητα της συγκεκριμένης μεταβλητής υποδηλώνει ότι το `churn` δεν είναι απλώς θέμα όγκου παραγγελιών ή αδράνειας, αλλά συνδέεται και με το πώς βιώνουν οι χρήστες το κατάστημα μέσα στην πλατφόρμα. Χαμηλότερη αξιολόγηση μπορεί να συνδέεται με αδύναμη εξυπηρέτηση, με χαμηλότερη ικανοποίηση χρηστών και τελικά με μειωμένη βιωσιμότητα της συνεργασίας. Στον ίδιο άξονα κινείται και η μεταβλητή `avg_delivery_time_minutes`, η οποία αναδεικνύεται ως μία από τις τρεις ισχυρότερες μεταβλητές. Η λειτουργική απόδοση, και ειδικά η ταχύτητα εξυπηρέτησης, φαίνεται επομένως να αποτελεί ουσιαστικό παράγοντα που σχετίζεται με τη σταθερότητα της παρουσίας ενός καταστήματος στην πλατφόρμα.

Στις σημαντικότερες μεταβλητές εμφανίζονται επίσης η `shop_has_pinata_service` και το `avg_shop_monthly_gmv`. Η παρουσία της πρώτης υποδηλώνει ότι ο βαθμός ενσωμάτωσης ενός καταστήματος σε επιμέρους υπηρεσίες της πλατφόρμας μπορεί να σχετίζεται με την πιθανότητα αποχώρησης, ενώ η δεύτερη δείχνει ότι η οικονομική και εμπορική επίδοση του καταστήματος παραμένει κεντρικός παράγοντας για τη συνολική του σταθερότητα. Με άλλα λόγια, το μοντέλο δεν αξιοποιεί μόνο δείκτες δραστηριότητας ή εμπειρίας πελάτη, αλλά και μεταβλητές που αποτυπώνουν τη συνολική εμπορική δυναμική του συνεργάτη.

Οι υπόλοιπες `input` μεταβλητές συμπληρώνουν την εικόνα και ενισχύουν συγκεκριμένους επιχειρησιακούς άξονες. Μεταβλητές όπως οι `avg_shop_monthly_commission`, `avg_shop_monthly_orders` και `minimum_order_value` αποτυπώνουν περισσότερο την οικονομική και εμπορική πλευρά της συνεργασίας.

Για να μεταφραστεί η καθαρά στατιστική κατάταξη των μεταβλητών σε πιο ουσιαστικούς επιχειρησιακούς όρους, οι σημαντικότερες μεταβλητές ομαδοποιήθηκαν σε βασικούς άξονες ερμηνείας. Η ομαδοποίηση αυτή παρουσιάζεται στον πίνακα της Εικόνας 37.

Επιχειρησιακός άξονας	Σημαντικές μεταβλητές	Ερμηνεία
Ένταση δραστηριότητας / συμμετοχής	rated_orders, avg_shop_monthly_orders	Αποτυπώνεται το επίπεδο ενεργής συμμετοχής του καταστήματος στο οικοσύστημα της πλατφόρμας και η ισχύς της βάσης ζήτησης.
Ποιότητα και λειτουργική επίδοση	avg_rating_on_platform, avg_delivery_time_minutes	Αντανακλάται η εμπειρία του πελάτη και η λειτουργική συνέπεια του καταστήματος.
Οικονομική και εμπορική δυναμική	avg_shop_monthly_gmv, avg_shop_monthly_commission, minimum_order_value	Αποτυπώνεται το οικονομικό αποτύπωμα και η εμπορική βιωσιμότητα της συνεργασίας.
Βαθμός ενσωμάτωσης στην πλατφόρμα	shop_has_pinata_service, shop_is_promoted	Αποτυπώνεται ο βαθμός ενσωμάτωσης του καταστήματος σε υπηρεσίες, εργαλεία και εμπορικές λειτουργίες της πλατφόρμας..
Ανταγωνιστικό και δομικό περιβάλλον	competitor_active	Εκφράζονται οι εξωτερικές πιέσεις και τα οργανωτικά χαρακτηριστικά που επηρεάζουν τη σταθερότητα της συνεργασίας.

Εικόνα 37 – Ομαδοποίηση των Σημαντικότερων Μεταβλητών του Champion Model σε Βασικούς Επιχειρησιακούς Άξονες

Η Εικόνα 37 βοηθά να αποδοθεί η στατιστική σημασία των μεταβλητών σε πιο κατανοητούς επιχειρησιακούς όρους. Συνολικά, το champion model δείχνει ότι το churn των καταστημάτων δεν εξαρτάται από έναν μόνο παράγοντα, αλλά από έναν συνδυασμό στοιχείων που σχετίζονται με τη δραστηριότητα του καταστήματος, την ποιότητα της υπηρεσίας, την οικονομική του εικόνα, τον βαθμό ενσωμάτωσής του στην πλατφόρμα και το ανταγωνιστικό περιβάλλον. Με άλλα λόγια, τα καταστήματα με υψηλότερο κίνδυνο αποχώρησης δεν είναι απλώς εκείνα με τις λιγότερες παραγγελίες, αλλά εκείνα που εμφανίζουν συνολικά πιο αδύναμη παρουσία και μικρότερη σταθερότητα στη συνεργασία τους με την πλατφόρμα. Παράλληλα, η ερμηνεία της σχετικής σημαντικότητας των μεταβλητών πρέπει να γίνεται με προσοχή, καθώς δείχνει ποιοι παράγοντες συμβάλλουν περισσότερο στην πρόβλεψη, χωρίς όμως να αποδεικνύει από μόνη της αιτιακή σχέση. Παρ' όλα αυτά, η συνολική εικόνα του μοντέλου είναι αρκετά σαφής και προσφέρει μια χρήσιμη βάση για την κατανόηση των βασικών διαστάσεων που συνδέονται με το churn.

7.2.4 Επιχειρησιακή αξιοποίηση των αποτελεσμάτων

Η αξία του μοντέλου πρόβλεψης αποχώρησης δεν περιορίζεται μόνο στην τεχνική αξιολόγηση της απόδοσής του, αλλά επεκτείνεται κυρίως στην πρακτική του αξιοποίηση από την πλατφόρμα. Στο συγκεκριμένο πλαίσιο, το Gradient Boosting μπορεί να χρησιμοποιηθεί ως εργαλείο έγκαιρου εντοπισμού καταστημάτων αυξημένου κινδύνου αποχώρησης, επιτρέποντας πιο στοχευμένη και οργανωμένη διαχείριση του δικτύου συνεργατών. Αντί η πλατφόρμα να αντιμετωπίζει όλους τους συνεργάτες με τον ίδιο τρόπο, μπορεί να αξιοποιήσει τον δείκτη κινδύνου αποχώρησης για να ιεραρχεί τα καταστήματα από τα πιο ασφαλή προς τα πιο επισφαλή και να κατευθύνει ανάλογα τις ενέργειες διατήρησής.

Με βάση αυτή τη λογική, το μοντέλο μπορεί να χρησιμοποιηθεί όχι μόνο ως μηχανισμός δυαδικής πρόβλεψης, αλλά και ως εργαλείο κατάταξης κινδύνου. Η πληροφορία αυτή είναι ιδιαίτερα χρήσιμη, επειδή στην πράξη η πλατφόρμα δεν χρειάζεται πάντα να μετατρέπει την πρόβλεψη σε απόλυτη απόφαση τύπου «θα αποχωρήσει / δεν θα αποχωρήσει». Σε πολλές περιπτώσεις αρκεί να

γνωρίζει ποια καταστήματα εμφανίζουν υψηλότερη πιθανότητα αποχώρησης, ώστε να δίνει προτεραιότητα στις αντίστοιχες παρεμβάσεις. Με αυτόν τον τρόπο, οι διαθέσιμοι πόροι μπορούν να κατευθυνθούν πιο στοχευμένα, χωρίς να απαιτείται ενιαία αντιμετώπιση όλων των συνεργατών.

Η παραπάνω λογική μπορεί να αποτυπωθεί και σε ένα απλό επιχειρησιακό σχήμα διαβάθμισης κινδύνου. Ένα ενδεικτικό παράδειγμα παρουσιάζεται στην Εικόνα 38.

Ζώνη κινδύνου	Ενδεικτική περιγραφή	Πιθανές ενέργειες διαχείρισης
Υψηλός κίνδυνος	Καταστήματα με πολύ υψηλό predicted churn risk ή θετική ταξινόμηση στο πιο «ευαίσθητο» threshold.	Στοχευμένη επικοινωνία από account managers, διερεύνηση αιτίων αδράνειας, προσωποποιημένες εμπορικές ή λειτουργικές παρεμβάσεις.
Μεσαίος κίνδυνος	Καταστήματα με αυξημένο αλλά όχι ακραίο risk score, τα οποία ενδέχεται να βρίσκονται σε φάση σταδιακής υποχώρησης.	Monitoring, ήπιες ενέργειες retention, βελτιστοποίηση προβολής ή συμμετοχής σε προωθητικές ενέργειες, επαναξιολόγηση σε επόμενο κύκλο scoring.
Χαμηλός κίνδυνος	Καταστήματα με σταθερή παρουσία και χαμηλή πιθανότητα churn.	Περιοδική παρακολούθηση, χωρίς ανάγκη εντατικών ενεργειών διατήρησης, εκτός αν εμφανιστούν νέα σήματα επιδείνωσης.

Εικόνα 38 – Ενδεικτική Επιχειρησιακή Αξιοποίηση του Churn Score ανά Ζώνη Κινδύνου

Ο πίνακας αυτός δείχνει έναν πρακτικό τρόπο με τον οποίο η πλατφόρμα μπορεί να αξιοποιήσει το churn score. Στην πράξη, τα καταστήματα μπορούν να χωριστούν σε ζώνες κινδύνου και να αντιμετωπιστούν με διαφορετικό τρόπο, ανάλογα με το πόσο πιθανό είναι να αποχωρήσουν. Έτσι, το μοντέλο δεν λειτουργεί μόνο ως εργαλείο πρόβλεψης, αλλά και ως μέσο ιεράρχησης προτεραιοτήτων, ώστε η πλατφόρμα να δίνει περισσότερη προσοχή στα καταστήματα που φαίνεται να τη χρειάζονται περισσότερο. Παράλληλα, η τελική επιλογή του threshold δεν είναι μόνο τεχνικό θέμα, αλλά εξαρτάται και από το κόστος των ενεργειών διατήρησης, καθώς και από το πόσο ευρεία κάλυψη θέλει να πετύχει η πλατφόρμα. Επιπλέον, το churn score μπορεί να συνδυαστεί με τα αποτελέσματα του shop clustering, ώστε να υπάρχει πιο ολοκληρωμένη εικόνα όχι μόνο για το ποια καταστήματα κινδυνεύουν περισσότερο, αλλά και για το τι είδους καταστήματα είναι αυτά. Με αυτόν τον τρόπο, το churn prediction μπορεί να αποτελέσει μια χρήσιμη βάση για πιο στοχευμένη και ουσιαστική διαχείριση του δικτύου συνεργατών.

8. Συμπεράσματα – Προτάσεις

Η παρούσα μεταπτυχιακή διατριβή είχε ως βασικό στόχο την αξιοποίηση τεχνικών ανάλυσης δεδομένων και μηχανικής μάθησης για την κατανόηση της λειτουργίας και της απόδοσης των συνεργαζόμενων καταστημάτων σε μια ηλεκτρονική πλατφόρμα διανομής φαγητού. Η εργασία οργανώθηκε σε δύο βασικούς άξονες: την τμηματοποίηση καταστημάτων μέσω τεχνικών clustering και την πρόβλεψη αποχώρησης καταστημάτων μέσω επιβλεπόμενων μοντέλων ταξινόμησης. Μέσα από αυτή τη διπλή προσέγγιση διαμορφώθηκε ένα πρακτικό πλαίσιο shop analytics, το οποίο μπορεί να στηρίξει αποφάσεις που αφορούν τη διαχείριση συνεργατών, τη βελτίωση της λειτουργικής επίδοσης και τη διατήρηση συνεργασιών.

Η διερευνητική ανάλυση έδειξε ότι το δίκτυο των συνεργαζόμενων καταστημάτων χαρακτηρίζεται από έντονη ετερογένεια. Οι διαφοροποιήσεις ως προς τη γεωγραφική κατανομή, την κατηγορία κουζίνας, τον χρόνο παράδοσης, τις αξιολογήσεις και την οικονομική επίδοση επιβεβαιώνουν ότι τα καταστήματα δεν αποτελούν έναν ομοιογενή πληθυσμό. Το εύρημα αυτό δικαιολογεί την ανάγκη αξιοποίησης αναλυτικών μεθόδων, καθώς η ενιαία ή χειροκίνητη διαχείριση χιλιάδων συνεργατών δεν είναι ρεαλιστική ούτε αποτελεσματική σε επιχειρησιακό επίπεδο.

Η ανάλυση clustering ανέδειξε πέντε διακριτά και επιχειρησιακά ερμηνεύσιμα προφίλ καταστημάτων, τα οποία διαφοροποιούνται ως προς τον όγκο δραστηριότητας, τη λειτουργική απόδοση και την ωριμότητα παρουσίας στην πλατφόρμα. Το αποτέλεσμα αυτό επιβεβαιώνει ότι διαφορετικές κατηγορίες συνεργατών χρειάζονται διαφορετικό τρόπο διαχείρισης. Καταστήματα υψηλής ζήτησης αλλά λειτουργικά επιβαρυμένα δεν μπορούν να υποστηριχθούν με τον ίδιο τρόπο όπως νέα καταστήματα χαμηλής δραστηριότητας ή ώριμα καταστήματα υψηλής απόδοσης. Συνεπώς, η τμηματοποίηση λειτουργεί ως βάση για πιο στοχευμένες στρατηγικές συνεργασίας και υποστήριξης.

Στο δεύτερο μέρος της εργασίας, η πρόβλεψη αποχώρησης των καταστημάτων προσεγγίστηκε ως πρόβλημα δυαδικής ταξινόμησης. Επειδή στο διαθέσιμο σύνολο δεδομένων δεν υπήρχε ρητή ένδειξη τερματισμού συνεργασίας, ο ορισμός του churn βασίστηκε σε proxy αδράνειας, με χρήση της μεταβλητής `days_inactive` και προσέγγιση βασισμένη στα ίδια τα δεδομένα για τον καθορισμό του τελικού `threshold`. Η επιλογή αυτή ήταν μεθοδολογικά αναγκαία και επέτρεψε τη δημιουργία μιας τεκμηριωμένης μεταβλητής-στόχου πάνω στην οποία αναπτύχθηκαν και συγκρίθηκαν τα μοντέλα πρόβλεψης.

Η συγκριτική αξιολόγηση των μοντέλων έδειξε ότι το Gradient Boosting αποτέλεσε την ισχυρότερη προσέγγιση για το συγκεκριμένο πρόβλημα, καθώς πέτυχε την καλύτερη επίδοση στο `validation partition` και εμφάνισε μικρότερη ένδειξη υπερπροσαρμογής σε σχέση με το Forest. Παράλληλα, η ανάλυση της σημαντικότητας των μεταβλητών ανέδειξε ότι ο κίνδυνος churn δεν εξαρτάται από έναν μόνο δείκτη, αλλά συνδέεται με ένα πολυπαραγοντικό σύνολο χαρακτηριστικών που περιλαμβάνει την ένταση δραστηριότητας, την ποιότητα εξυπηρέτησης, τη λειτουργική επίδοση, την οικονομική δυναμική και τον βαθμό ενσωμάτωσης του καταστήματος στην πλατφόρμα.

Ένα από τα βασικά συμπεράσματα της εργασίας είναι ότι το shop clustering και το churn prediction λειτουργούν συμπληρωματικά. Η τμηματοποίηση απαντά στο ερώτημα τι είδους κατάσταση είναι ο συνεργάτης, ενώ η πρόβλεψη churn απαντά στο ποιο κατάσταση εμφανίζει αυξημένη πιθανότητα αποχώρησης. Ο συνδυασμός των δύο δίνει στην πλατφόρμα τη δυνατότητα να περάσει από μια γενική λογική διαχείρισης συνεργατών σε μια περισσότερο στοχευμένη και προληπτική προσέγγιση.

Παρά τη συνολικά θετική εικόνα των αποτελεσμάτων, η εργασία παρουσιάζει και ορισμένους περιορισμούς. Ο σημαντικότερος είναι ότι το churn ορίστηκε μέσω συμπεριφορικού proxy και όχι μέσω πραγματικής ένδειξης τερματισμού συνεργασίας. Επιπλέον, η ανάλυση πραγματοποιήθηκε σε συγκεντρωτικό επίπεδο καταστήματος και όχι σε αναλυτικά χρονικά δεδομένα παραγγελιών, ενώ η τελική αξιολόγηση των μοντέλων βασίστηκε σε `training` και `validation partitions` χωρίς ανεξάρτητο `test set`. Οι περιορισμοί αυτοί δεν αναιρούν την αξία της εργασίας, αλλά πρέπει να λαμβάνονται υπόψη στην ερμηνεία των ευρημάτων.

Με βάση τα παραπάνω, προκύπτουν ορισμένες βασικές προτάσεις για την πρακτική αξιοποίηση και τη μελλοντική επέκταση της ανάλυσης:

- Η πλατφόρμα μπορεί να αξιοποιήσει τα αποτελέσματα του clustering για διαφοροποιημένη διαχείριση συνεργατών. Τα καταστήματα δεν εμφανίζουν κοινό προφίλ και, επομένως, δεν είναι σκόπιμο να αντιμετωπίζονται με ενιαίες πολιτικές. Καταστήματα υψηλής ζήτησης αλλά λειτουργικά επιβαρυμένα χρειάζονται διαφορετικό τρόπο υποστήριξης από νέα ή χαμηλής δραστηριότητας καταστήματα, ενώ τα ώριμα και σταθερά καταστήματα απαιτούν περισσότερο δράσεις διατήρησης και ενίσχυσης της συνεργασίας.
- Το churn score μπορεί να λειτουργήσει ως μηχανισμός έγκαιρης προειδοποίησης (early warning system). Αντί να χρησιμοποιείται μόνο ως τεχνική πρόβλεψη, μπορεί να ενσωματωθεί σε διαδικασίες προτεραιοποίησης, ώστε τα καταστήματα υψηλού κινδύνου να εντοπίζονται έγκαιρα και να λαμβάνουν πιο στοχευμένη υποστήριξη ή εμπορική διαχείριση. Ανάλογα με το κόστος των ενεργειών διατήρησης, η πλατφόρμα μπορεί να επιλέγει πιο συντηρητικά ή πιο ευαίσθητα `thresholds`.

- Ιδιαίτερη έμφαση θα πρέπει να δοθεί στη βελτίωση λειτουργικών και ποιοτικών δεικτών. Η ανάλυση έδειξε ότι μεταβλητές όπως οι αξιολογήσεις και ο χρόνος παράδοσης σχετίζονται έντονα με την πιθανότητα churn. Συνεπώς, ενέργειες που βελτιώνουν την εμπειρία του τελικού χρήστη, την ταχύτητα εξυπηρέτησης, την ακρίβεια των παραγγελιών και τη συνολική αξιοπιστία του καταστήματος μπορούν να λειτουργήσουν προληπτικά ως προς την αποχώρηση.
- Σε μελλοντική φάση, η ανάλυση θα μπορούσε να ενισχυθεί με πιο πλούσια και χρονικά αναλυτικά δεδομένα. Η ύπαρξη πραγματικής ένδειξης τερματισμού συνεργασίας, ιστορικού παραγγελιών σε διαδοχικές χρονικές περιόδους, μεταβλητών ακυρώσεων, χρόνου προετοιμασίας, αποδοχής παραγγελιών ή στοιχείων προβολής στην πλατφόρμα θα επέτρεπε ακριβέστερο ορισμό του churn και ισχυρότερα μοντέλα πρόβλεψης.

Συνολικά, η παρούσα διατριβή δείχνει ότι η ανάλυση δεδομένων μπορεί να αποτελέσει ουσιαστικό εργαλείο επιχειρησιακής υποστήριξης για πλατφόρμες online food delivery. Η τμηματοποίηση καταστημάτων και η πρόβλεψη αποχώρησης δεν αποτελούν δύο ανεξάρτητες τεχνικές, αλλά δύο συμπληρωματικούς μηχανισμούς κατανόησης και διαχείρισης του δικτύου συνεργατών. Μέσα από τα ευρήματα της εργασίας αναδεικνύεται ότι μια πιο αναλυτική, στοχευμένη και προληπτική διαχείριση συνεργατών μπορεί να ενισχύσει τόσο τη λειτουργική αποτελεσματικότητα όσο και τη συνολική ανταγωνιστικότητα της πλατφόρμας.

Βιβλιογραφία

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group Research Note.

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. Harvard Business Review, 90(10), 60–68.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137–144.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19, 171–209.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. MIS Quarterly, 36(4), 1165–1188.

Davenport, T. H., & Harris, J. G. (2007). Competing on Analytics: The New Science of Winning. Harvard Business School Press.

Sharda, R., Delen, D., & Turban, E. (2017). Business Intelligence, Analytics, and Data Science: A Managerial Perspective. Pearson Education.

Wixom, B. H., Watson, H. J., & Werner, T. (2011). Developing an enterprise business intelligence capability: The Norfolk Southern Journey. MIS Quarterly Executive, 10(2), 61–71.

Baecker, J., Böttcher, T. P., & Weking, J. (2021). How Companies Create Value From Data – A Taxonomy on Data, Approaches, and Resulting Business Value. ECIS 2021 Research Papers, 124.

Kshetri, N. (2014). Big Data's impact on privacy, security and consumer welfare. Telecommunications Policy, 38, 1134–1145.

Sakr, S. (2016). Introduction. In Big Data 2.0 Processing Systems: A Survey. Springer International Publishing.

Vassakis, K., Petrakis, E., & Kopanakis, I. (2018). Big data analytics: applications, prospects and challenges. In Mobile Big Data. Springer, Cham.

Gartner. (n.d.). Gartner IT Glossary: Big Data. Gartner. Available at: <https://www.gartner.com/en/information-technology/glossary>

SAP Greece. (n.d.). Τι είναι τα Μεγάλα Δεδομένα; | Ορισμός, Πώς Λειτουργεί, και Χρήσεις. SAP. Available at: <https://www.sap.com/greece/products/technology-platform/what-is-big-data.html>

Badman, A., & Kosinski, M. (n.d.). What Is Big Data? IBM Think. Available at: <https://www.ibm.com/think/topics/big-data>

Doken, S. (2025, January 28). The Technology Behind Modern Food Delivery Platforms. Adeia Blog. Available at: <https://adeia.com/blog/the-technology-behind-modern-food-delivery-platforms>

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B, 63(2), 411–423.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236–244.

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD 2013)*. Springer.

Khan, K., Rehman, S. U., Aziz, K., Fong, S., & Sarasvady, S. (2014). DBSCAN: Past, present and future. In *The 5th International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)* (pp. 232–237). IEEE.

Tran, T. N., Drab, K., & Daszykowski, M. (2013). Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemometrics and Intelligent Laboratory Systems*, 120, 92–96.

Kanagala, H. K., & Krishnaiah, V. J. R. (2016). A comparative study of K-Means, DBSCAN and OPTICS. In *2016 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1–6). IEEE.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On Spectral Clustering: Analysis and an Algorithm. In *Advances in Neural Information Processing Systems (NIPS)*.

Yadav, J., & Sharma, M. (2013). A Review of K-mean Algorithm. *International Journal of Engineering Trends and Technology*, 4(7), 2972–2976.

Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., & Song, A. (2015). Efficient agglomerative hierarchical clustering. *Expert Systems with Applications*, 42(5), 2785–2797.

Sasirekha, K., & Baby, P. (2013). Agglomerative Hierarchical Clustering Algorithm – A Review. *International Journal of Scientific and Research Publications*, 3(3).

Agarwal, K., Jain, P., & Rajnayak, M. (2019). Comparative Analysis of Store Clustering Techniques in the Retail Industry. In *Proceedings of the 8th International Conference on Data Science, Technology and Applications (DATA 2019)* (pp. 65–73). SciTePress.

Prajapati, J., Maurya, V., & Chaudhuri, B. (2021). Advanced Retail Store Feature-based Clustering for the Large Multi-brand Retailer. *MRSI 2021 Conference Paper* (Unique ID: 202120024).

Ünal, A., Önal, M., Kaya, T., & Özcan, T. (2022). Store Segmentation in Retail Industry Using Clustering Algorithms. In *Intelligent and Fuzzy Systems – Digital Acceleration and The New Normal: Proceedings of the INFUS 2022 Conference, Volume 2 (Lecture Notes in Networks and Systems, Vol. 505, pp. 409–416)*. Springer.

Bilgiç, E., Kantardzic, M., & Cakir, O. (2015). Retail Store Segmentation for Target Marketing. In P. Perner (Ed.), *Advances in Data Mining: Applications and Theoretical Aspects (ICDM 2015)*, LNCS 9165, 32–44. Springer.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2016).

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. In Advances in Neural Information Processing Systems.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273–297.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Advances in Large Margin Classifiers (pp. 61–74). MIT Press.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning (ICML 2005).

Kumar, S., & Chandrakala, D. (2016). A Survey on Customer Churn Prediction using Machine Learning Techniques. International Journal of Computer Applications, 154(10).

Sweidan, D., Johansson, U., Gidenstam, A., & Alenljung, B. (2022). Predicting Customer Churn in Retailing. In ICMLA 2022 (pp. 635–640).

Sari, R. P., Febriyanto, F., & Adi, A. C. (2023). Analysis Implementation of the Ensemble Algorithm in Predicting Customer Churn in Telco Data: A Comparative Study. Informatica, 47(7), 95–108.

Imani, M., Joudaki, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Customer Churn Prediction: A Systematic Review of Recent Advances, Trends, and Challenges in Machine Learning and Deep Learning. Machine Learning and Knowledge Extraction, 7(3), 105.

SAS Institute Inc. (n.d.). The Machine Learning Landscape: A Quick Guide to the Different Types of Learning. SAS White Paper. Available at: <https://www.sas.com/en/whitepapers/machine-learning-landscape-109909.html>