ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΣΤΗ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ

Διπλωματική Εργασία

ΠΟΣΟΤΙΚΗ ΠΡΟΣΕΓΓΙΣΗ ΣΤΗ ΔΙΕΡΕΥΝΗΣΗ ΤΗΣ ΑΠΩΛΕΙΑΣ B2B ΠΕΛΑΤΩΝ ΣΤΗ ΒΙΟΜΗΧΑΝΙΑ ΜΗΧΑΝΟΛΟΓΙΚΩΝ ΚΑΤΑΣΚΕΥΩΝ

του:

ΚΩΝΣΤΑΝΤΙΝΟΥ Α. ΒΑΣΣΟΥ

Αριθμός Μητρώου: mbx22022

Επιβλέπων: ΑΝΔΡΕΑΣ ΓΕΩΡΓΙΟΥ, ΚΑΘΗΓΗΤΗΣ

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος στη Διοίκηση Επιχειρήσεων

Ιανουάριος 2024

Εγνατία 156, 54636 Θεσσαλονίκη

Τηλ.: 2310 891530 https://www.uom.gr/mba e-mail : mba@uom.edu.gr

Ευχαριστίες

Θα ήθελα να ευχαριστήσω την οικογένειά μου Λένα και Αλέξανδρο για τη στήριξη και την υπομονή τους κατά τη διάρκεια του μεταπτυχιακού προγράμματος, δε θα τα κατάφερνα χωρίς εσάς! Να ευχαριστήσω πολύ και τον επιβλέποντα καθηγητή της εργασίας, Καθηγητή κ. Γεωργίου Ανδρέα για τη βοήθεια και υποστήριξή του τις στιγμές που αντιμετώπισα δυσκολίες με τη διπλωματική εργασία. Η συμβολή και καθοδήγησή του με επιστημονικό και χιουμοριστικό τρόπο έκαναν όλη αυτήν την προσπάθεια εποικοδομητική και διασκεδαστική ταυτόχρονα.

Περίληψη

Η ανάλυση της απώλειας πελατών (customer churn) σε επιχειρήσεις με πωλήσεις B2B (Business to Business) μπορεί να αποτελέσει σημαντικό εργαλείο στη διατήρηση μεριδίου αγοράς με στοχευμένα μέτρα και πολιτικές διατήρησης πελατών. Η διπλωματική εργασία επικεντρώνεται στην έρευνα της απώλειας B2B πελατών στη βιομηχανία μηχανολογικών κατασκευών, με τη χρήση μεθόδων μηχανικής μάθησης (Machine Learning), και πιο συγκεκριμένα Δέντρα Αποφάσεων (Decision Trees), Νευρωνικά Δίκτυα (Neural Networks), και Λογιστική Παλινδρόμηση (Logistic Regression). Πραγματοποιήθηκε επίσης περιγραφική ανάλυση των αρχικών δεδομένων και τμηματοποίηση πελατών (customer segmentation) με χρήση ανάλυσης RFM και K-means clustering για την αναγνώριση ομάδων πελατών με μεγαλύτερη πιθανότητα αποχώρησης (churn). Η ανάλυση ανέδειξε ότι και τα τρία μοντέλα μπορούν να προβλέψουν με καλή ακρίβεια την απώλεια πελατών με καλύτερο αυτό της Λογιστικής Παλινδρόμησης. Η αποτελεσματικότητα της χρήσης προηγμένων αλγορίθμων μηχανικής μάθησης στην κατανόηση και πρόβλεψη των παραγόντων που επηρεάζουν την απώλεια πελατών στις B2B εταιρείες μπορεί να βοηθήσει στη βιωσιμότητα και ανάπτυξή τους.

Λέξεις κλειδιά: Απώλεια Πελατών, Μηχανική Μάθηση, πωλήσεις B2B, Ανάλυση RFM

Abstract

The analysis of customer churn in businesses with B2B sales can be a significant tool in retaining market share through targeted measures and customer retention policies. The thesis focuses on researching B2B customer churn in the machinery industry, utilizing machine learning methods, and more specifically Decision Trees, Neural Networks, and Logistic Regression. Moreover, a descriptive analysis of the initial data was conducted, and customers were clustered using RFM analysis and K-means clustering to identify groups with higher likelihood of churn. The analysis revealed that all three models can predict customer churn with high accuracy, with the Logistic Regression model performing the best. The effectiveness of using advanced machine learning algorithms in understanding and predicting the factors affecting customer churn in B2B companies can contribute towards their sustainability and growth.

Key words: Customer churn, Machine Learning, B2B sales, RFM analysis

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Διαγραμμάτων

Κατάλογος Εικόνων

1. Εισαγωγή

Ένα από τα διαχρονικά προβλήματα των επιχειρήσεων παγκοσμίως και ταυτοχρόνως ένας αρνητικός παράγοντας για την πορεία - εξέλιξη της επιχείρησης είναι η απώλεια πελατών είτε πρόκειται για φυσικά πρόσωπα είτε για νομικά (εταιρείες). Σε ένα ιδιαίτερα ανταγωνιστικό και ευμετάβλητο επιχειρηματικό περιβάλλον, οι εταιρείες για να αυξήσουν τον τζίρο και να μεγιστοποιήσουν την κερδοφορία τους, επενδύουν όλο και περισσότερα κεφάλαια στην απόκτηση νέων πελατών. Αυτός ο τρόπος δεν είναι πάντοτε ο αποτελεσματικότερος ούτε και ο πιο γρήγορος καθώς για να εδραιωθεί η συνεργασία με ένα νέο πελάτη και να γίνει κερδοφόρος, απαιτείται χρόνος.

Ο σημαντικότερος παράγοντας ανάπτυξης μιας επιχείρησης είναι η διατήρηση των υπαρχόντων πελατών και η μετατροπή τους σε πιστούς πελάτες που διατηρούν μακροχρόνιες συνεργασίες και προχωρούν σε επαναλαμβανόμενες συναλλαγές. Επομένως, η μελέτη των παραγόντων που οδηγούν σε απώλεια πελατών και η πρόβλεψη της πιθανότητας, είναι πολύ σημαντική για τη λειτουργία και ανάπτυξη μιας επιχείρησης (Gordini & Veglio, 2017).

Ο όρος απώλεια πελατών (customer churn), αναφέρεται σε πελάτες που αποχωρούν προς μια ανταγωνίστρια επιχείρηση ή πάροχο υπηρεσιών. Η αποχώρηση αυτή μπορεί να γίνεται για λόγους όπως η καλύτερη ποιότητα υπηρεσιών, προσφορών ή/και προνομίων. Ο ρυθμός της απώλειας πελατών είναι ένας σημαντικός δείκτης που όλες οι εταιρείες επιθυμούν να ελαχιστοποιήσουν. Για τον λόγο αυτόν, η πρόβλεψη της απώλειας πελατών είναι ένα αναπόσπαστο μέρος ενός προληπτικού σχεδιασμού διατήρησης πελατών. Η πρόβλεψη της απώλειας πελατών περιλαμβάνει μοντέλα προβλεπτικής ανάλυσης και εξόρυξης δεδομένων για να προσδιοριστούν οι πελάτες με τη μεγαλύτερη πιθανότητα αποχώρησης. Τα μοντέλα αυτά αναλύουν προσωπικά και συμπεριφορικά δεδομένα πελατών ώστε να δημιουργηθούν στοχευμένες και πελατοκεντρικές καμπάνιες μάρκετινγκ διατήρησης πελατών (Sabbeh, 2018).

Η πρόβλεψη απώλειας πελατών έχει εφαρμοστεί σε διάφορους τομείς όπως οι εκδόσεις, οι τραπεζικές υπηρεσίες, οι ασφάλειες, το ηλεκτρονικό εμπόριο, οι οικονομικές υπηρεσίες, οι τηλεπικοινωνίες, οι λιανικές πωλήσεις, η εφοδιαστική αλυσίδα και οι συνδρομητικές υπηρεσίες. Οι περισσότερες μελέτες αφορούν στις τηλεπικοινωνίες

καθώς είναι ένας ιδιαίτερα ανταγωνιστικός κλάδος με τους πελάτες να έχουν την ελευθερία να επιλέξουν διαφορετικό πάροχο οποτεδήποτε (Van Haver, 2017).

Το κόστος απόκτησης ενός νέου πελάτη στον τομέα των τηλεπικοινωνιών μπορεί να είναι 5 ή περισσότερες φορές μεγαλύτερο από το κόστος διατήρησης ενός υπάρχοντος πελάτη. Σε κλάδους όπου απαιτείται η παρουσία εκπροσώπων της επιχείρησης σε μια ξένη χώρα, επαγγελματικά ταξίδια, συμμετοχή σε τοπικές εκθέσεις καθώς και εκπαίδευση τοπικών συνεργατών, το κόστος απόκτησης νέου πελάτη μπορεί να πολλαπλασιαστεί. Είναι κρίσιμο επομένως να γίνεται σωστή ανάλυση της πιθανότητας απώλειας πελατών ώστε να δημιουργηθούν στοχευμένες ενέργειες μάρκετινγκ ώστε να διατηρηθεί η πελατειακή βάση που θα βοηθήσει την επιχείρηση να συνεχίσει να αναπτύσσεται (Hughes, 2007).

Η διατήρηση πελατών (customer retention) είναι ο βασικός στόχος ενός Συστήματος Διαχείρισης Πελατών CRM (Customer Relationship Management) καθώς το να διατηρηθεί ένας υπάρχων πελάτης μπορεί να κοστίσει 5 έως και 20 φορές λιγότερο από το να αποκτηθεί κάποιος νέος. Η πολιτική διατήρησης πελατών περιλαμβάνει όλες τις δράσεις από μεριάς επιχείρησης ώστε να εξασφαλιστεί η αφοσίωση (loyalty) και να μειωθεί η απώλεια πελατών (Sabbeh, 2018).

Συγκεκριμένα, η φιλοσοφία του Συστήματος Διαχείρισης Πελατών στοχεύει:

- στη δημιουργία διαχρονικά πιστών πελατών και ταυτοχρόνως
- στη μείωση των πιθανοτήτων απώλειά τους,
- στον εντοπισμό των σημαντικότερων πελατών,
- στη διαρκή προσπάθεια διασφάλισης των αναγκών τους και
- στον έλεγχο της ικανοποίησης – εξυπηρέτησης των πελατών (Παξιμάδης, χ.χ.).

Οι σύγχρονες εταιρείες είτε διαθέτουν ένα σύστημα CRM είτε όχι, έχουν ως βασικό εργαλείο για την τμηματοποίηση των πελατών (customer segmentation), την πρόβλεψη απώλειας πελατών (customer churn prediction). Αξιοποιώντας δεδομένα πωλήσεων και χαρακτηριστικά των πελατών και εφαρμόζοντας την κατάλληλη μέθοδο, μπορούν να προβλέψουν την πιθανότητα κάποιος πελάτης να αποχωρήσει στο προσεχές μέλλον. Μια πτυχή του CRM είναι η εκ των προτέρων αναγνώριση πελατών για τους οποίους υπάρχει κίνδυνος να αποχωρήσουν και η εφαρμογή μέτρων διατήρησης τους (Gattermann-Itschert & Thonemann, 2022).

Η πρόβλεψη της απώλειας πελατών είναι περισσότερο κρίσιμη σε περιπτώσεις πωλήσεων B2B εξαιτίας του διεθνούς ανταγωνιστικού περιβάλλοντος των επιχειρήσεων, την αυξανόμενη χρήση του διαδικτύου για σύγκριση προϊόντων και τιμών και κυρίως της μέσης αξίας πελατών B2B σε σχέση με πελάτες B2C (Business to Customer). Σε περιβάλλον πωλήσεων B2B όπου οι πελάτες είναι λιγότεροι σε αριθμό, έχουν μεγαλύτερο τζίρο, αγοράζουν με μεγαλύτερη συχνότητα και επομένως είναι περισσότερο πολύτιμοι για μια επιχείρηση, η διαχείριση της απώλειας πελατών θεωρείται βασικό εργαλείο στην ανάπτυξη πελατειακών σχέσεων. Επιπρόσθετα, λόγω του μεγάλου όγκου συναλλαγών που τυπικά πραγματοποιούν οι B2B πελάτες σε μια επιχείρηση, η διατήρηση τους μπορεί να αποφέρει υψηλές οικονομικές ανταμοιβές σε όποια επιχείρηση λειτουργεί σε αυτό το πλαίσιο (Gordini & Veglio, 2017).

Με την ανάλυση πρόβλεψης της απώλειας πελατών, μια επιχείρηση μπορεί καταρχήν να αναγνωρίσει του πελάτες αυτούς και να κατανοήσει τους λόγους που οδηγούν στην πιθανή απώλεια. Αναγνωρίζοντας τους πελάτες και τους λόγους αυτούς, η επιχείρηση μπορεί να μειώσει το ποσοστό απώλειας παρέχοντας νέα κίνητρα και βελτιώνοντας το προϊόν ή την υπηρεσία που προσφέρουν ικανοποιώντας τους πελάτες αυτούς αλλά και νέους πελάτες εκ των προτέρων (Hughes, 2007).

2. Απώλεια πελατών – Βιβλιογραφική επισκόπηση

2.1. Εισαγωγή – ορισμοί

Η απώλεια πελατών ορίζεται ως ο αριθμός ή το ποσοστό των πελατών που διακόπτουν τις σχέσεις και τις συναλλαγές τους με μια επιχείρηση-προμηθευτή. Η απώλεια αυτή μπορεί να είναι μερική ή ολική. Η μερική απώλεια πελατών αφορά στην αλλαγή ορισμένων συναλλαγών του πελάτη σε μια άλλη επιχείρηση-προμηθευτή, ενώ η ολική απώλεια αφορά στην αλλαγή όλων των συναλλαγών (Van Haver, 2017).

Ο όρος απώλεια πελατών περιγράφει τη διαδικασία κατά την οποία ένας πελάτης διακόπτει την υπηρεσία ή τη συνδρομή που παρέχει ένας προμηθευτής/πάροχος. Πέρα όμως από τη διακοπή μιας υπηρεσίας, ως απώλεια πελατών μπορεί να θεωρηθεί και η αδράνεια πελατών για ένα ορισμένο χρονικό διάστημα που ποικίλλει ανάλογα με τον κλάδο δραστηριότητας (Silpa & Chandran, 2020).

Η απώλεια πελατών μπορεί να είναι εκούσια, όταν ο πελάτης επιλέγει να αλλάξει έναν προμηθευτή ή πάροχο. Μπορεί να είναι τυχαία, όπως για παράδειγμα σε περίπτωση θανάτου του χρήστη αλλά τις περισσότερες φορές είναι εσκεμμένη, όταν ο πελάτης επιλέγει συνειδητά να διακόψει τις συναλλαγές με μια επιχείρηση. Ωστόσο, υπάρχει και η ακούσια απώλεια πελατών όταν η ίδια η επιχείρηση καταργεί ένα προϊόν, απενεργοποιεί μια υπηρεσία ή ακυρώνει ένα συμβόλαιο. Οι λόγοι που μπορεί μια επιχείρηση να ακυρώσει ένα συμβόλαιο μπορεί να είναι η κατάχρηση μια υπηρεσίας ή η αδυναμία αποπληρωμής των συνδρομών από την πλευρά των πελατών. Η ανάλυση της πιθανότητας απώλειας πελατών μπορεί να προλαμβάνει τέτοια φαινόμενα ώστε η επιχείρηση να παρακολουθεί την πορεία των προβληματικών πελατών και να προτείνει λύσεις (Hadden, Tiwari, Roy, & Ruta, 2007).

Πολλές φορές η απώλεια πελατών έχει να κάνει με τη σύναψη ή μη, σύμβασης συνεργασίας μεταξύ πελάτη και επιχείρησης. Όταν ο πελάτης δεν ανανεώνει μια ισχύουσα σύμβαση για μια υπηρεσία που του παρέχεται και αποχωρεί ή ειδοποιεί τον πάροχο της υπηρεσίας ότι επιθυμεί να διακόψει τη σύμβαση, τότε ονομάζεται απώλεια πελατών συμβολαίου (contractual churn) και παρατηρείται σε κλάδους παροχής υπηρεσιών όπως τραπεζικές υπηρεσίες, υπηρεσίες τηλεφωνίας και διαδικτύου, συνδρομητικές υπηρεσίες όπως π.χ. αναπαραγωγή μουσικής, προβολής ταινιών, συνδρομές σε γυμναστήρια κ.α.. Καθότι η πρόθεση του πελάτη για μη ανανέωση καθώς και αυτή καθαυτή η αποχώρηση καθορίζονται χρονικά επακριβώς ως η ημερομηνία διακοπής της σύμβασης, η επιχείρηση που παρέχει την υπηρεσία έχει τη δυνατότητα να αντιδράσει εγκαίρως και να προσπαθήσει να διατηρήσει τον πελάτη (Gattermann-Itschert & Thonemann, 2022).

Όταν δεν υπάρχει σύμβαση μεταξύ πελάτη και επιχείρησης, ο πελάτης μπορεί να διακόψει τη χρήση της υπηρεσίας ή να σταματήσει να αγοράζει προϊόντα από μια επιχείρηση οποτεδήποτε. Καθώς δεν υπάρχει ενεργή σύμβαση, η απώλεια πελατών σε αυτήν την περίπτωση χαρακτηρίζεται ως απώλεια χωρίς δέσμευση συμβολαίου (non contractual churn). Η απώλεια πελάτη σε αυτές τις περιπτώσεις είναι δυσκολότερο να οριστεί καθώς δεν υπάρχει σύμβαση ορισμένου χρόνου και πρέπει να γίνει μια παραδοχή για το πότε ο πελάτης θεωρείται χαμένος. Για παράδειγμα ένας πελάτης θεωρείται χαμένος αν δεν αγοράσει κάποιο προϊόν για 3, 6 ή 12 μήνες. Η αναγνώριση μια πιθανής απώλειας πελάτη εγκαίρως ώστε η επιχείρηση να έχει χρόνο να αντιδράσει και να προσπαθήσει να διατηρήσει τον πελάτη έχει προκλήσεις και η ανάπτυξη προβλεπτικών

μοντέλων παρότι μπορεί να είναι πολύπλοκη, είναι ιδιαίτερα χρήσιμη για τις επιχειρήσεις (Gattermann-Itschert & Thonemann, 2022).

2.2. Παράγοντες απώλειας πελατών

Οι σύγχρονες εταιρίες καλούνται να αντιμετωπίσουν το πρόβλημα της απώλειας πελατών σε ένα ιδιαίτερα ανταγωνιστικό περιβάλλον όπου οι πελάτες είναι πλέον καλά πληροφορημένοι και μπορούν να αλλάξουν προμηθευτή ή πάροχο με μεγάλη ευκολία. Ο κυριότερος λόγος που ένας πελάτης αποφασίζει να αλλάξει προμηθευτή ή πάροχο, είναι η τιμή της υπηρεσίας. Οι πελάτες αναζητούν τη χαμηλότερη δυνατή τιμή και συγκρίνουν την τιμή του υφιστάμενου προμηθευτή με αυτές των ανταγωνιστών. Πέρα από την τιμή αυτή καθαυτή, αναζητούν την καλύτερη σχέση ποιότητας-τιμής και συγκρίνουν την τιμή που πληρώνουν με την αντιλαμβανόμενη αξία που λαμβάνουν. Οι προμηθευτές από τη μεριά τους, προσπαθούν να μειώσουν τις τιμές τους ώστε να διατηρήσουν το πελατολόγιο τους και να προσελκύσουν πελάτες από τον ανταγωνισμό. Είναι γενικά αποδεκτό ότι οι υψηλές τιμές έχουν αρνητική επίπτωση στις πωλήσεις και συμβάλουν στην αύξηση της απώλειας πελατών (Caigny, Coussement, Verbeke, Idbenjra, & Phan, 2021).

Όταν ένας πελάτης είναι δυσαρεστημένος από έναν προμηθευτή ή πάροχο και αποφασίσει να σταματήσει τις συναλλαγές και να στραφεί σε κάποιο νέο, θα πρέπει να συνυπολογίσει το κόστος αλλαγής. Το κόστος αλλαγής προμηθευτή μπορεί να είναι οικονομικό, χρονικό ή και συναισθηματικό. Πολλές εταιρείες προσφέρουν προγράμματα ανταμοιβής ή συλλογής πόντων που ο πελάτης μπορεί να εξαργυρώσει και με τον τρόπο αυτό τους δεσμεύουν για περισσότερο χρόνο καθώς η αποχώρησή τους θα σημάνει και την απώλεια των πόντων ανταμοιβής. Υπάρχουν περιπτώσεις πελατών που παρότι δεν είναι ευχαριστημένοι με τον τωρινό τους προμηθευτή, προτιμούν να μην αλλάξουν καθώς το κόστος της αλλαγής είναι υψηλό (Ahh, Han, & Lee, 2006).

Παράγοντας απώλειας πελατών για μια επιχείρηση μπορεί να είναι οι ανταγωνιστές που προσφέρουν προϊόντα ή/και υπηρεσίες υψηλότερης τεχνολογίας. Ειδικότερα στον τομέα παροχής υπηρεσιών, οι πελάτες μπορεί να αποχωρήσουν ευκολότερα όταν ο ανταγωνισμός μπορεί να προσφέρει καλύτερες υπηρεσίες σε πιο ανταγωνιστικές τιμές (Gattermann-Itschert & Thonemann, 2022).

Πέρα από την τιμή που θεωρείται ο κύριος λόγος απώλειας πελατών, η ποιότητα είναι ένας επίσης πολύ σημαντικός λόγος. Η αντιλαμβανόμενη ποιότητα είναι η διαφορά μεταξύ των προσδοκιών που έχει ένας πελάτης πριν αγοράσει το προϊόν/υπηρεσία και αυτού που τελικά εισπράττει μετά τη χρήση. Κατ' αντιστοιχία, η ικανοποίηση ενός πελάτη είναι η διαφορά στην αξία που λαμβάνει ως προς τις προσδοκίες του για το προϊόν. Ένας από τους σημαντικότερους παράγοντες ικανοποίησης και διατήρησης πελατών είναι η ποιότητα καθώς σε διάφορες μελέτες έχει αποδειχτεί ότι πάνω από το 40% των πελατών συνδρομητικών υπηρεσιών αλλάζει πάροχο εξαιτίας της χαμηλής ποιότητας. Αν ο πελάτης δε μείνει ικανοποιημένος από την όλη εμπειρία αγοράς, χρήσης και αντιλαμβανόμενης αξίας τότε είναι πιθανόν να αποχωρήσει (Ahh, Han, & Lee, 2006). Τέλος, παράγοντες απώλειας πελατών αποτελούν η ασφάλεια και η διαφήμιση. Οι πελάτες θέλουν να νιώθουν ασφαλείς ως προς τα προσωπικά δεδομένα και αν υπάρχει ανησυχία, αυτή προκαλείται από την έλλειψη εμπιστοσύνης προς τον προμηθευτή/πάροχο. Η διαφήμιση βοηθάει στην καθιέρωση της επιχείρησης στην αγορά, δημιουργεί πιστούς πελάτες, προσελκύει νέους και αποτρέπει την απώλεια πελατών (Oghojafor, Mesike, Bakarea, Omoera, & Adeleke, 2012).

Για να αντιμετωπίσουν οι εταιρείες την απώλεια πελατών, εφαρμόζουν καμπάνιες μάρκετινγκ ώστε να διατηρήσουν το πελατολόγιό τους. Οι καμπάνιες αυτές είτε είναι στοχευμένες σε ένα μικρό αριθμό πελατών με κοινά χαρακτηριστικά, είτε είναι πιο γενικές που εφαρμόζονται σε όλο το πελατολόγιο οριζόντια. Θεωρώντας ότι μια επιχείρηση διαθέτει περιορισμένους πόρους, οι στοχευμένες καμπάνιες μάρκετινγκ είναι πολύ πιο αποτελεσματικές (Van Haver, 2017).

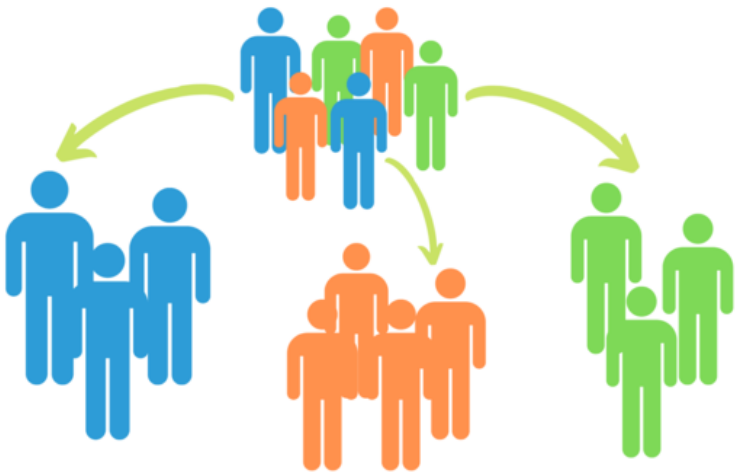
Οι τεχνικές μηχανικής μάθησης μπορούν να βοηθήσουν μια επιχείρηση να αναγνωρίσει ποιοι πελάτες έχουν τη μεγαλύτερη πιθανότητα να αποχωρήσουν και έτσι να εστιάσει τις προωθητικές ενέργειες σε αυτούς. Εκπαιδεύοντας τα μοντέλα με τη βοήθεια διαχρονικών δεδομένων πελατών, μπορούν να αναγνωριστούν μοτίβα και τάσεις πελατών που πιθανώς αποχωρούν. Έτσι, τα μοντέλα είναι σε θέση να προβλέπουν την πιθανότητα μελλοντικής απώλειας πελατών. Είναι προφανές λοιπόν, ότι χρειάζονται αποτελεσματικά και ακριβή μοντέλα πρόβλεψης απώλειας πελατών ώστε να αναγνωρίζεται αξιόπιστα η πιθανότητα αυτής της απώλειας (Gattermann-Itschert & Thonemann, 2022).

2.3. RFM Ανάλυση

Η αύξηση του ανταγωνισμού στον επιχειρηματικό κόσμο αναδεικνύει τη σημασία ανάπτυξης στρατηγικών που θα διασφαλίζουν τη σταθερότητα και την εξέλιξη της κάθε επιχείρησης. Σε ένα περιβάλλον που διακρίνεται από αλλαγές στις προτιμήσεις των καταναλωτών, η διατήρηση του υπάρχοντος πελατολογίου είναι ζωτικής σημασίας για τη μακροπρόθεσμη επιτυχία μιας επιχείρησης, όχι μόνο για τη διασφάλιση σταθερών εσόδων αλλά και για την εξασφάλιση της πελατειακής βάσης ως βασική πηγή προτεραιότητας και συνεχούς ανάπτυξης.

Αναμφίβολα, η διατήρηση των υπαρχόντων πελατών είναι πιο σημαντικό εγχείρημα από την εύρεση νέων. Καθώς όπως αναφέρουν οι Christy, Umamakeswari, Priyatharsini, & Neyaa (2021), σύμφωνα με την αρχή του Ιταλού οικονομολόγου Pareto (Νόμος του 80-20), μόλις το 20% των πελατών αποφέρουν το 80% του συνόλου των εσόδων μιας επιχείρησης.

Η χρήση τεχνικών εξόρυξης δεδομένων αποτελεί μία από τις πιο αποτελεσματικές προσεγγίσεις για την ανάλυση της συμπεριφοράς των πελατών καθώς μέσω αυτών των τεχνικών υλοποιείται η τμηματοποίηση των πελατών και έπειτα η ανάπτυξη κατάλληλων πολιτικών για τη διαχείριση των σχέσεων με τους πελάτες (Sheikh, Ghanbarpour, & Gholamiangonabadi, 2019). Μερικά από τα πιο κοινά χαρακτηριστικά που χρησιμοποιούνται για την τμηματοποίηση των πελατών είναι η τοποθεσία, η ηλικία, το φύλο, το εισόδημα, ο τρόπος ζωής και η προηγούμενη αγοραστική συμπεριφορά (Christy, Umamakeswari, Priyatharsini, & Neyaa, 2021).



Εικόνα 1: Τμηματοποίηση πελατών.

Πηγή: (Roshan, 2020).

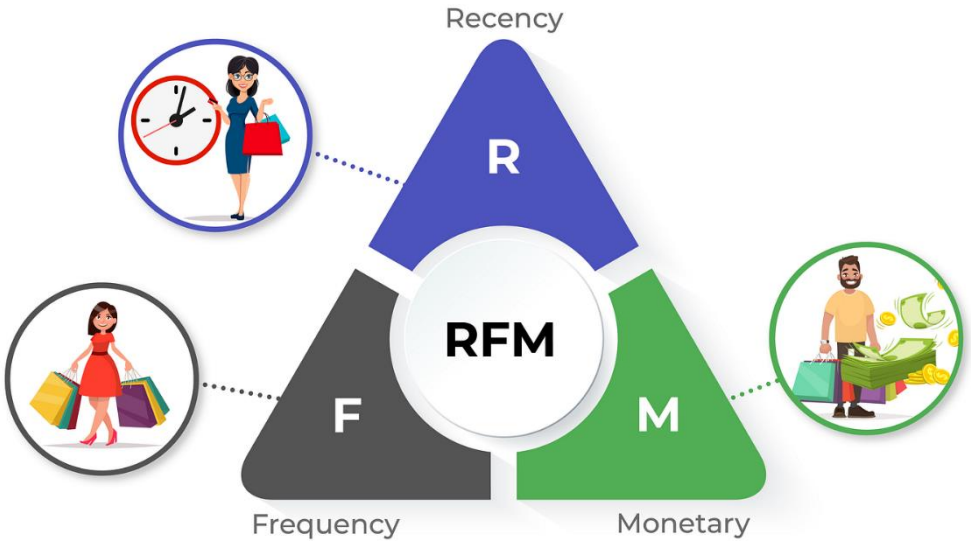
Η ανάλυση της αξίας πελατών είναι μια διαδικασία που χρησιμοποιείται από επιχειρήσεις και οργανισμούς ώστε να βρεθούν σχέσεις και μοτίβα που περιγράφουν την αξία του πελατολογίου. Με την ανάλυση αξίας πελατών, η επιχείρηση μπορεί να εντοπίσει τους πελάτες που συνεισφέρουν το μεγαλύτερο τζίρο ή/και κέρδος σε συγκεκριμένο χρονικό διάστημα. Η ανάλυση RFM (Recency, Frequency, Monetary Value) αποτελεί την πιο διαδεδομένη τεχνική ανάλυσης πελατών, η οποία χρησιμοποιείται για την αξιολόγηση των πελατών βάσει της αγοραστικής τους συμπεριφοράς και στη συνέχεια για την πραγματοποίηση προβλέψεων βάσει αυτής της συμπεριφοράς (Chen, Hu, & Hsieh, 2014), (Christy, Umamakeswari, Priyatharsini, & Neyaa, 2021).

Οι Bult και Wansbeek (1995) ήταν οι πρώτοι που εισήγαγαν το μοντέλο RFM και έκτοτε χρησιμοποιείται ως εργαλείο στην ανάπτυξη στρατηγικών μάρκετινγκ που επεξεργάζονται τις βάσεις δεδομένων τους (Birant, 2011). Η παραπάνω ανάλυση χρησιμοποιήθηκε και στην παρούσα εργασία και βασίζεται στην ανάλυση της συμπεριφοράς των πελατών σε μια περίοδο αναφοράς.

Η ανάλυση RFM είναι το αρκτικόλεξο των λέξεων:

- Recency (R) → η πιο πρόσφατη αγορά του πελάτη
- Frequency (F) → η συχνότητα των αγορών του πελάτη
- Monetary Value (M) → η χρηματική αξία της αγοράς του πελάτη.

Ειδικότερα, η μεταβλητή Recency (R) προσδίδει το διάστημα μεταξύ του χρόνου που λαμβάνει χώρα η αγορά με το σήμερα. Εκφράζεται σε μονάδες όπως ημέρες, μήνες ή έτη και δείχνει κατά πόσο ο πελάτης είναι επίκαιρος καθώς είθισται να θεωρείται ότι οι πελάτες που έχουν αγοράσει πρόσφατα είναι πολύ πιθανόν να επαναλάβουν ξανά μια αγορά σε σχέση με τους λιγότερο πρόσφατους αγοραστές. Όσον αφορά στη δεύτερη διάσταση της ανάλυσης RFM, Frequency (F), εξετάζεται ο συνολικός αριθμός των συναλλαγών που εκτελεί ο πελάτης για την περίοδο αναφοράς που διεξάγεται η ανάλυση, που ομοίως με την προηγούμενη μεταβλητή, οι πελάτες με τις περισσότερες αγορές είναι πιο πιθανόν να αγοράσουν ξανά από τους πελάτες με λιγότερες αγορές. Η τελευταία μεταβλητή Monetary Value (M), αφορά στο σωρευτικό σύνολο των χρημάτων που απέδωσε ο κάθε πελάτης στο εξεταζόμενο διάστημα αναφοράς και φέρει νομισματική μονάδα (Εικόνα 2), (Birant, 2011).



Εικόνα 2: Σχηματική απεικόνιση της RFM ανάλυσης.

Πηγή: (Berkay, 2021)

Το μοντέλο RFM περιλαμβάνει τις παραπάνω τρεις μεταβλητές, οι οποίες μετατρέπονται σε διακριτές τιμές. Συγκεκριμένα, η κάθε μεταβλητή χωρίζεται σε πέντε ίσα μέρη (πεμπτημόρια) στα οποία αποδίδεται μία από τις διακριτές τιμές 1, 2, 3, 4 και 5 (Πίνακας 1). Η τιμή 5 εκχωρείται στο κορυφαίο πεμπτημόριο (20%) της κάθε μεταβλητής, στο αμέσως επόμενο πεμπτημόριο εκχωρείται η τιμή 4 και ούτω καθεξής με την τιμή 1 να αποδίδεται στο τελευταίο και «χειρότερο» πεμπτημόριο της κάθε μεταβλητής (Wei, Lin, & Wu, 2010), (Christy, Umamakeswari, Priyatharsini, & Neyaa, 2021).

Πίνακας 1: Περιγραφικός πίνακας της κλίμακας της RFM ανάλυσης.

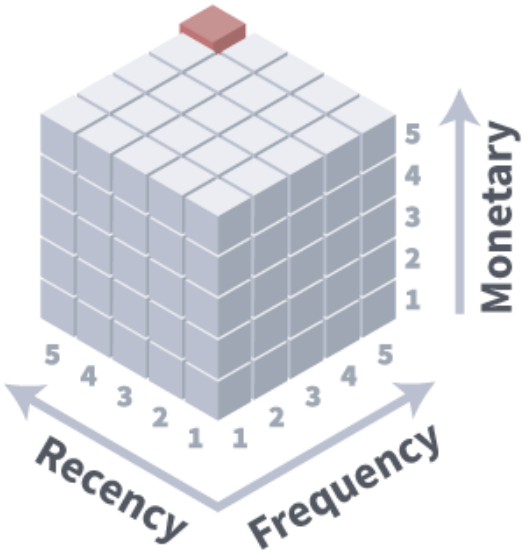
Score	Characteristics
5	Potential
4	Promising
3	Can’t-lose them
2	At risk
1	Lost

Πηγή: (Christy, Umamakeswari, Priyatharsini, & Neyaa, 2021)

Πιο συγκεκριμένα, η διαδικασία για την ποσοτικοποίηση της συμπεριφοράς των πελατών μέσω της RFM ανάλυσης αρχίζει με την ταξινόμηση της βάσης δεδομένων κατά κάθε διάσταση των μεταβλητών RFM. Στη συνέχεια, διαιρείται η κάθε λίστα σε πέντε ίσα τμήματα καθώς η μέθοδος είθισται να έχει το ίδιο μέγεθος στις κλάσεις της τάξης του 20%. Όσον αφορά στη μεταβλητή R, που δείχνει κατά πόσο μία αγορά είναι πρόσφατη, οι πελάτες ταξινομούνται κατά ημερομηνίες αγοράς, χωρίζονται σε πέντε ίσες ομάδες βάσει των ημερομηνιών αυτών και τους αποδίδονται οι αντίστοιχες πέντε διακριτές τιμές.

Για παράδειγμα, το 20% των πελατών που αγόρασαν πιο πρόσφατα λαμβάνουν τη διακριτή τιμή 5 και συνεπώς οι πελάτες αυτοί που έχουν υψηλή βαθμολογία στη μεταβλητή R είναι πιο πιθανό να επαναλάβουν την αγορά (Wei, Lin, & Wu, 2010).

Έπειτα, από το μετασχηματισμό των μεταβλητών στη νέα κλίμακα του πίνακα 1, όλοι οι πελάτες παρουσιάζονται βαθμολογημένοι με 555, 554, 553, …, 111 δημιουργώντας 125 (5x5x5) κελιά RFM (Εικόνα 3). Οι πελάτες με βαθμολογία 555 μπορούν να χαρακτηριστούν ως πιθανοί πελάτες της επιχείρησης, καθώς είναι πιθανό να αποφέρουν περισσότερα κέρδη στην επιχείρηση και αντίστροφα αυτοί με βαθμολογία 111 αποτελεί το λιγότερο κερδοφόρο κομμάτι των πελατών της επιχείρησης (Wei, Lin, & Wu, 2010), (Christy, Umamakeswari, Priyatharsini, & Neyaa, 2021).



Εικόνα 3: Το σύνολο των κελιών RFM χρησιμοποιώντας την κλίμακα από το 1 έως το 5. Πηγή: (loginom, 2021).

Η ευρεία αποδοχή αυτής της μεθόδου έγκειται στο γεγονός ότι ο υπολογισμός των μεταβλητών της ανάλυσης είναι εξαιρετικά απλός, καθώς μπορούν να εξαχθούν εύκολα και γρήγορα από μια βάση δεδομένων που κρατά το ιστορικό των αγορών. Είναι ευρέως κατανοητές έννοιες και ταυτοχρόνως δίνουν τη δυνατότητα της πρόγνωσης της συμπεριφοράς των πελατών (Kohavi & Parekh, 2004). Μέσω της τμηματοποίησης των πελατών σε διάφορες ομάδες η κάθε επιχείρηση δύναται να εντοπίσει τους καλύτερους πελάτες και να οργανώσει μελλοντικές εξατομικευμένες υπηρεσίες σε επιλεγμένους πελάτες που πιθανόν θα ανταποκριθούν στις εκάστοτε προσφορές και να αυξήσει τα κέρδη της. Επίσης, αυτή η γνώση που δίνεται στις επιχειρήσεις, αποσκοπεί στην

καλύτερη εξυπηρέτηση των πελατών της σύμφωνα με τις εξατομικευμένες ανάγκες τους, ώστε να χτιστεί μια σχέση εμπιστοσύνης και αφοσίωσης μεταξύ των δύο μελών (Birant, 2011).

Επομένως, οι πελάτες των οποίων η τελευταία τους αγορά ήταν πρόσφατη, αγοράζουν συχνά και έχουν ξοδέψει αρκετά χρήματα, είναι πιο πιθανό να αγοράσουν ξανά και κατά συνέπεια μέσω της ανάλυσης αυτής κατατάσσονται στην κορυφή του καταλόγου της εκάστοτε επιχείρησης. Η συγκεκριμένη τμηματοποίηση επιτρέπει στην επιχείρηση να αναγνωρίσει τους καλύτερους πελάτες της που συνήθως αντιστοιχούν στο 20% του συνόλου των πελατών της. Χρησιμοποιώντας τις πληροφορίες αυτές που αφορούν στην αγοραστική συμπεριφορά των πελατών η επιχείρηση δύναται να ξεχωρίσει εκείνους που είναι αποδεδειγμένα καλές πηγές εσόδων και έχουν τις καλύτερες αγοραστικές προοπτικές (Lamb, Hair, & McDaniel, 2011).

2.4. Πλεονεκτήματα και μειονεκτήματα της ανάλυσης RFM

Υπάρχουν πολλοί λόγοι για τους οποίους η ανάλυση RFM αναδεικνύεται ως η πιο δημοφιλής και αποτελεσματική μέθοδος στην τμηματοποίηση των πελατών στο χώρο του μάρκετινγκ και των πωλήσεων. Αρχικά, είναι πολύ αποδοτική μέθοδος στην άντληση πληροφοριών σχετικά με τη συμπεριφορά των πελατών. Η ευκολία ποσοτικοποίησης της συμπεριφοράς των πελατών, σε συνδυασμό με την εξόρυξη δεδομένων που αφορούν στις συναλλαγές τους σε προσβάσιμες ηλεκτρονικές μορφές, το καθιστά ένα πρακτικό και αποτελεσματικό εργαλείο. Η απλότητα της μεθόδου διασφαλίζει ότι οι υπεύθυνοι λήψης αποφάσεων μπορούν εύκολα να την κατανοήσουν και να την εφαρμόσουν. Χρησιμοποιεί ένα ελάχιστο αριθμό μεταβλητών και μέσω αυτών επιτυγχάνεται η σύνοψη της αγοραστικής συμπεριφοράς του κάθε πελάτη (Birant, 2011).

Επίσης, η δυνατότητα πρόβλεψης της απώλειας των πελατών και γενικότερα η προγνωστική ικανότητα του μοντέλου δίνει τη δυνατότητα βραχυπρόθεσμα αύξησης των κερδών της εκάστοτε επιχείρησης. Επιπρόσθετα, αξιοποιείται το ιστορικό συναλλαγών για κάθε πελάτη, το οποίο μετατρέπεται με τη μέθοδο αυτή σε μια μεταβλητή συγκεντρωτικού επιπέδου. Αυτή η ιδιότητα επιτρέπει στις επιχειρήσεις να πραγματοποιήσουν στοχευμένες στρατηγικές μάρκετινγκ ώστε να επικεντρωθούν σε μεμονωμένους πελάτες βάσει της αγοραστικής τους συμπεριφοράς. Τέλος, η ανάλυση

RFM ξεχωρίζει για την ικανότητά της να εντοπίζει και να ιεραρχεί τους σημαντικότερους πελάτες (Wei, Lin, & Wu, 2010).

Στον αντίποδα ένα από τα σημαντικότερα μειονεκτήματα της μεθόδου είναι το γεγονός ότι επικεντρώνεται στους καλύτερους πελάτες, και συνεπώς δεν μπορεί να εφαρμοστεί σε νέες επιχειρήσεις οι οποίες έχουν ιδρυθεί σε σύντομο χρονικό διάστημα με πελάτες που έχουν αγοράσει μία μόνο φορά και η παραγγελία τους είναι μικρής αξίας. Επιπρόσθετα, το γεγονός ότι η μέθοδος αξιοποιεί δεδομένα με τους υπάρχοντες πελάτες, αποκλείει μελλοντικούς πελάτες καθώς δεν υπάρχουν τα αντίστοιχα δεδομένα για ανάλυση. Τέλος, ένα από τα πλεονεκτήματα της συγκεκριμένης μεθόδου ταυτόχρονα αποτελεί και μειονέκτημα της ίδιας. Ειδικότερα, λόγω των λιγοστών μεταβλητών που χρησιμοποιεί η μέθοδος, αδυνατεί να περιγράψει σύνθετες σχέσεις με τους πελάτες και περιορίζεται κυρίως στην περιγραφική αναλυτική (descriptive analytics) παρά στην προγνωστική (predictive analytics) (Wei, Lin, & Wu, 2010).

2.5. Διαδικασία πρόβλεψης απώλειας πελατών

Οι περισσότερες εταιρείες για να παρακολουθούν τη συνεργασία με τους πελάτες τους, χρησιμοποιούν λογισμικά CRM και προχωρούν σε ενέργειες βάσει των ευρημάτων της ανάλυσης απώλειας πελατών (churn analysis). Η ανάλυση υποδεικνύει ποιοι πελάτες είναι πιθανόν να αποχωρήσουν το προσεχές διάστημα και τους λόγους πίσω από αυτήν την αποχώρηση. Με τη βοήθεια της ανάλυσης και των προγραμμάτων CRM, καταστρώνονται σχέδια για τη διατήρηση των πελατών (Van Haver, 2017). Η μελέτη της πιθανότητας απώλειας πελάτη είναι σημαντική για τις εταιρείες πρωτίστως για οικονομικούς λόγους καθώς η απώλεια πελατών ισοδυναμεί με μείωση στο τζίρο αλλά και αυξημένα κόστη για την απόκτηση νέων πελατών. Επίσης σημαντικός λόγος για την παρακολούθηση της πιθανότητας απώλειας πελατών είναι η φήμη και η αξιοπιστία των επιχειρήσεων (Silpa & Chandran, 2020).

Στα προβλήματα πρόβλεψης απώλειας πελατών, χρησιμοποιούνται αλγόριθμοι μηχανικής μάθησης που μοντελοποιούν τη σχέση μεταξύ των ανεξάρτητων μεταβλητών ώστε να προβλεφθεί η πιθανότητα αποχώρησης ενός πελάτη. Η τυπική διαδικασία της μελέτης πρόβλεψης απώλειας πελατών περιλαμβάνει αρχικά τη συλλογή δεδομένων και στη συνέχεια την επεξεργασία τους όπως για παράδειγμα καθαρισμός (cleaning), εξομάλυνση (smoothing), κανονικοποίηση (normalization), τυποποίηση

(standardization) κ.α. Μετά την επεξεργασία, τα δεδομένα χρησιμοποιούνται σε μοντέλα για την πρόβλεψη της απώλειας πελατών, γίνεται πιθανή βελτιστοποίηση (optimization) και τέλος τα μοντέλα αξιολογούνται βάσει σχετικών μετρικών αξιολόγησης (evaluation metrics) (Silpa & Chandran, 2020).

Όσον αφορά στη συλλογή δεδομένων, αυτά μπορούν να είναι είτε πρωτογενή είτε δευτερογενή. Πρωτογενή δεδομένα είναι αυτά που συλλέγονται άμεσα και χωρίς επεξεργασία από μια ή περισσότερες πηγές και θεωρούνται αντικειμενικά. Τα δευτερογενή δεδομένα είναι δεδομένα που έχουν ήδη υποστεί επεξεργασία και παρότι θεωρούνται λιγότερο αξιόπιστα, είναι ιδιαίτερα χρήσιμα καθώς το κόστος απόκτησής τους είναι σαφώς μικρότερο (Πρίφτης, 2021). Η συλλογή δεδομένων αποτελεί σημαντικό κομμάτι της έρευνας και είναι απαραίτητο να υπάρχει μια πηγή δεδομένων με σχετικά και χρήσιμα χαρακτηριστικά. Τα δεδομένα που θα συλλεχθούν, θα πρέπει να μετατραπούν σε κατάλληλη μορφή, μια διαδικασία που ονομάζεται επεξεργασία δεδομένων (Silpa & Chandran, 2020).

Η επεξεργασία των δεδομένων περιλαμβάνει τη συγχώνευση, τη διαγραφή, την προσθήκη και τη μετατροπή τους ώστε τα δεδομένα να είναι πλήρως αξιοποιήσιμα κατά την ανάλυση του προβλήματος. Αν τα δεδομένα προέρχονται από διαφορετικές πηγές, το πρώτο βήμα που θα πρέπει να γίνει είναι η συγχώνευση τους σε ένα αρχείο (dataset) ώστε να είναι εύκολη η επεξεργασία. Αν μετά τη συγχώνευση παρατηρηθεί ότι κάποιες μεταβλητές (χαρακτηριστικά) παρουσιάζουν ελλείπουσες τιμές (missing values) τότε θα πρέπει να διορθωθούν. Η πιο απλή λύση είναι οι ελλείπουσες τιμές να διαγραφούν με τον τρόπο αυτό όμως μειώνεται το πλήθος των δεδομένων που θα χρησιμοποιηθούν στην ανάλυση με αποτέλεσμα στα μικρά δείγματα να μην είναι δυνατή η διαγραφή (Kule, Brentari, & Alberici, 2022). Μια άλλη μέθοδος διαχείρισης των ελλειπουσών τιμών είναι να αντικατασταθούν είτε με το μέσο όρο είτε με τη διάμεσο ενώ μπορούν να χρησιμοποιηθούν και μέθοδοι πρόβλεψης των τιμών αυτών. Τέλος, η προσθήκη και μετατροπή των μεταβλητών περιλαμβάνει την προσθήκη ψευδομεταβλητών (dummy variables) σε περιπτώσεις κατηγορικών μεταβλητών, τη μετατροπή ημερομηνιών σε αριθμό ημερών ή τη δημιουργία ποσοστιαίων χαρακτηριστικών (Μητρόπουλος, 2022).

Για την εφαρμογή του μοντέλου πολύ σημαντικός παράγοντας αποτελεί η επιλογή των ανεξάρτητων μεταβλητών (features) καθώς μπορεί να οδηγήσει σε μεγαλύτερη ακρίβεια με τις λιγότερες δυνατές μεταβλητές. Αφού επιλεγούν οι μεταβλητές και τα μοντέλα που

θα χρησιμοποιηθούν στην ανάλυση, τα δεδομένα χωρίζονται σε δεδομένα εκπαίδευσης (training data) που εκπαιδεύουν τα μοντέλα και δεδομένα επαλήθευσης (validation data) που χρησιμοποιούνται για την αξιολόγηση των προβλέψεων των μοντέλων (Müller & Guido, 2017).

Το τελευταίο στάδιο της μελέτης πρόβλεψης απώλειας πελατών αποτελεί η αξιολόγηση και σύγκριση των αποτελεσμάτων. Η αξιολόγηση αλγορίθμων πρόβλεψης απώλειας πελατών είναι κρίσιμη για την κατανόηση της απόδοσής τους και την επιλογή του κατάλληλου αλγορίθμου για την εφαρμογή σε μια συγκεκριμένη επιχειρηματική περίπτωση που θα μπορεί να προβλέπει με τη μεγαλύτερη ακρίβεια στο μέλλον (Kule, Brentari, & Alberici, 2022).

2.6. Πρόβλεψη απώλειας πελατών

Η πρόβλεψη της απώλειας πελατών εμπίπτει σε πρόβλημα ταξινόμησης όπου ένας πελάτης κατατάσσεται ως χαμένος (churn) ή ενεργός (non-churn). Υπάρχουν διάφοροι αλγόριθμοι ταξινόμησης που είναι δημοφιλείς για τέτοιου είδους προβλήματα, ενώ υπάρχουν και υβριδικοί αλγόριθμοι που συνδυάζουν περισσότερους από έναν αλγόριθμους και πηγαίνουν την ανάλυση σε βαθύτερο επίπεδο (Silpa & Chandran, 2020). Στη διεθνή βιβλιογραφία, οι περισσότερες εφαρμογές για την απώλεια πελατών αφορούν σε πωλήσεις B2C, ενώ για περιπτώσεις B2B οι μελέτες περίπτωσης είναι πολύ λιγότερες. Ο περιορισμένος αυτός αριθμός σε συνδυασμό με την όχι και τόσο ξεκάθαρη εφαρμογή των μεθόδων B2C σε οποιοδήποτε σετ δεδομένων, καθιστούν την πρόβλεψη απώλειας πελατών για B2B περιπτώσεις ακόμα πιο σύνθετη διαδικασία (Van Haver, 2017).

Η Aleksandrova (2018), χρησιμοποιώντας την ανάλυση RFM σε συνδυασμό με τις μεταβλητές τύπος πελάτη (ιδιώτης, επιχείρηση) και τον αριθμό των διαφόρων προϊόντων που εμπορεύονται, εφάρμοσαν διάφορους αλγορίθμους πρόβλεψης της απώλειας πελατών. Αρχικά, η ανάλυση RFM έγινε στο πρόγραμμα SPSS, κατηγοριοποιώντας τους πελάτες σε 5 κατηγορίες ανάλογα με το RFM σκορ. Στη συνέχεια, με τη βοήθεια του λογισμικού Azure Machine Learning Studio, εφαρμόστηκαν οι ακόλουθοι αλγόριθμοι δυαδικής ταξινόμησης: Two-Class Boosted Decision Tree, Two-Class Decision Jungle, Two-Class Decision Forest, Two-Class Support Vector Machine, Two-Class Neural Network, Two-Class Logistic Regression (Aleksandrova, 2018). Στη μελέτη αυτή, έγιναν δοκιμές με διαφορετικούς συνδυασμούς μεταβλητών για να γίνει ανάλυση της

πρόβλεψης της απώλειας πελατών, καθορίζοντας την πιθανότητα ένα πελάτης να πραγματοποιήσει συναλλαγή με την επιχείρηση τους επόμενους έξι μήνες. Η ανάλυση των αποτελεσμάτων, απέδειξε ότι ακόμα και όταν χρησιμοποιήθηκαν μόνον οι τρείς βασικές μεταβλητές recency, frequency και monetary, οι αλγόριθμοι πέτυχαν σχετικά καλή πρόβλεψη της απώλειας πελατών. Τα αποτελέσματα έδειξαν ότι οι αλγόριθμοι Two-Class SVM, Two-Class Neural Networks and Two-Class Logistic Regression είχαν καλύτερα αποτελέσματα όταν οι μεταβλητές recency, frequency και monetary χρησιμοποιήθηκαν ως συνεχείς μεταβλητές. Οι αλγόριθμοι Two-Class Decision Jungle, Two-Class Boosted Decision Trees, έδωσαν καλύτερα αποτελέσματα όταν συνδυάστηκαν τα RFM σκορ με μεταβλητές που έχουν να κάνουν με το προφίλ του πελάτη (Aleksandrova, 2018).

Στο άρθρο τους, οι Chen, Hu και Hsieh (2014) ανέλυσαν δεδομένα από 106.747 πελάτες μιας επιχείρησης εφοδιαστικής αλυσίδας (logistics) που αφορούσαν μια περίοδο δύο ετών. Πριν γίνει η ανάλυση της απώλειας πελατών, καθορίστηκαν οι ομάδες των ενεργών και των χαμένων πελατών βασισμένες στο αν ένας πελάτης πραγματοποίησε ή όχι, συναλλαγή με την επιχείρηση τον τελευταίο μήνα. Η κατηγορία των χαμένων πελατών περιλαμβάνει τέσσερις υπό-κατηγορίες: αυτούς που άλλαξαν περιοχή, αυτούς που χρεοκόπησαν, αυτούς που είχαν χρέος προς την επιχείρηση και τέλος αυτούς που έφυγαν οικειοθελώς προς ανταγωνίστρια επιχείρηση. Η ανάλυση έγινε για την τελευταία υποκατηγορία και περιλαμβάνει την ανάλυση αξίας πελατών (Customer Lifetime Value) και την πρόβλεψη απώλειας πελατών. Για την ανάλυση της αξίας πελατών, χρησιμοποιήθηκε η μέθοδος RFM (Recency, Frequency, Monetary) και επιπλέον οι μεταβλητές διάρκεια συνεργασίας (Longevity) και το κέρδος (Profit). Οι μεταβλητές έλαβαν διαφορετικά βάρη με τη μέθοδο αναλυτικής ιεράρχησης (Analytic Hierarchy Process) με τη βοήθεια των προϊσταμένων πωλήσεων της επιχείρησης. Αφού οι πελάτες χωρίστηκαν σε πολύτιμους (valuable) και μη, η ανάλυση απώλειας πελατών εστίασε μόνο στους πολύτιμους πελάτες καθώς συνεισφέρουν περισσότερο στην κερδοφορία της επιχείρησης. Η πρόβλεψη της απώλειας πελατών πραγματοποιήθηκε με τους αλγόριθμους Decision Tree, Multilayer Perceptron (Neural Network), Support Vector Machine και Logistic Regression. Η σύγκριση έδειξε ότι ο αλγόριθμος Decision Tree είχε την καλύτερη ακρίβεια (accuracy) με 93,1%, ενώ ο αλγόριθμος Logistic Regression είχε τη μικρότερη ακρίβεια με 87,6%. Οι μεταβλητές που είχαν τη μεγαλύτερη επίδραση στην πρόβλεψη της απώλειας πελατών ήταν η πιο πρόσφατη συναλλαγή (recency), η διάρκεια

της συνεργασίας (longevity) και τέλος ο συνολικός τζίρος του πελάτη (monetary) (Chen, Hu, & Hsieh, 2014).

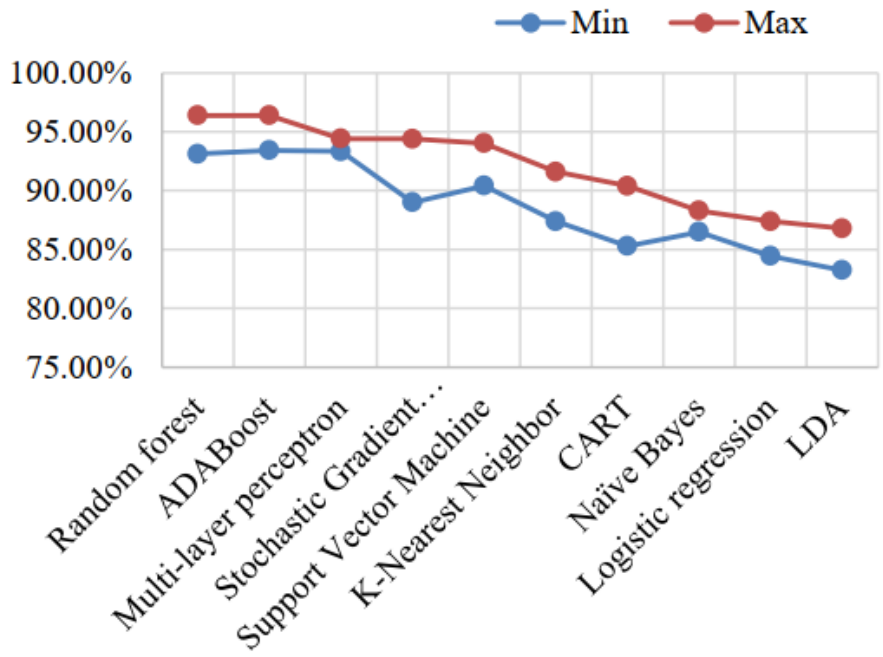
Οι Alwis, Kumara, & Hapuarachchi (2018) μελετούν την απώλεια πελατών σε τηλεπικοινωνιακούς παρόχους χρησιμοποιώντας αλγόριθμους πρόβλεψης. Στόχος της μελέτης είναι να βρεθούν εκείνοι οι πελάτες που είναι πιθανόν να αποχωρήσουν καθώς και να γίνει μια ταξινόμηση πελατών σε κατηγορίες ώστε να μπορέσουν να εφαρμοστούν ενέργειες μάρκετινγκ. Η ταξινόμηση έγινε σε τέσσερις κατηγορίες με τους πελάτες της κάθε κατηγορίας να έχουν παρόμοια χαρακτηριστικά. Όπως προέκυψε, οι πελάτες που συνήθως αποχωρούν είναι άνδρες επιχειρηματίες ή ιδιωτικοί υπάλληλοι και γυναίκες δημόσιοι υπάλληλοι. Αντιθέτως, οι πελάτες που συνήθως παραμένουν σε έναν τηλεπικοινωνιακό πάροχο είναι γυναίκες που πηγαίνουν σχολείο ή πανεπιστήμιο. Για τους αλγόριθμους πρόβλεψης, τα δεδομένα βασίστηκαν σε 200 ερωτηματολόγια από τα οποία δημιουργήθηκαν 14 μεταβλητές όπως ηλικία, φύλο, επάγγελμα, εισόδημα, κατηγορία συμβολαίου, διάρκεια συνεργασίας, αξία λογαριασμού, χρήση ίντερνετ κ.α. Από τις 14 αρχικές μεταβλητές και με εφαρμογή του τεστ συσχέτισης x^2 Pearson, μόνο οι 11 σχετίζονται τελικά με την απώλεια ή μη των πελατών και είναι αυτές που χρησιμοποιήθηκαν στην ανάλυση. Στη συγκεκριμένη μελέτη, χρησιμοποιήθηκαν τέσσερις διαφορετικοί αλγόριθμοι στο λογισμικό SPSS Modeler 18.0 και συγκεκριμένα οι C5.0 tree, Bayesian network, Neural Network και Logistic regression. Η ανάλυση έδειξε ότι ο αλγόριθμος C5.0 tree που βασίζεται σε Decision Trees, είχε τη μεγαλύτερη ακρίβεια πρόβλεψης με 85% και τιμή AUC (Area Under Curve) 0,888 (Alwis, Kumara, & Hapuarachchi, 2018).

Στο άρθρο Sabbeh (2018), πραγματοποιείται μια συγκριτική μελέτη ανάμεσα στους πιο διαδεδομένους αλγόριθμους πρόβλεψης απώλειας πελατών. Ο στόχος είναι να αναλυθεί και να συγκριθεί η απόδοση των αλγορίθμων που εμφανίζονται περισσότερο συχνά στη βιβλιογραφία οι οποίοι είναι οι παρακάτω:

1. Logistic regression
2. Decision tree (CART)
3. Naive Bayesian
4. Support Vector Machine
5. *K*-nearest Neighbor
6. *A*da Boost
7. Stochastic Gradient Boost

8. Random Forest
9. Artificial neural network (Multi-layer Perceptron)
10. Linear Discriminant Analysis (Sabbeh, 2018).

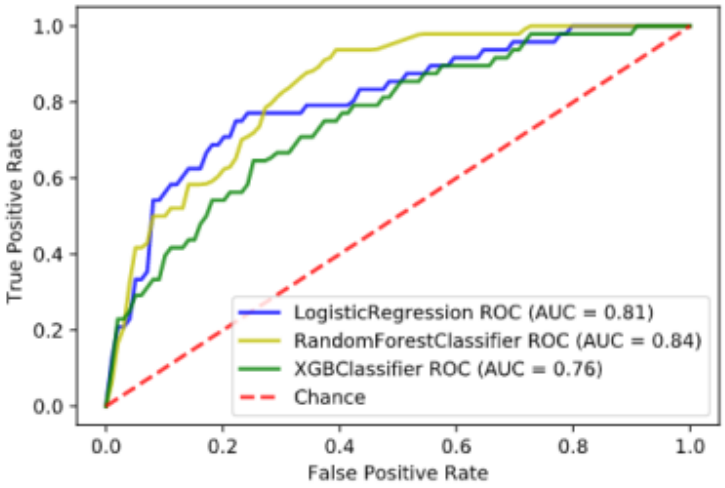
Τα δεδομένα που χρησιμοποιήθηκαν, προήλθαν από τη βάση δεδομένων μια επιχείρησης τηλεπικοινωνιών και περιλάμβαναν 17 μεταβλητές. Για την επιλογή των μεταβλητών που θα χρησιμοποιούνταν στην ανάλυση, οι μεταβλητές κατανεμήθηκαν με φθίνουσα σειρά σπουδαιότητας βάσει της τεχνικής Random Forest και Boruta και τελικά επιλέχτηκαν οι 13. Με το τρόπο αυτό, το τελικό σετ δεδομένων περιλάμβανε 3333 εγγραφές με 13 μεταβλητές πρόβλεψης και την κατηγορική μεταβλητή churn. Τα δεδομένα στη συνέχεια χωρίστηκαν τυχαία σε δεδομένα εκπαίδευσης και δεδομένα επαλήθευσης σε ποσοστό 60% και 40% αντίστοιχα. Μετά την εφαρμογή όλων των αλγορίθμων, τα αποτελέσματα έδειξαν ότι οι αλγόριθμοι Random Forest και AdaBoost πέτυχαν την καλύτερη προβλεπτική απόδοση με ποσοστό ακρίβειας (accuracy) 96%. Αντίστοιχα καλό ποσοστό είχαν και οι αλγόριθμοι Neural Networks και Support Vector Machine με ακρίβεια πρόβλεψης 94%. Χαμηλότερα ποσοστά ακρίβειας παρουσίασαν ο αλγόριθμος Decision Trees με 90%, Naive Bayes με 88% και τέλος οι Logistic Regression και Linear Discriminant Analysis με 86,7%. Τα αποτελέσματα φαίνονται συνοπτικά στην εικόνα 4 (Sabbeh, 2018).



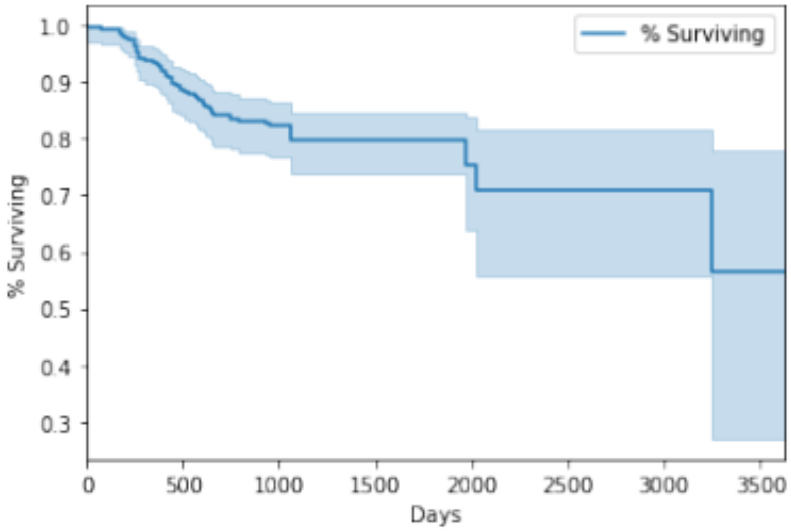
Εικόνα 4: Ακρίβεια αλγορίθμων της μελέτης Sabbeh (2018).

Πηγή: (Sabbeh, 2018).

Η μελέτη των Hills, Daniel, Lu, Schaer, & Adams (2020) είναι ένα παράδειγμα πρόβλεψης απώλειας πελατών σε επιχείρηση με πωλήσεις B2B. Οι πελάτες που περιλαμβάνονται στην ανάλυση είναι 242 και οι μεταβλητές που χρησιμοποιούνται είναι 16 μεταξύ των οποίων ο αριθμός εργαζομένων του πελάτη, ο τζίρος, η διάρκεια της συνεργασίας, ο αριθμός επισκέψεων στην ιστοσελίδα, αριθμός τηλεφωνημάτων για τεχνική υποστήριξη κ.α. Η ανάλυση περιλαμβάνει δύο στάδια, μια κλασική προσέγγιση πρόβλεψης απώλειας πελατών με τρεις διαφορετικούς αλγόριθμους και μια προσέγγιση που ερευνά την παράμετρο του χρόνου στην απώλειας πελατών χρησιμοποιώντας μοντελοποίηση επιβίωσης (survival modelling). Το μοντέλο επιβίωσης ουσιαστικά δείχνει το χρόνο που ένας πελάτης παρέμεινε ενεργός πριν αποχωρήσει. Για τον καθορισμό των μεταβλητών που χρησιμοποιήθηκαν στους αλγορίθμους πρόβλεψης, έγινε ένας πίνακας συσχέτισης Pearson που δείχνει ποιες μεταβλητές συσχετίζονται μεταξύ τους ενώ έγινε και ιεράρχηση της σπουδαιότητας τους. Από τις 16 αρχικές μεταβλητές, τελικά επιλέχτηκαν οι 15. Εξαιτίας του μικρού μεγέθους του δείγματος, οι ερευνητές δε χρησιμοποίησαν δεδομένα εκπαίδευσης (train) και δεδομένα δοκιμής (test), εναλλακτικά επικύρωσαν με τριπλή διασταύρωση δεδομένων τις μετρικές αξιολόγησης των αλγορίθμων. Από τους τρεις αλγόριθμους που χρησιμοποιήθηκαν, το μοντέλο Random Forest είχε το καλύτερο ποσοστό πρόβλεψης με AUC 84%, ακολουθούμενο από το μοντέλο Logistic Regression με 81% και τέλος το μοντέλο XGBoost με 76%. Δεδομένης της απλότητας αλλά και της ευκολίας κατανόησης του μοντέλου Logistic Regression, οι ερευνητές συμπέραναν ότι το μοντέλο αυτό είναι το κατάλληλο για τη δεδομένη περίπτωση. Το γράφημα με τα αποτελέσματα ROC/AUC των τριών αλγορίθμων, φαίνονται στην εικόνα 5. Όσον αφορά στο μοντέλο επιβίωσης, έγινε εφαρμογή του αλγορίθμου Kaplan-Meier Estimate το οποίο έδειξε ότι οι πελάτες είναι πιθανόν να αποχωρήσουν μετά από 600 μέρες συνεργασίας με την επιχείρηση και προτείνουν ενέργειες μάρκετινγκ διατήρησης πελατών περίπου στον ενάμιση χρόνο συνεργασίας. Επίσης παρατηρείται μια μεγάλη πτώση στο ποσοστό επιβίωσης των πελατών μετά τις 1000 μέρες και η επόμενη δραστική πτώση του ποσοστού εμφανίζεται περίπου στις 2000 μέρες. Στην εικόνα 6 φαίνεται διαχρονικά το ποσοστό επιβίωσης του πελάτη για το συγκεκριμένο παράδειγμα (Hills, Daniel, Lu, Schaer, & Adams, 2020).



Εικόνα 5: Αποτελέσματα ROC/AUC για τους 3 αλγορίθμους.
Πηγή: (Hills, Daniel, Lu, Schaer, & Adams, 2020)



Εικόνα 6: Διαχρονική εξέλιξη επιβίωσης πελάτη.
Πηγή: (Hills, Daniel, Lu, Schaer, & Adams, 2020)

Οι Gattermann-Itschert και Thonemann (2022) μελετούν την απώλεια B2B πελατών σε μια μεγάλη επιχείρηση χονδρικού εμπορίου στην Ευρώπη. Χρησιμοποιώντας μοντέλα μηχανικής μάθησης, γίνεται πρόβλεψη της απώλειας πελατών με μεγάλη ακρίβεια χρησιμοποιώντας την ανάλυση RFM (Recency, Frequency, Monetary) καθώς και άλλων χαρακτηριστικών όπως για παράδειγμα ο χρόνος από την τελευταία επικοινωνία του πελάτη με κάποιον εκπρόσωπο της επιχείρησης. Οι αλγόριθμοι που χρησιμοποιήθηκαν είναι Logistic Regression, Support Vector Machines και Random Forests. Από την ανάλυση των αποτελεσμάτων και της προβλεπτικής απόδοσης, η μελέτη έδειξε ότι ο αλγόριθμος Random Forest είχε την καλύτερη ακρίβεια με 0,758 AUC (Area Under

Curve). Οι μεταβλητές που επηρεάζουν περισσότερο την πρόβλεψη της απώλειας αποδείχτηκε ότι είναι οι τρεις μεταβλητές RFM και επίσης ο χρόνος από την τελευταία επικοινωνία με εκπρόσωπο της επιχείρησης. Η έρευνα χρησιμοποιώντας τα αποτελέσματα των προβλέψεων, στόχευσε στους πελάτες με τη μεγαλύτερη πιθανότητα να αποχωρήσουν για ένα διάστημα τριών μηνών με καμπάνιες μάρκετινγκ. Στοχεύοντας σε μόλις 15% των πελατών αυτών, η επιχείρηση κατάφερε να μειώσει το ποσοστό απώλειας πελατών από 12,9% σε 11,6%. Συγκριτικά με τυχαία επικοινωνία με το πελατολόγιο, η στοχευμένη προσέγγιση πελατών κατάφερε διπλάσιο αριθμό τηλεφωνημάτων προς την επιχείρηση και διπλάσιες επισκέψεις εκ μέρους των πελατών (Gattermann-Itschert & Thonemann, 2022).

Στη μελέτη των Gordini & Veglio, αναλύεται η απώλεια πελατών σε μια επιχείρηση ηλεκτρονικού χονδρικού εμπορίου στην Ιταλία με 80.000 πελάτες και δεδομένα συναλλαγών ενός έτους. Οι μεταβλητές που χρησιμοποιήθηκαν στις διαφορετικές μεθόδους περιλαμβάνουν δημογραφικά στοιχεία των πελατών και στοιχεία συναλλαγών. Οι αλγόριθμοι της ανάλυσης περιλάμβαναν Support Vector Machines, Neural Networks και Logistic Regression. Σε όλους τους αλγορίθμους, η μεταβλητή με τη μεγαλύτερη βαρύτητα ήταν η πιο πρόσφατη συναλλαγή (Recency) ενώ σε δύο από τους τρεις αλγορίθμους, η δεύτερη σημαντικότερη μεταβλητή ήταν η συχνότητα συναλλαγών επιβεβαιώνοντας τη σημαντικότητα της RFM ανάλυσης. Η σύγκριση των αποτελεσμάτων ανέδειξε ότι ο αλγόριθμος SVM_{auc} με αντιστοίχιση μεταβλητών είχε την υψηλότερη ακρίβεια ποσοστού σωστών προβλέψεων με 89,67%, ακολουθούμενος από τον SVM_{acc} με 86,13%, το Neural Networks με 85,1% και το Logistic Regression με 83,8%. Η ίδια εικόνα παρατηρήθηκε και για τα αποτελέσματα AUC και top decile. Συμπερασματικά, ο αλγόριθμος SVM_{auc} δίνει καλά αποτελέσματα σε εφαρμογές B2B, ιδιαίτερα σε κλάδους με υψηλό ρυθμό απώλειας πελατών, κάτι που επιβεβαιώνει και η βιβλιογραφία κατά τα λεγόμενα των συγγραφέων (Gordini & Veglio, 2017).

Οι Bagul, Surana, Berad, & Khachane, 2021, στην έρευνά τους πραγματοποίησαν τμηματοποίηση πελατών με τον αλγόριθμο K-Means βάσει της RFM ανάλυσης με στόχο να διατηρηθούν οι πελάτες που έχουν μεγαλύτερη πιθανότητα να αποχωρήσουν. Πιο συγκεκριμένα, χρησιμοποιήθηκαν τα δεδομένα συναλλαγών μιας επιχείρησης λιανικού εμπορίου στο Ηνωμένο Βασίλειο που περιλάμβαναν μεταξύ άλλων, στοιχεία για την αξία της συναλλαγής και την ημερομηνία. Από τα πρωτογενή δεδομένα δημιουργήθηκαν οι μεταβλητές Recency που αναφέρεται στην πιο πρόσφατη συναλλαγή, Frequency που

αναφέρεται στο συνολικό αριθμό συναλλαγών ενός συγκεκριμένου πελάτη και Monetary που αναφέρεται στο συνολικό τζίρο του πελάτη. Δίνοντας τιμές 1-5 σε κάθε μεταβλητή και αθροίζοντας τις τιμές των τριών μεταβλητών, προέκυψε το συνολικό RFM score το οποίο χρησιμοποιήθηκε για να γίνει τμηματοποίηση των πελατών με τη βοήθεια του αλγορίθμου K-Means Clustering. Ο βέλτιστος αριθμός των ομάδων/κλάσεων των πελατών πραγματοποιήθηκε με τη μέθοδο Elbow που υπολογίζει το άθροισμα των τετραγώνων των αποστάσεων του κάθε σημείου από το κέντρο της κάθε κλάσης. Η μέθοδος υπέδειξε τρεις κλάσεις που ταξινομούν τους πελάτες σε τρεις κατηγορίες με παρόμοια χαρακτηριστικά. Η πρώτη κλάση περιλαμβάνει τους καλύτερους πελάτες, η δεύτερη τους πιστούς πελάτες και η τρίτη τους πελάτες που βρίσκονται σε κίνδυνο να αποχωρήσουν (churn). Η τρίτη κλάση, είναι ιδιαιτέρως χρήσιμη στην επιχείρηση καθώς μπορούν να εφαρμοστούν συγκεκριμένες τεχνικές με στόχο τη διατήρηση των πελατών αυτών (Bagul, Surana, Berad, & Khachane, 2021).

3. Μέθοδοι Μηχανικής Μάθησης

3.1. Μηχανική Μάθηση (Machine Learning)

Η Μηχανική Μάθηση είναι ένα πεδίο έρευνας που συνδυάζει την επιστήμη της Στατιστικής, την Τεχνητή Νοημοσύνη και την Επιστήμη Υπολογιστών με στόχο την ανάπτυξη συστημάτων που εκπαιδεύονται και βελτιώνονται αυτόματα. Όταν η Μηχανική Μάθηση χρησιμοποιεί ένα πλαίσιο στατιστικών αναλύσεων ονομάζεται εναλλακτικά Στατιστική Μάθηση (Statistical Learning) (Sarker, 2021).

Η έννοια της Μηχανικής Μάθησης αναφέρεται στη μελέτη αλγορίθμων που βοηθούν ένα πρόγραμμα/μηχανή να πραγματοποιήσει μια συγκεκριμένη εργασία χωρίς την παρέμβαση του ανθρώπου. Το πρόγραμμα εκπαιδεύεται με ένα συγκεκριμένο αριθμό χαρακτηριστικών και στη συνέχεια προσπαθεί να κάνει το ίδιο βασιζόμενο σε συμπεράσματα και μοτίβα που έχει μάθει κατά την εκπαίδευση (Silpa & Chandran, 2020).

Οι αλγόριθμοι Μηχανικής Μάθησης μπορούν να λύνουν προβλήματα παλινδρόμησης (Regression), ταξινόμησης (Classification), ομαδοποίησης (Clustering), αναγνώρισης ανωμαλιών (Anomaly Detection), μείωσης διαστάσεων (Dimensionality Reduction),

μάθηση κανόνων συσχέτισης (Association Rule Learning) και άλλα. Η παλινδρόμηση χρησιμοποιείται σε προβλήματα πρόβλεψης μιας συνεχούς μεταβλητής, η ταξινόμηση σε προβλήματα ταξινόμησης δεδομένων σε δύο ή περισσότερες κλάσεις, η ομαδοποίηση σε προβλήματα που χρειάζεται να βρεθεί ένα μοτίβο που ομαδοποιεί δεδομένα με παρόμοια χαρακτηριστικά, η αναγνώριση ανωμαλιών χρησιμοποιείται σε προβλήματα που επιχειρούν να αναγνωρίσουν αποκλίνουσες τιμές από ένα σύνολο δεδομένων, η μείωση διαστάσεων σε προβλήματα που χρειάζεται να μειωθούν οι απαραίτητες μεταβλητές εισόδου και τέλος η μάθηση κανόνων συσχέτισης για προβλήματα συσχέτισης δεδομένων με τη λογική αν x τότε y (If-Then). Οι αλγόριθμοι Μηχανικής Μάθησης έχουν ευρύ φάσμα εφαρμογών όπως κυβερνοασφάλεια, πρόβλεψη κυκλοφορίας οχημάτων, αυτόνομη οδήγηση, ιατρική διάγνωση, αναγνώριση εικόνας και φωνής, σύσταση προϊόντος και άλλα (Sarker, 2021), (Carleo, et al., 2019).

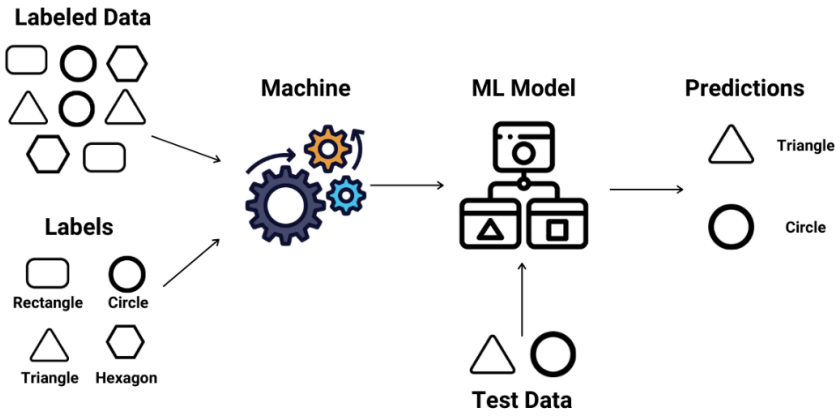
Οι δύο βασικές προσεγγίσεις των τεχνικών μηχανικής μάθησης είναι η επιβλεπόμενη μάθηση (Supervised) και η μη επιβλεπόμενη μάθηση (Unsupervised), ενώ υπάρχει και μια τρίτη κατηγορία αυτή της ενισχυτικής μάθησης (Reinforcement) η οποία είναι λιγότερο διαδεδομένη (Μητρόπουλος, 2022).

3.1.1. Επιβλεπόμενη μάθηση

Οι πιο διαδεδομένοι αλγόριθμοι Μηχανικής Μάθησης είναι αυτοί που χρησιμοποιούνται σε διαδικασίες λήψης αποφάσεων, γενικεύοντας μοτίβα από ήδη γνωστά παραδείγματα (Sarker, 2021). Στην επιβλεπόμενη μάθηση, ο εκπαιδευτής έχοντας γνώση του περιβάλλοντος δίνει χειροκίνητα μια δεδομένη τιμή απόκρισης (Label) στην εξαρτημένη μεταβλητή βάσει των δεδομένων που έχει. Στόχος είναι η εύρεση της κατάλληλης σχέσης με την οποία αντιστοιχίζονται τα δεδομένα εισόδου με τα ήδη χαρακτηρισμένα δεδομένα (Labeled). Με τον τρόπο αυτό, η επιβλεπόμενη μάθηση προβλέπει μια εξαρτημένη μεταβλητή y βάσει ανεξάρτητων μεταβλητών x (Μητρόπουλος, 2022). Πιο συγκεκριμένα, ο αλγόριθμος επιβλεπόμενης μάθησης έχει την ικανότητα να αποδίδει μία τιμή στην εξαρτημένη μεταβλητή για ένα σύνολο ανεξάρτητων μεταβλητών που δεν έχει ξανά συναντήσει, χωρίς οποιαδήποτε βοήθεια από άνθρωπο (Sarker, 2021).

Στην περίπτωση που η εξαρτημένη μεταβλητή y είναι κατηγορική παίρνει συγκεκριμένες τιμές από ένα δεδομένο σύνολο τιμών που υποδεικνύουν τις κλάσεις του προβλήματος

και η διαδικασία ονομάζεται ταξινόμηση. Αν η ταξινόμηση γίνεται σε ακριβώς δύο κλάσεις, τότε η ταξινόμηση ονομάζεται δυαδική (Binary Classification) ενώ αν γίνεται σε πάνω από δύο κλάσεις τότε ονομάζεται πολλαπλών κλάσεων (Multiclass Classification). Αν η εξαρτημένη μεταβλητή y παίρνει συνεχείς τιμές, τότε η διαδικασία ονομάζεται παλινδρόμηση. Η διαδικασία της επιβλεπόμενης Μηχανικής Μάθησης απεικονίζεται συνοπτικά στην εικόνα 7 (Müller & Guido, 2017).



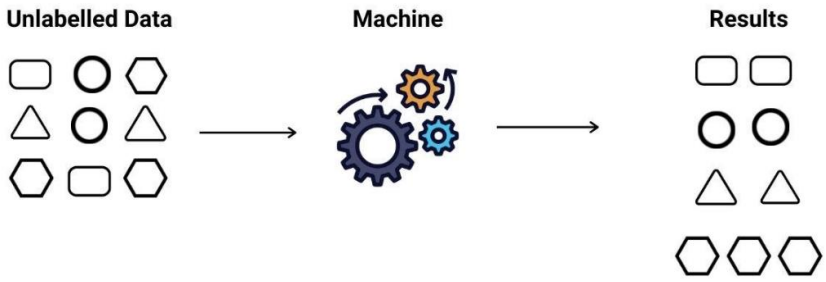
Εικόνα 7: Διαδικασία επιβλεπόμενης Μηχανικής Μάθησης.
Πηγή: (Ravish)

3.1.2. Μη επιβλεπόμενη μάθηση

Στη μη επιβλεπόμενη μάθηση δεν υπάρχει εκπαιδευτής που μπορεί να δώσει την κατάλληλη τιμή απόκρισης (Label) και η μάθηση οργανώνεται αυτόνομα ακολουθώντας συγκεκριμένα μοτίβα. Χρησιμοποιώντας τα δεδομένα, ο αλγόριθμος εκμάθησης προσπαθεί να ανακαλύψει μια δομή και ομαδοποιεί την αταξινόμητη πληροφορία βάσει ομοιοτήτων και διαφορών μεταξύ των δεδομένων (Silpa & Chandran, 2020). Οι κυριότερες κατηγορίες μη επιβλεπόμενης μάθησης είναι η τμηματοποίηση,

η ομαδοποίηση και οι μετασχηματισμοί δεδομένων (Dataset Transformations) όπου στη μεν πρώτη ομαδοποιούνται δεδομένα με κοινά χαρακτηριστικά ενώ στη δεύτερη τα δεδομένα αναπαρίστανται με λιγότερα χαρακτηριστικά και με πιο κατανοητό τρόπο (Müller & Guido, 2017). Δευτερευόντως απαντώνται και οι κατηγορίες ανίχνευσης ανωμαλιών που στόχο έχει να βρει αποκλίνουσες τιμές από ένα σύνολο δεδομένων καθώς και των αυτόματων κωδικοποιητών (Autoencoders) που χρησιμοποιούνται σε εργασίες

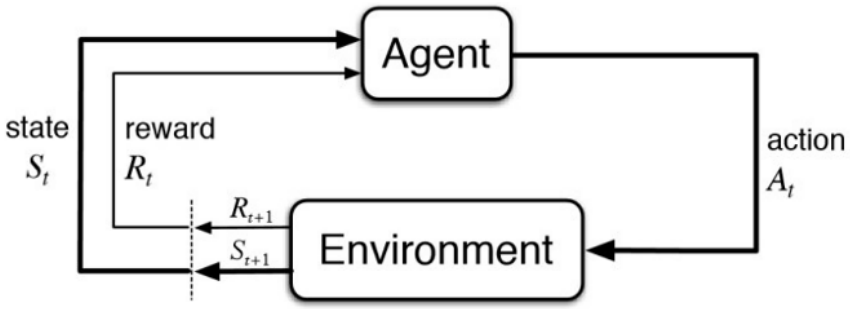
μάθησης αναπαράστασης. Η διαδικασία της μη επιβλεπόμενης Μηχανικής Μάθησης απεικονίζεται συνοπτικά στην εικόνα 8 (Μητρόπουλος, 2022).



Εικόνα 8: Διαδικασία μη επιβλεπόμενης Μηχανικής Μάθησης.
Πηγή: (Ravish)

3.1.3. Ενισχυτική μάθηση

Η ενισχυτική μάθηση (Reinforcement Learning) είναι μία υποκατηγορία της μηχανικής μάθησης που επιτρέπει σε ένα σύστημα να μάθει να προσαρμόζεται σε ένα περιβάλλον μέσω διαδικασιών δοκιμής και σφάλματος (trial & error). Στην ενισχυτική μάθηση, ένας πράκτορας (agent) αλληλοεπιδρά σε ένα περιβάλλον, λαμβάνοντας αποφάσεις και εκτελώντας ενέργειες προκειμένου να μεγιστοποιήσει μια ανταμοιβή (reward) ή να επιτύχει έναν συγκεκριμένο στόχο. Κάθε ενέργεια (action) του πράκτορα, μεταβάλει την κατάσταση (state) του περιβάλλοντος και στη συνέχεια συλλέγει νέες παρατηρήσεις για νέα εν δυνάμει ανταμοιβή σε μια επαναληπτική διαδικασία όπως απεικονίζεται στην εικόνα 9. Βάσει των παρατηρήσεων ο πράκτορας αποφασίζει για την επόμενη ενέργεια φιλτράροντας στρατηγικές που μεγιστοποιούν το κέρδος. Αυτός ο τύπος μηχανικής μάθησης εφαρμόζεται σε περιπτώσεις όπου ο μόνος τρόπος για να διερευνηθούν οι ιδιότητες ενός περιβάλλοντος είναι να υπάρξει αλληλεπίδραση με τον πράκτορα. Η βασική αρχή της ενισχυτικής μάθησης είναι η εναλλαγή μεταξύ βέλτιστων στρατηγικών που έχουν αναγνωριστεί ήδη και τη διερεύνηση νέων καλύτερων στρατηγικών (Carleo, et al., 2019).



Εικόνα 9: Διαγραμματική απεικόνιση ενισχυτικής μάθησης.
Πηγή: (Schwartz, 2014)

3.2. Αλγόριθμοι Επιβλεπόμενης Μηχανικής Μάθησης για την πρόβλεψη της απώλειας πελατών

Στην επιβλεπόμενη Μηχανική Μάθηση υπάρχουν δύο βασικοί τύποι αλγορίθμων, οι αλγόριθμοι ταξινόμησης και οι αλγόριθμοι παλινδρόμησης. Οι πιο διαδεδομένοι μέθοδοι πρόβλεψης απώλειας πελατών που συναντώνται στη βιβλιογραφία είναι η Λογιστική Παλινδρόμηση (Logistic Regression), τα Δέντρα Αποφάσεων (Decision Trees), τα Νευρωνικά Δίκτυα (Neural Networks) οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines), ο αλγόριθμος Κ-εγγύτερων γειτόνων (K-Nearest Neigbors) και ο αλγόριθμος της αφελούς υπόθεσης του Bayes (Naive Bayes) (James, Witten, Hastie, & Tibshirani, 2023).

3.2.1. Λογιστική Παλινδρόμηση (Logistic Regression)

Όσον αφορά στη Λογιστική Παλινδρόμηση, πρόκειται για έναν αλγόριθμο μηχανικής μάθησης που χρησιμοποιείται ευρέως για την πρόβλεψη διαφόρων προβλημάτων, όπως για την πρόβλεψη απώλειας πελατών, βάσει μιας ή περισσοτέρων μεταβλητών. Ο αλγόριθμος είναι απλός και ευέλικτος και χρησιμοποιεί τη Λογιστική συνάρτηση που μετατρέπει το γραμμικό συνδυασμό των χαρακτηριστικών σε μια πιθανότητα με τιμές μεταξύ 0 και 1 (Silpa & Chandran, 2020).

Η Λογιστική Παλινδρόμηση χρησιμοποιείται σε περιπτώσεις δυαδικής ταξινόμησης όπου μια εξαρτημένη μεταβλητή (dependent variable) μπορεί να πάρει δύο τιμές που στην περίπτωση της απώλειας πελατών είναι απώλεια ή όχι. Ο αλγόριθμος δίνει μια πιθανότητα από 0 έως 1 και ταξινομεί τους πελάτες που έχουν πάνω από 0,5 ως χαμένους

και κάτω από 0,5 ως ενεργούς βάσει ανεξάρτητων μεταβλητών (independent variables) (James, Witten, Hastie, & Tibshirani, 2023).

Τα κύρια πλεονεκτήματα στη χρήση γραμμικών μοντέλων όπως η Λογιστική Παλινδρόμηση είναι ότι είναι απλά, απαιτούν ελάχιστο χρόνο εκπαίδευσης και είναι γρήγορα στην πρόβλεψή τους (Gupta, 2020). Η απλότητα τους, τα κάνει χρήσιμα στη σύγκριση της απόδοσης πιο πολύπλοκων μοντέλων ως μοντέλο βάσης (baseline) (Silpa & Chandran, 2020). Είναι εύκολα ερμηνεύσιμα και κατανοητά στη σχέση μεταξύ των ανεξάρτητων μεταβλητών και των εξαρτημένων και είναι ιδιαίτερα χρήσιμα όταν υπάρχει πολύ μεγάλος όγκος δεδομένων. Τέλος, σε δεδομένα με μεγαλύτερο αριθμό παραμέτρων, τα γραμμικά μοντέλα, έχουν υψηλή προβλεπτική ικανότητα (Sarker, 2021).

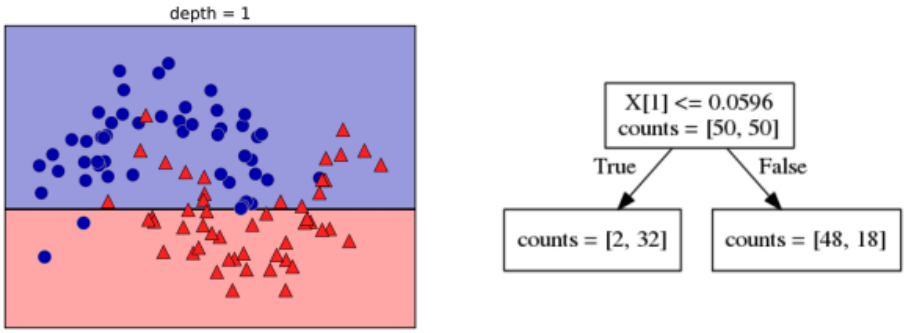
Τα μειονεκτήματα των γραμμικών μοντέλων και συγκεκριμένα η Λογιστική Παλινδρόμηση, είναι ότι έχουν περιορισμένες δυνατότητες στην ανάλυση μη γραμμικών σχέσεων, όπως επίσης και σε δεδομένα με μεταβλητές που είναι συσχετισμένες καθώς είναι δύσκολο να ερμηνεύσουν τη βαρύτητα των συντελεστών. Σε δεδομένα με λιγότερες διαστάσεις έχουν χαμηλότερη απόδοση γενίκευσης σε σχέση με πιο πολύπλοκα μοντέλα (Gupta, 2020), (Müller & Guido, 2017).

3.2.2. Δέντρα Αποφάσεων (Decision Trees)

Τα Δέντρα Αποφάσεων είναι μια μη παραμετρική εποπτευόμενη μέθοδος μάθησης που χρησιμοποιείται για ταξινόμηση και παλινδρόμηση. Ο στόχος είναι να δημιουργηθεί ένα μοντέλο σε δομή δέντρου που προβλέπει την τιμή μιας μεταβλητής στόχου μαθαίνοντας απλούς κανόνες απόφασης που συνάγονται από τα χαρακτηριστικά δεδομένων. Ουσιαστικά τα Δέντρα Αποφάσεων μαθαίνουν μέσα από μια ιεραρχία ερωτήσεων αν/αλλιώς (if/else) να οδηγηθούν σε μια απόφαση με τις λιγότερες δυνατές ερωτήσεις ή αλλιώς split tests (Mitchell, 1997).

Για την δόμηση του Δέντρου Αποφάσεων, ο αλγόριθμος αναζητεί από όλα τα πιθανά split tests, αυτό που περιγράφει καλύτερα τη μεταβλητή στόχο. Η διαδικασία επαναλαμβάνεται παράγοντας ένα δυαδικό δέντρο αποφάσεων μέχρις ότου η απόφαση καταλήξει σε μια δυνατή τιμή. Κάθε επαναληπτική διαδικασία ονομάζεται φύλλο (leaf)

του δέντρου ενώ όταν το φύλλο καταλήξει να έχει μόνο μια δυνατή τιμή ονομάζεται αγνό (pure) (James, Witten, Hastie, & Tibshirani, 2023).



Εικόνα 10: Δέντρο Αποφάσεων με βάθος 1 ερώτηση (αριστερά) και το αντίστοιχο δέντρο (δεξιά). Πηγή: (Müller & Guido, 2017)

Τα Δέντρων Αποφάσεων είναι σχετικά εύκολο να εφαρμοστούν και να οπτικοποιηθούν γεγονός που οδηγεί σε υψηλό βαθμό ερμηνευσιμότητας (James, Witten, Hastie, & Tibshirani, 2023). Επιπρόσθετα, μπορούν να ανιχνεύουν μη γραμμικές σχέσεις και πολύπλοκες αλληλεπιδράσεις μεταξύ των μεταβλητών ενώ παρέχουν κατάταξη των μεταβλητών που επηρεάζουν περισσότερο την πρόβλεψη. Όσον αφορά στο πρακτικό κομμάτι της διαχείρισης των μεταβλητών, τα Δέντρα Αποφάσεων μπορούν να διαχειριστούν συνεχείς, κατηγορικές ή δυαδικές μεταβλητές χωρίς να είναι απαραίτητη η κανονικοποίηση τους ή η χρήση ψευδομεταβλητών. Επίσης, μπορούν να διαχειριστούν ελλείπουσες τιμές ακολουθώντας διαφορετικό φύλλο αν μια τιμή μεταβλητής λείπει (Müller & Guido, 2017).

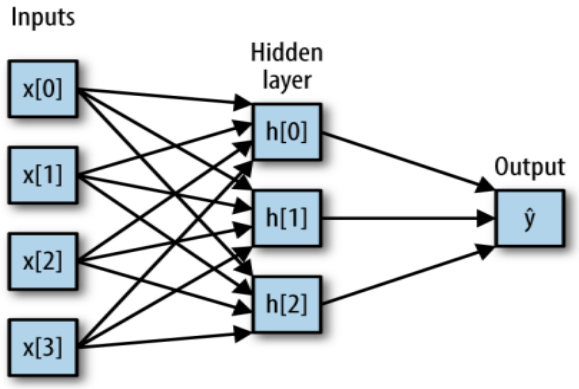
Το βασικότερο μειονέκτημα των Δέντρων Αποφάσεων είναι ότι ακόμα και μετά από τροποποιήσεις, εξακολουθούν να εμφανίζουν υπερπροσαρμογή στα δεδομένα εκπαίδευσης με αποτελέσμα τη χαμηλή γενίκευση όταν τροφοδοτούνται με νέα δεδομένα (Mitchell, 1997). Επίσης, δεν έχουν μεγάλη συνοχή καθώς μια μικρή αλλαγή στα δεδομένα μπορεί να προκαλέσει μεγάλη αλλαγή στη δομή του δέντρου ανάλυση (James, Witten, Hastie, & Tibshirani, 2023).

3.2.3. Νευρωνικά Δίκτυα (Neural Networks)

Τα Νευρωνικά Δίκτυα παρομοιάζονται και εμπνέονται από τη δομή και τη λειτουργία των νευρικών δικτύων του ανθρώπινου εγκεφάλου. Συγκεκριμένα, τα τεχνητά νευρωνικά

δίκτυα μπορούν να χρησιμοποιηθούν σε προβλήματα αναγνώρισης προτύπων, ταξινόμησης, παλινδρόμησης και άλλων πολύπλοκων αναλύσεων δεδομένων (Mitchell, 1997).

Η λειτουργία του Νευρωνικού Δικτύου, βασίζεται στο να υπολογίζει τα σταθμισμένα αθροίσματα σε μια επαναληπτική διαδικασία όπου πρώτα υπολογίζει κρυφές μεταβλητές ως ενδιάμεσο βήμα υπολογισμού πριν καταλήξει στο τελικό αποτέλεσμα. Το μοντέλο αυτό έχει να μάθει πολλούς περισσότερους συντελεστές καθώς υπάρχει ένας μεταξύ κάθε ανεξάρτητης μεταβλητής και κρυφού επιπέδου και ένας μεταξύ κρυφού επιπέδου και εξαρτημένης μεταβλητής (Εικόνα 11). Το κρυφό επίπεδο, πραγματοποιεί υπολογισμούς που είναι μη γραμμικοί μετασχηματισμοί των γραμμικών συνδυασμών των ανεξάρτητων μεταβλητών, επομένως δεν είναι άμεσα ορατό στη διαδικασία. Οι μετασχηματισμοί αυτοί δεν είναι ορισμένοι εξ αρχής και μαθαίνονται κατά τη διαδικασία εκπαίδευσης του δικτύου. (James, Witten, Hastie, & Tibshirani, 2023).



Εικόνα 11: Γραφική απεικόνιση Νευρωνικού Δικτύου με ένα κρυφό επίπεδο.
Πηγή: (Müller & Guido, 2017)

Μια σημαντική παράμετρος που πρέπει να οριστεί από το χρήστη, είναι ο αριθμός των κόμβων στο κρυφό επίπεδο. Αυτοί μπορεί να κυμαίνονται από μερικές δεκάδες μέχρι και χιλιάδες για να περιγράψουν πολύ σύνθετα δεδομένα ενώ μπορούν να προστεθούν περισσότερα κρυφά επίπεδα. Η ύπαρξη πολλών τέτοιων κρυφών επιπέδων για τον υπολογισμό των Νευρωνικών Δικτύων, ενέπνευσε τον όρο Deep Learning (Müller & Guido, 2017).

Το βασικό πλεονέκτημα των Νευρωνικών Δικτύων είναι ότι έχουν την ικανότητα να συλλέξουν πληροφορίες που εμπεριέχονται σε μεγάλους όγκους δεδομένων και να

χτίσουν ιδιαίτερα πολύπλοκα μοντέλα. Αν υπάρχουν αρκετά δεδομένα, αρκετός χρόνος υπολογισμού και κατάλληλη προσαρμογή των παραμέτρων, τα Νευρωνικά Δίκτυα μπορούν να έχουν καλύτερα αποτελέσματα σε σχέση με άλλες μεθόδους μηχανικής μάθησης. Τα πλεονεκτήματα αποτελούν κάποιες φορές μειονεκτήματα καθώς τα μεγάλα και σύνθετα Νευρωνικά Δίκτυα απαιτούν μεγάλο χρόνο εκπαίδευσης, μεγάλο αριθμό δεδομένων και προσεκτική προ επεξεργασία. Τέλος, παρουσιάζουν χαμηλή προβλεπτική ικανότητα όταν τα δεδομένα δεν είναι ομοιογενή (Sarker, 2021), (Mitchell, 1997).

3.2.4. Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

Οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) είναι ένας αλγόριθμος όπου κάθε στοιχείο του συνόλου δεδομένων σχεδιάζεται ως σημείο σε χώρο n διαστάσεων (όπου n είναι ο αριθμός των χαρακτηριστικών) με την τιμή κάθε χαρακτηριστικού να είναι η τιμή μιας συγκεκριμένης συντεταγμένης. Έπειτα, η ταξινόμηση πραγματοποιείται εντοπίζοντας το βέλτιστο υπερεπίπεδο (hyper-plane) που διαχωρίζει το σύνολο των δεδομένων με βάση το μέγιστο περιθώριο των κλάσεων (Silpa & Chandran, 2020). Ο στόχος των SVM είναι να βρεθεί ένα υπερεπίπεδο που χωρίζει καλύτερα τα δεδομένα στις δύο κατηγορίες. Όταν ο χώρος είναι δισδιάστατος το υπερεπίπεδο είναι μια γραμμή, όταν είναι τρισδιάστατος είναι ένα επίπεδο ενώ σε μεγαλύτερες διαστάσεις είναι δύσκολο να οπτικοποιηθεί (James, Witten, Hastie, & Tibshirani, 2023).

Οι Μηχανές Διανυσμάτων Υποστήριξης αποτελούν επέκταση των Γραμμικών (Linear) SVM όπου ο χώρος των μεταβλητών μεγαλώνει με συγκεκριμένο τρόπο χρησιμοποιώντας πυρήνες (kernels) (James, Witten, Hastie, & Tibshirani, 2023). Τα kernels είναι μια συνάρτηση που επιτρέπει να μετασχηματιστούν τα δεδομένα εισόδου από τον αρχικό χώρο σε έναν χώρο υψηλότερων διαστάσεων χωρίς να γίνει ο υπολογισμός της νέας και μεγαλύτερης απεικόνισης (Müller & Guido, 2017). Οι πιο διαδεδομένοι τύποι kernel που μπορούν να χρησιμοποιηθούν στα SVM είναι τα Linear, Polynomial, Radial Basis Function (RBF) και Sigmoid (Sarker, 2021).

Κατά την εκπαίδευση, τα SVM μαθαίνουν τη σημαντικότητα των δεδομένων που απεικονίζουν το όριο μεταξύ των δύο κλάσεων. Τυπικά, μόνο ένα μέρος των δεδομένων εκπαίδευσης καθορίζει τα όρια της απόφασης και είναι αυτά που βρίσκονται στα όρια μεταξύ των κλάσεων και ονομάζονται διανύσματα υποστήριξης (Support Vectors). Για

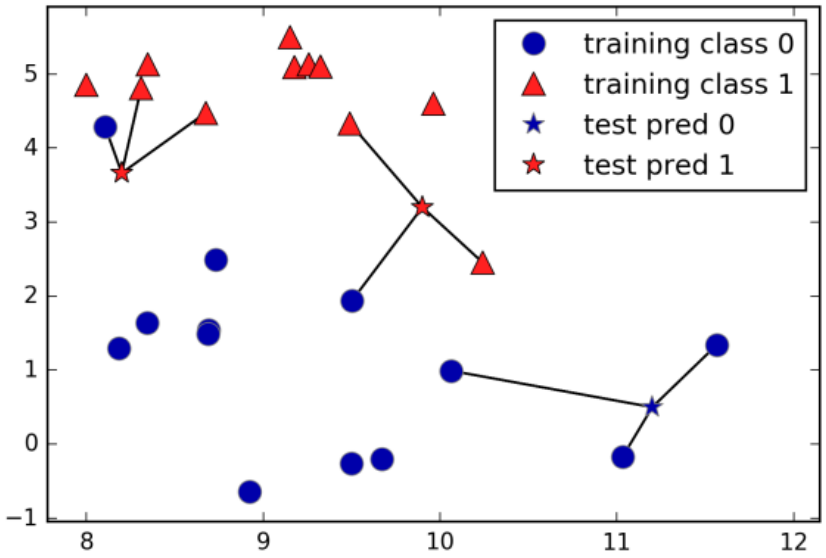
να γίνει μια πρόβλεψη μετριέται η απόσταση από κάθε διάνυσμα υποστήριξης και της σημαντικότητας τους (Müller & Guido, 2017).

Τα SVM είναι αλγόριθμοι που παρουσιάζουν καλά αποτελέσματα σε ταξινομήσεις με σύνθετα όρια ακόμα και αν υπάρχουν λίγες μεταβλητές ενώ αποδίδουν εξίσου καλά σε χαμηλών διαστάσεων και υψηλών διαστάσεων δεδομένα. Ένα βασικό μειονέκτημα είναι ότι όσο το σύνολο των δεδομένων αυξάνει τότε ο χρόνος υπολογισμού αυξάνεται πολύ (Müller & Guido, 2017). Επίσης, τα SVM απαιτούν προσεκτική προ-επεξεργασία δεδομένων και ρύθμιση των παραμέτρων ενώ τέλος δεν είναι εύκολη η επεξήγηση των προβλέψεων και για το λόγο αυτόν θεωρούνται από κάποιους «μαύρα κουτιά» (Van Haver, 2017).

3.2.5. Κ-πλησιέστερων γειτόνων (K-Nearest Neighbors)

Ένας σχετικά απλός αλγόριθμος μηχανικής μάθησης για την πρόβλεψη της απώλειας πελατών είναι των Κ-πλησιέστερων γειτόνων. Ο αλγόριθμος υποθέτει ότι όλες οι παρατηρήσεις αντιστοιχίζονται σε σημεία στο n διαστάσεων χώρο και κάθε νέα παρατήρηση ταξινομείται βρίσκοντας την εγγύτερη παρατήρηση (nearest neighbor) από τα δεδομένα εκπαίδευσης. Για τον υπολογισμό της απόστασης μεταξύ της προς ταξινόμηση παρατήρησης και της ήδη ταξινομημένης παρατήρησης χρησιμοποιείται η Ευκλείδεια απόσταση (Mitchell, 1997).

Στην πιο απλή της μορφή, ο αλγόριθμος υπολογίζει μόνο την πιο κοντινή παρατήρηση, ωστόσο μπορεί να οριστεί Κ αριθμός γειτόνων για την ταξινόμηση. Όταν υπάρχουν παραπάνω από μια γειτονική παρατήρηση, τότε ο αλγόριθμος μετράει πόσες γειτονικές παρατηρήσεις ανήκουν στην κλάση 0 και πόσες στην κλάση 1 και η υπό ταξινόμηση παρατήρηση αντιστοιχίζεται στην κλάση με τη μεγαλύτερη συχνότητα. Ένα παράδειγμα ταξινόμησης με K=3 φαίνεται στην εικόνα 12 (Müller & Guido, 2017). Ο αριθμός των Κ γειτόνων έχει επίπτωση στη διακύμανση των αποτελεσμάτων ταξινόμησης και τη μεροληψία του αλγορίθμου. Πολύ μικρός αριθμός Κ προσδίδει μεγάλη διακύμανση και χαμηλή μεροληψία, ενώ μεγάλος αριθμός Κ το αντίθετο καθώς ο αλγόριθμος γίνεται λιγότερο ευέλικτος και τα όρια της απόφασης ταξινόμησης γίνονται σχεδόν γραμμικά (James, Witten, Hastie, & Tibshirani, 2023).



Εικόνα 12: Ταξινόμηση K-Nearest Neighbors με K=3.
Πηγή: (Müller & Guido, 2017)

Το μεγαλύτερο πλεονέκτημα του αλγόριθμου K-NN είναι ότι είναι εύκολα κατανοητός και συχνά δίνει λογικά αποτελέσματα χωρίς πολλές τροποποιήσεις. Η απλότητα και ευκολία του αλγόριθμου τον καθιστά κατάλληλο για βασικό μέτρο σύγκρισης προτού δοκιμαστούν πιο προχωρημένες μέθοδοι ταξινόμησης. Στον αντίποδα, η εγγενής μεροληψία (bias) του αλγόριθμου που υποθέτει ότι μια παρατήρηση θα ταξινομηθεί βάσει της εγγύτητας της στις ήδη ταξινομημένες παρατηρήσεις, κάνει τον αλγόριθμο ευαίσθητο σε ακραίες τιμές (outliers). Επίσης, όταν το σύνολο δεδομένων εκπαίδευσης είναι αρκετά μεγάλο, ο αλγόριθμός είναι αργός στις προβλέψεις ενώ παρουσιάζει φτωχά αποτελέσματα όταν υπάρχουν πολλές μεταβλητές (Müller & Guido, 2017).

3.2.6. Naïve Bayes

Η μέθοδος Naïve Bayes εμφανίζεται συχνά στη βιβλιογραφία που συνδυάζει την εύκολη κατανόηση και εφαρμογή με ικανοποιητική ακρίβεια στην πρόβλεψη απώλειας πελατών (Van Haver, 2017). Ο αλγόριθμος υπολογίζει την πιθανότητα ότι ένα γεγονός θα συμβεί, βάσει της πρότερης γνώσης των μεταβλητών που σχετίζονται με αυτό και υποθέτει ότι οι μεταβλητές που χρησιμοποιούνται είναι ανεξάρτητες μεταξύ τους (Sabbeh, 2018).

Ο αλγόριθμος ταξινόμησης Naïve Bayes βασίζεται στο θεώρημα του Bayes και την (αφελή) υπόθεση ότι τα χαρακτηριστικά των δεδομένων είναι ανεξάρτητα μεταξύ τους.

Η υπόθεση της ανεξαρτησίας μεταξύ των μεταβλητών σημαίνει ότι η πιθανότητα της εξαρτημένης μεταβλητής μπορεί να εκφραστεί ως το γινόμενο της κάθε ανεξάρτητης μεταβλητής (Mitchell, 1997). Για να γίνει μια πρόβλεψη, ένα σημείο των δεδομένων συγκρίνεται στατιστικά με κάθε κλάση και ο αλγόριθμος προβλέπει την κλάση στην οποία ταιριάζει περισσότερο. Υπάρχουν τρεις παραλλαγές του αλγορίθμου Naïve Bayes, Gaussian για συνεχείς μεταβλητές, Bernoulli για δυαδικές μεταβλητές και Multinomial για διακριτές μεταβλητές (Sarker, 2021).

Στα πλεονεκτήματα του αλγορίθμου συγκαταλέγονται η ευκολία κατανόησης, η απλότητα στον υπολογισμό των μεταβλητών και η ταχύτητα εκπαίδευσης και πρόβλεψης. Είναι χρήσιμος ως μέτρο σύγκρισης με πιο προχωρημένες μεθόδους και τέλος παρουσιάζει καλά αποτελέσματα με δεδομένα πολλών διαστάσεων. Βασικό μειονέκτημα αποτελεί η θεμελιώδης υπόθεση ότι οι μεταβλητές είναι ανεξάρτητες μεταξύ τους καθώς στον πραγματικό κόσμο κάτι τέτοιο δεν ισχύει πάντα. Επιπρόσθετα, έχουν συγκριτικά χαμηλότερη απόδοση γενίκευσης σε σχέση με άλλους αλγορίθμους ταξινόμησης (Müller & Guido, 2017).

3.2.7. Αλγόριθμοι ταξινόμησης συνόλων (ensemble)

Τέλος, αξίζει να αναφερθούν οι αλγόριθμοι που εμπίπτουν στην κατηγορία ταξινόμησης συνόλων που στόχο έχουν να βελτιώσουν την προβλεπτική δύναμη των απλών μεθόδων ταξινόμησης. Οι αλγόριθμοι αυτοί από μόνοι τους μπορεί να οδηγήσουν σε μέτρια αποτελέσματα για αυτό και ονομάζονται αδύναμοι ταξινομητές. (Sabbeh, 2018), (James, Witten, Hastie, & Tibshirani, 2023).

Ο αλγόριθμος Τυχαίων Δέντρων (Random Forests) ανήκει στην κατηγορία ensemble μεθόδων και αντιμετωπίζει το πρόβλημα της υπερπροσαρμογής των δεδομένων εκπαίδευσης στα Δέντρα Αποφάσεων. Η μέθοδος προσαρμόζει πολλά Δέντρα Αποφάσεων παράλληλα σε διαφορετικά σετ δεδομένων και χρησιμοποιώντας τη μέθοδο της πλειοψηφίας ή του μέσου όρου καταλήγει σε αποτελέσματα με καλύτερη ακρίβεια πρόβλεψης (Sarker, 2021).

Τα Δέντρα Αποφάσεων Βαθμιαίας Ενίσχυσης (Gradient Boosting Decision Trees) είναι μια ensemble μέθοδος ταξινόμησης που συνδυάζει πολλαπλά Δέντρα Αποφάσεων με

στόχο την παραγωγή ενός ισχυρότερου μοντέλου. Τα δέντρα δομούνται διαδοχικά και κάθε δέντρο προσπαθεί να διορθώσει τα λάθη του προηγούμενου. Τα Gradient Boosting Decision Trees βασίζονται στο εκτεταμένο κλάδεμα (pre-pruning) που οδηγεί σε δέντρα με λίγα κλαδιά, γεγονός που κάνει τις απαιτήσεις σε υπολογιστική ισχύ και μνήμη λιγότερη και παράγει αποτελέσματα γρηγορότερα (James, Witten, Hastie, & Tibshirani, 2023).

Ο αλγόριθμος Bagging ή αλλιώς Bootstrap Aggregation είναι επίσης μια ensemble μέθοδος ταξινόμησης που βασικό στόχο έχει να μειώσει τη διακύμανση και συνήθως χρησιμοποιείται στα Δέντρα Αποφάσεων. Η λογική του αλγορίθμου είναι να παίρνει πολλά υποσύνολα δεδομένων, να δομεί ένα ξεχωριστό μοντέλο πρόβλεψης και μετά να παίρνει απόφαση βάσει της πλειοψηφίας των αποτελεσμάτων πρόβλεψης μειώνοντας ταυτόχρονα τη διακύμανση. (Müller & Guido, 2017).

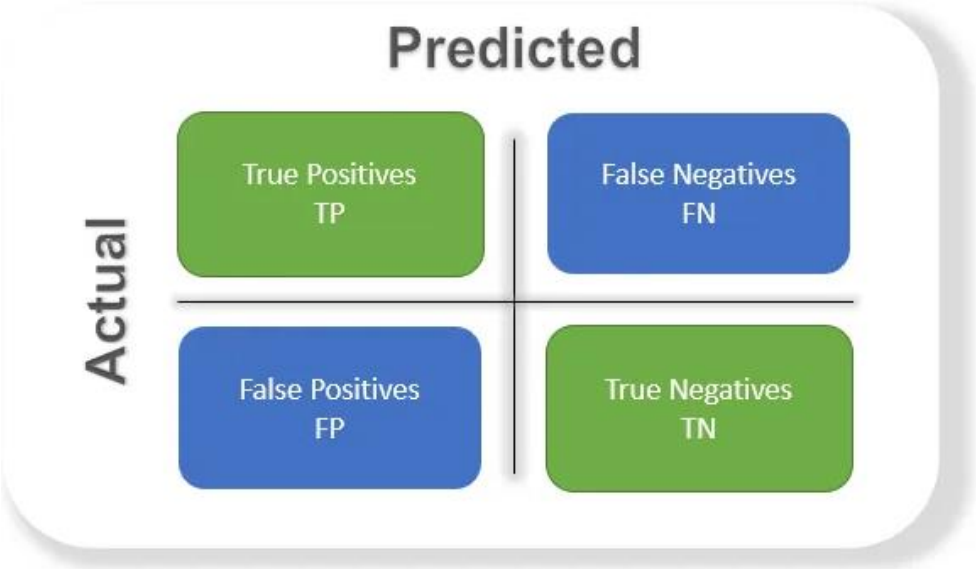
3.3. Μετρικές Αξιολόγησης (Evaluation Metrics)

Οι μετρικές αξιολόγησης αναφέρονται σε κριτήρια ή μεθόδους που χρησιμοποιούνται για να αξιολογήσουν την απόδοση, την ποιότητα ή άλλες πτυχές ενός συστήματος, μιας διαδικασίας ή ενός προϊόντος. Ομοίως και στη μηχανική μάθηση, χρησιμοποιούνται για να μετρήσουν την απόδοση ενός μοντέλου μηχανικής μάθησης σε συγκεκριμένες εργασίες. Ανάλογα με τον τύπο της εργασίας (π.χ. ταξινόμηση, παλινδρόμηση, σύσταση), υπάρχουν διάφορες μετρικές που μπορούν να χρησιμοποιηθούν (Mutuvi, 2019).

Ένα εργαλείο που χρησιμοποιείται για την αξιολόγηση της απόδοσης ενός αλγορίθμου ταξινόμησης είναι ο πίνακας σύγχυσης (confusion matrix). Το συγκεκριμένο εργαλείο παρουσιάζει την πραγματική κατηγορία των δειγμάτων σε σχέση με τις προβλέψεις που δίνει ο αλγόριθμος και έχει τη μορφή της εικόνας 13, όπου:

- True Positive (TP): Ο αριθμός των δειγμάτων που προβλέφθηκαν σωστά ως κλάση 1.
- False Positive (FP): Ο αριθμός των δειγμάτων που προβλέφθηκαν λανθασμένα ως κλάση 1, αλλά στην πραγματικότητα είναι κλάση 0 (Λάθος Θετικά). Στην Επιστήμη της Στατιστικής, τα λάθη αυτά ονομάζονται λάθη τύπου I.
- True Negative (TN): Ο αριθμός των δειγμάτων που προβλέφθηκαν σωστά ως κλάση 0.

- False Negative (FN): Ο αριθμός των δειγμάτων που προβλέφθηκαν λανθασμένα ως κλάση 0, αλλά στην πραγματικότητα είναι κλάση 1 (Λάθος Αρνητικά) Στην Επιστήμη της Στατιστικής, τα λάθη αυτά ονομάζονται λάθη τύπου II (Gattermann-Itschert & Thonemann, 2022).



Εικόνα 13: Σχηματική απεικόνιση του πίνακα σύγχυσης (confusion matrix).
Πηγή: (Nighania, 2018)

Από τον πίνακα σύγχυσης δύναται να υπολογιστούν διάφορες μετρικές αξιολόγησης, από τις οποίες οι πιο συνηθισμένες για προβλήματα δυαδικής ταξινόμησης είναι οι ακόλουθες:

- Accuracy που αποδίδει τον αριθμό των σωστών προβλέψεων δια το συνολικό αριθμό προβλέψεων (Silpa & Chandran, 2020). Υπολογίζεται από τον τύπο:

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

- Precision που αποδίδει τον αριθμό των σωστών θετικών προβλέψεων δια το συνολικό αριθμό προβλέψεων. Υπολογίζεται από τον τύπο:

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

Αυτός ο τύπος μέτρησης της ακρίβειας χρησιμοποιείται όταν ο στόχος είναι να περιοριστεί ο αριθμός των λαθών τύπου I (False Positive) και συναντάται στη βιβλιογραφία ως Τιμή Θετικών Προβλέψεων (Positive Predictive Value)

- Recall που αποδίδει το ποσοστό των σωστών θετικών προβλέψεων σε σχέση με το συνολικό αριθμό πραγματικά θετικών. Υπολογίζεται από τον τύπο:

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

Η ανάκληση είναι μια μετρική αξιολόγησης όταν χρειάζεται να αναγνωριστούν όλα τα θετικά δείγματα, δηλαδή είναι σημαντικό να αποφευχθούν τα λάθη τύπου II (False Negative). Εναλλακτικοί ορισμοί είναι ευαισθησία (sensitivity), ποσοστό επιτυχίας (hit rate) ή ποσοστό σωστά θετικών (true positive rate)

- Specificty που αποδίδει το ποσοστό των σωστών αρνητικών προβλέψεων σε σχέση με το συνολικό αριθμό πραγματικά αρνητικών. Υπολογίζεται από τον τύπο:

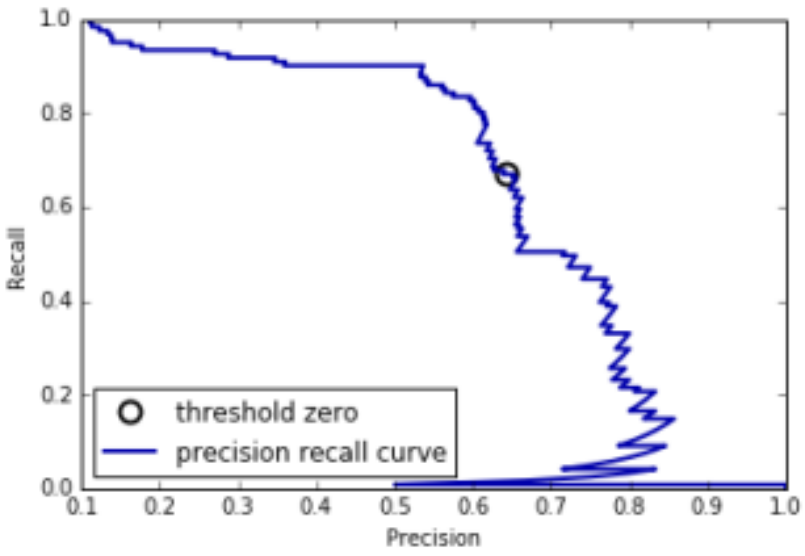
  $$\frac{TN}{TN + FP}$$

  Στη βιβλιογραφία συναντάται και ως ποσοστό σωστά αρνητικών (true negative rate)
- F_1-Score πρόκειται για ένα συνδυασμό της ακρίβειας και της ανάκλησης, που βοηθά στην αξιολόγηση της συνολικής απόδοσης και υπολογίζεται από τον τύπο:

  $$\frac{2 \text{ x Precision x Recall}}{\text{Precision } + \text{Recall}}$$

(Chen, Hu, & Hsieh, 2014) (Müller & Guido, 2017).

Μια ακόμη μέθοδος που μπορεί να προσαρμόσει την αναλογία ακρίβειας και ανάκλησης είναι η καμπύλη ακρίβειας-ανάκλησης (precision-recall curve). Η καμπύλη αυτή απεικονίζει την αναλογία ακρίβειας και ανάκλησης για ένα πρόβλημα ταξινόμησης και ο χρήστης μπορεί να επιλέξει το κατώφλι/όριο (threshold) που χρειάζεται. Για παράδειγμα, μπορεί για ένα πρόβλημα να χρειάζεται ο αλγόριθμος να χάνει μόνο το 10% των θετικών δειγμάτων οπότε το κατώφλι ανάκλησης πρέπει να οριστεί στο 90%. Φυσικά είναι δυνατόν να επιτύχει ένας αλγόριθμος 90% ανάκληση αλλά ταυτόχρονα θα πρέπει να έχει και λογικό ποσοστό ακρίβειας για να έχει νόημα η ταξινόμηση και αυτή η ισορροπία αποτελεί τη μεγαλύτερη δυσκολία. Η καμπύλη ακρίβειας-ανάκλησης δείχνει ακριβώς αυτό, το αντάλλαγμα σε ακρίβεια ή ανάκληση για κάθε πιθανό συνδυασμό σε ταξινομημένη σειρά ώστε να προκύψει μια καμπύλη όπως φαίνεται στο παράδειγμα της εικόνας 14. Το σημείο 0, δείχνει το βέλτιστο κατώφλι που χρησιμοποιεί ο αλγόριθμος ταξινόμησης (Müller & Guido, 2017).

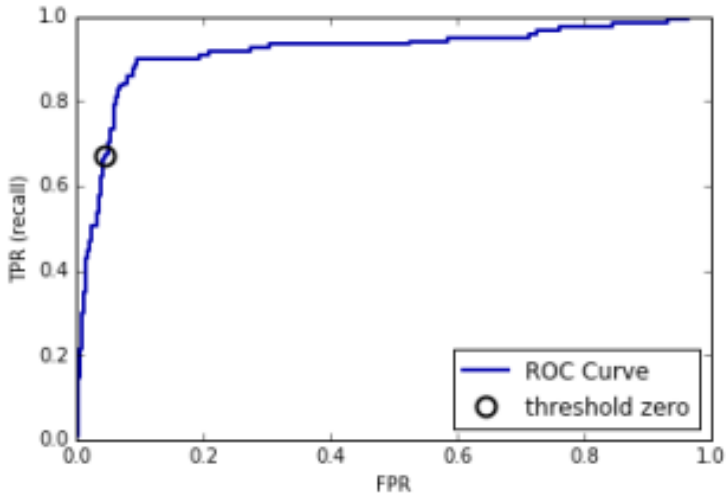


Εικόνα 14: Καμπύλη ακρίβειας - ανάκλησης.
Πηγή: (Müller & Guido, 2017)

Ένας ακόμη δείκτης που χρησιμοποιείται συχνά για να ελεγχθεί η συμπεριφορά ενός αλγόριθμου ταξινόμησης για διαφορετικά όρια ακρίβειας, αποτελεί η καμπύλη ROC (Receiver Operating Characteristics). Η καμπύλη αυτή δείχνει την αναλογία μεταξύ FPR (False Positive Rate) και TPR (True Positive Rate). Το TPR είναι η ανάκληση ενώ το FPR ή αλλιώς specificity είναι το ποσοστό των λανθασμένα θετικών προς το σύνολο των αρνητικών, δηλαδή:

$$FPR = \frac{FP}{FP + TN}$$

Το ιδανικό σημείο στην καμπύλη βρίσκεται στην πάνω αριστερή γωνία όπου ο αλγόριθμος προβλέπει υψηλό TPR και ταυτόχρονα κρατώντας χαμηλά το FPR. Παράδειγμα καμπύλης ROC φαίνεται στην εικόνα 15 ενώ ο κύκλος συμβολίζει το βέλτιστο κατώφλι που επιλέγει ο αλγόριθμος (Müller & Guido, 2017).



Εικόνα 15: Καμπύλη ROC.
Πηγή: (Müller & Guido, 2017)

4. Περιγραφική Ανάλυση

Τα δεδομένα της παρούσας μελέτης προέρχονται από μια εξειδικευμένη βιομηχανία μηχανολογικών κατασκευών, προσφέροντας ένα ενδελεχές υλικό για την ανάλυση των πωλήσεων. Πιο συγκεκριμένα, στη βάση δεδομένων της επιχείρησης υπάρχουν στοιχεία για κάθε παραγγελία που καταχωρείται. Τα δεδομένα καλύπτουν πτυχές όπως το προϊόν, το εργοστάσιο κατασκευής και η ημερομηνία παράδοσης του προϊόντος, πληροφορίες σχετικά με τον πελάτη, και άλλα στοιχεία που συμβάλλουν στον πλήρη χαρακτηρισμό και στην κατανόηση των πωλήσεων.

Από τα πληροφοριακά συστήματα της επιχείρησης εξάγονται πληροφορίες σχετικά με το αν ένας πελάτης έχει ως τρόπο πληρωμής πιστωτικό όριο ή όχι, αν έχει πρόσβαση στις ηλεκτρονικές πλατφόρμες Portal και E-shop, χαρακτηριστικά που δείχνουν τον βαθμό αφοσίωσης του πελάτη στην επιχείρηση. Τέλος, ζητήθηκε από το τμήμα πωλήσεων της επιχείρησης να δοθούν πληροφορίες για κάθε πελάτη όσον αφορά στον αριθμό εργαζομένων, τα χρόνια συνεργασίας και το αν είναι εγκαταστάτες μηχανολογικού εξοπλισμού ή όχι. Αν οι πελάτες είναι εγκαταστάτες αποτελεί ένδειξη ότι μπορεί να υπάρξει μακροχρόνια συνεργασία σε σχέση με πελάτες που δεν είναι εγκαταστάτες και μπορεί να αγοράσουν προϊόν μια φορά μόνο για συγκεκριμένο έργο.

Το πλήθος των δεδομένων ανέρχεται στις 21.099 γραμμές (εγγραφές) και εννέα στήλες με 1.716 διαφορετικούς πελάτες για τα έτη 2018 έως 2022. Ακολούθως, αναλύονται οι στήλες του αρχικού πίνακα δεδομένων.

Πίνακας 2: Επεξήγηση στηλών αρχικού πίνακα δεδομένων.

Στήλες αρχικού πίνακα δεδομένων	**Επεξήγηση στηλών**
Product_Code	Κωδικός Προϊόντος
Manufacturer_Code	Εργοστάσιο κατασκευής του προϊόντος
Sales_Person_Code	Κωδικός Πωλητή
Customer_Name_Code	Κωδικός Πελάτη
Country_Name	Χώρα
Total_Value	Συνολική αξία προϊόντος
Has_Credit_Limit	Τύπος δεδομένων αλήθειας (Boolean) για το Πιστωτικό όριο
Delivery_Date	Ημερομηνία παράδοσης προϊόντος
Portal_Access	Πρόσβαση στην ηλεκτρονική πλατφόρμα καταχώρησης παραγγελιών
Company_Size	Μέγεθος επιχείρησης
Installer	Τύπος δεδομένων αλήθειας (Boolean) για την ιδιότητα του εγκαταστάτη του προϊόντος
Years_Of_Collaboration	Χρόνια συνεργασίας με τον Πελάτη
Eshop_Users	Τύπος δεδομένων αλήθειας (Boolean) για τους χρήστες του ηλεκτρονικού καταστήματος
Missing_parts	Ελλείψεις ή ελλαττωματικά εξαρτήματα ανά πελάτη

Τα αρχικά δεδομένα αναλύθηκαν ανά έτος και συνολικά στην πενταετία σε συνάρτηση με την αξία των παραγγελιών ανά:

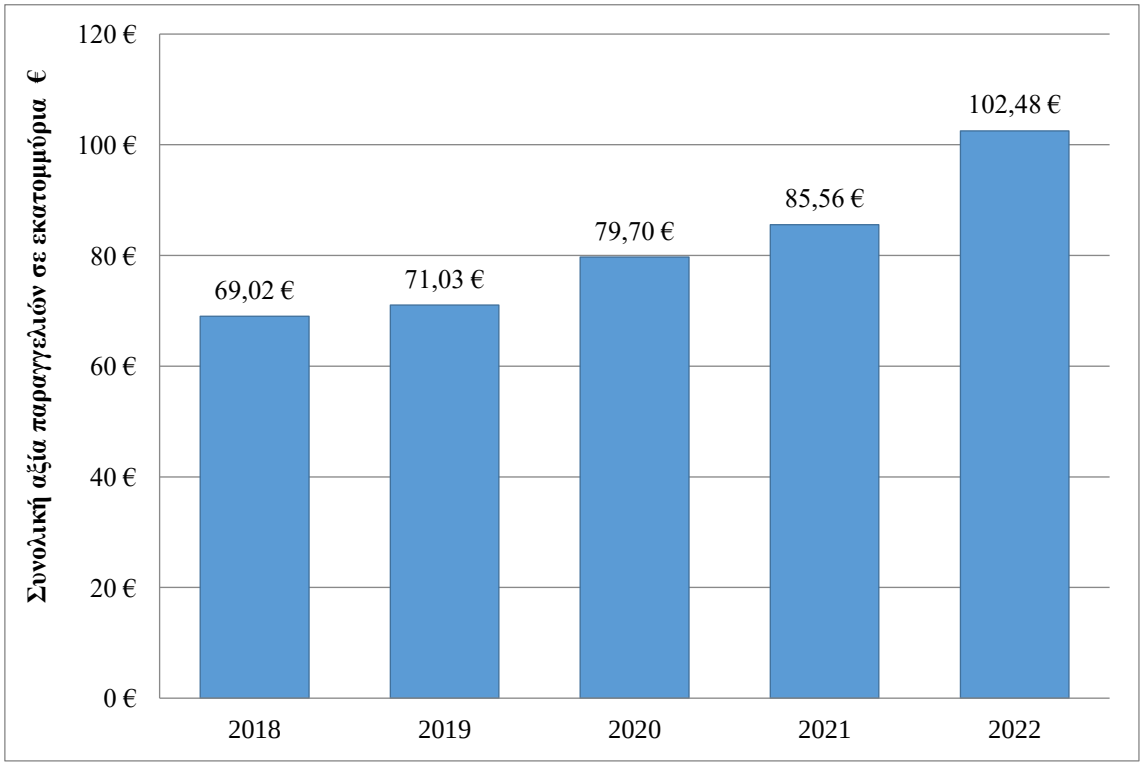
- Πελάτη
- Πωλητή
- Χώρα αγοράς παραγγελίας
- Εργοστάσιο κατασκευής παραγγελίας
- Προϊόν

Συγκεκριμένα, η μέση αξία των παραγγελιών ανέρχεται στα 19.327,49 €, ενώ η τυπική απόκλιση είναι 14.260,87 €. Τέλος, η ελάχιστη τιμή είναι 1.819,22 €, ενώ η μέγιστη είναι 567.194,00 €.

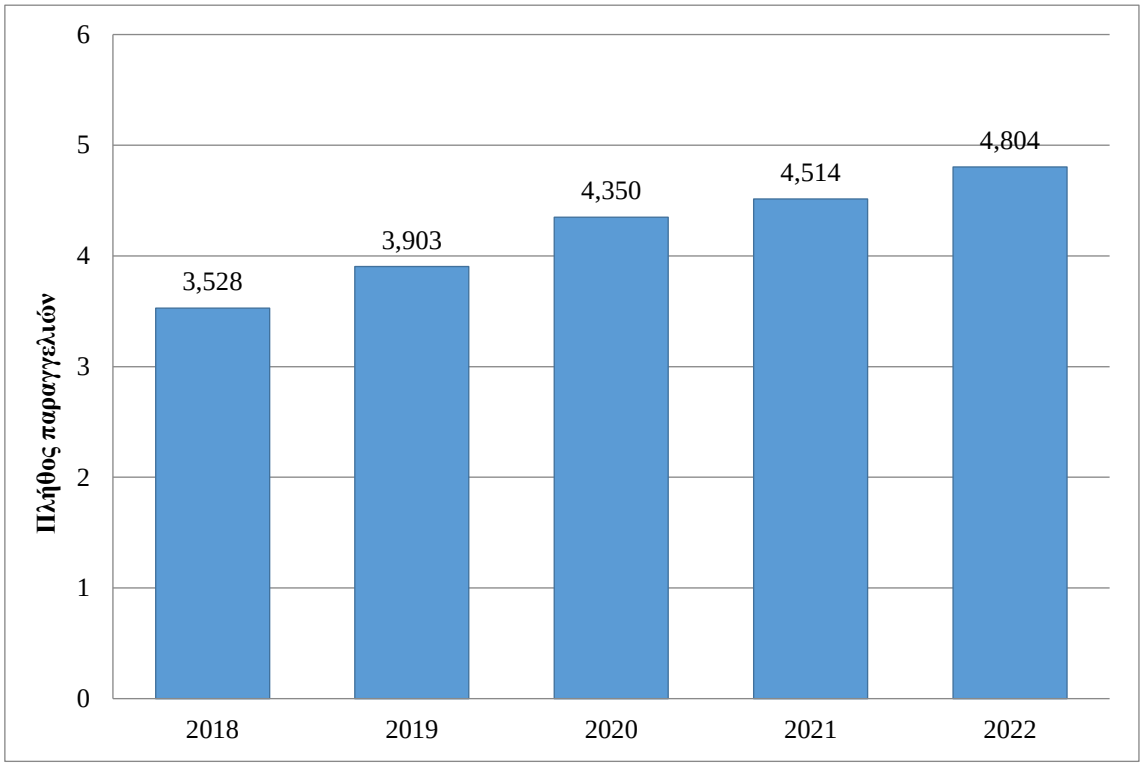
Πίνακας 3: Πλήθος, συνολική αξία παραγγελιών ανά έτος και οι αντίστοιχοι ρυθμοί ανάπτυξης.

Έτος	Πλήθος παραγγελιών	Συνολική αξία παραγγελιών	Ρυθμός ανάπτυξης πλήθους	Ρυθμός ανάπτυξης αξίας
2018	3528	69.023.227,01 €	-	-
2019	3903	71.032.331,32 €	11%	3%
2020	4350	79.698.562,07 €	11%	12%
2021	4514	85.557.597,98 €	4%	7%
2022	4804	102.478.909,55 €	6%	20%
2018 - 2022	**Σ = 21099**	**Σ = 407.790.627,93 €**	**Μ.Ο.: 8%**	**Μ.Ο.: 11%**

Επίσης, όπως φαίνεται από τον πίνακα 3 και από τα διαγράμματα 1 και 2 τόσο η αξία όσο και το αντίστοιχο πλήθος παραγγελιών παρουσιάζουν αυξητική τάση κάθε έτος με μέσο ρυθμό ανάπτυξης πλήθους και αξίας παραγγελιών 8% και 11% αντίστοιχα.

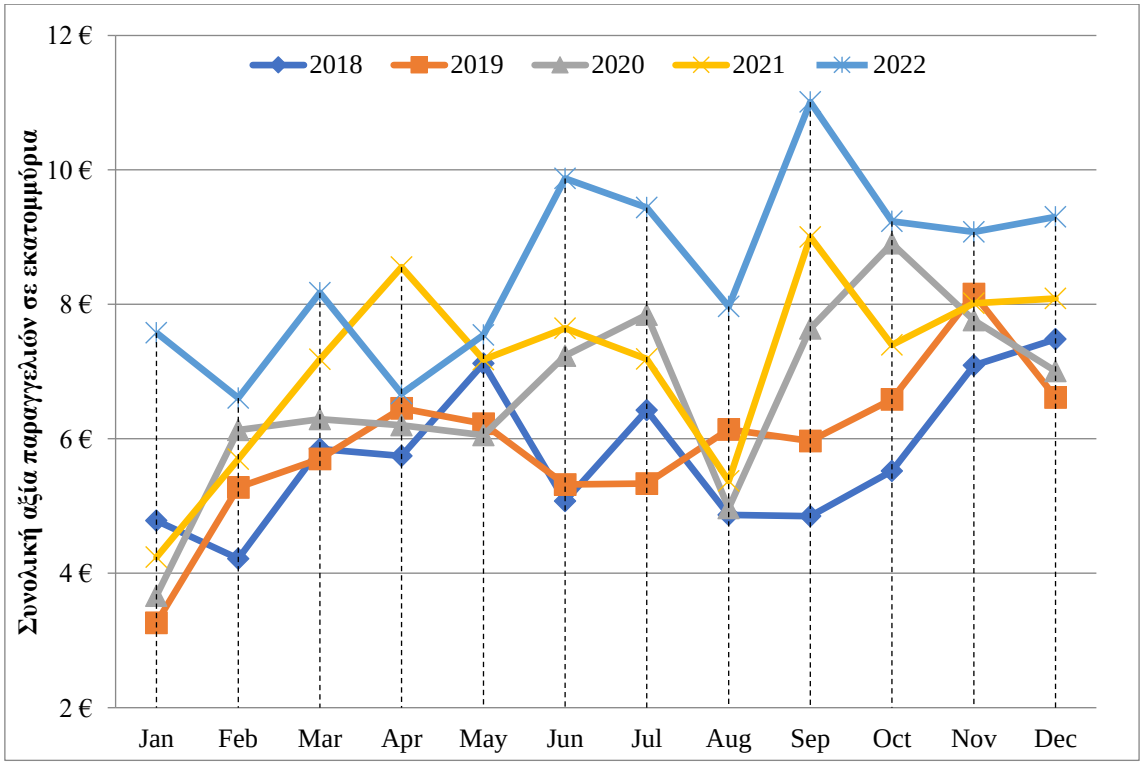


Διάγραμμα 1: Συνολική αξία παραγγελιών ανά έτος σε εκατομμύρια.

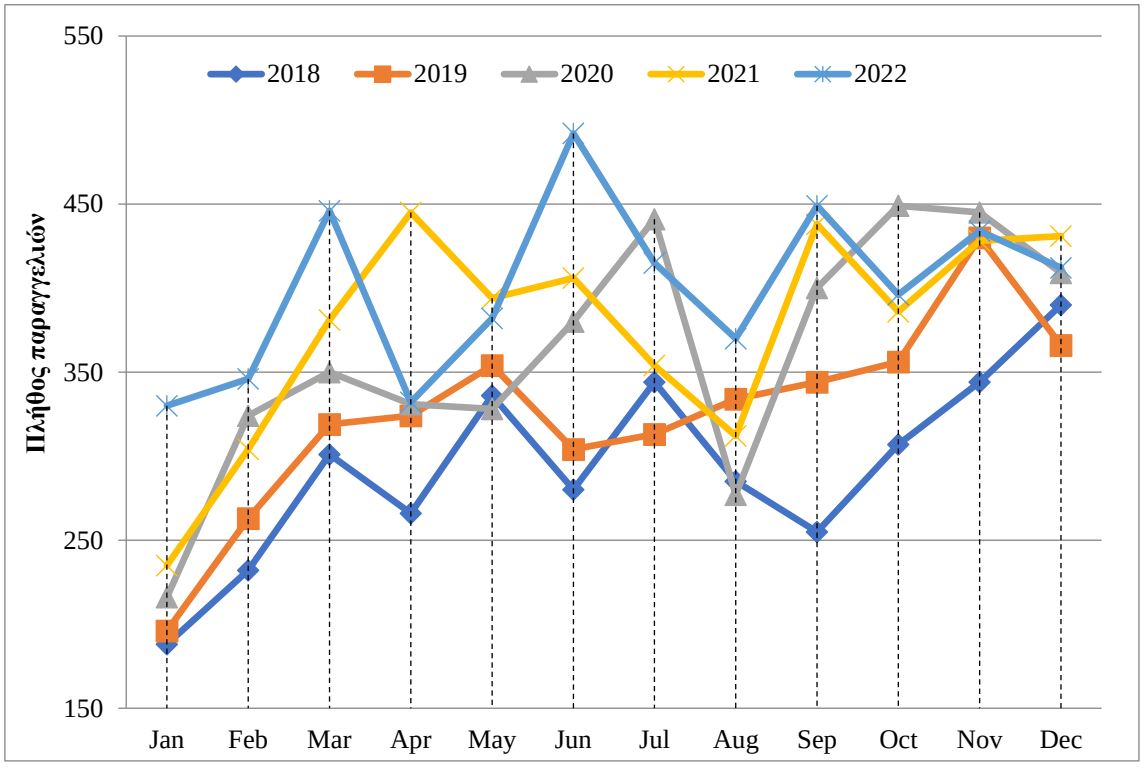


Διάγραμμα 2: Πλήθος παραγγελιών ανά έτος σε χιλιάδες.

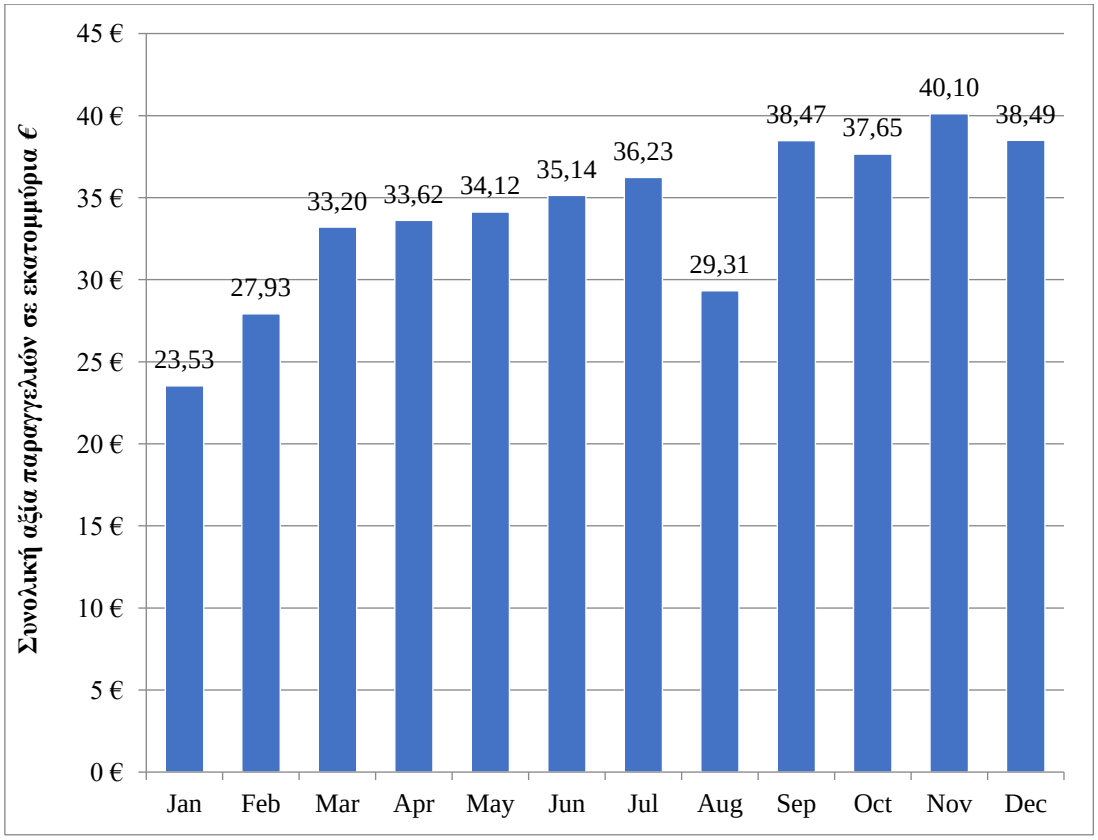
Αναλύοντας τα δεδομένα σε επίπεδο μήνα, οι μήνες Ιανουάριος και Φεβρουάριος δεν παρουσιάζουν μεγάλο αριθμό και αξία παραγγελιών και η κατανομή τους μαζί με αυτή του Αυγούστου, αποτελούν τις χαμηλότερες κατανομές παραγγελιών τόσο στην πενταετία όσο και ανά έτος (Διάγραμμα 3, Διάγραμμα 4 και Διάγραμμα 5). Ο Αύγουστος, συστηματικά παρουσιάζει πτωτική τάση αξίας και πλήθους παραγγελιών (Διάγραμμα 3 και Διάγραμμα 4) καθώς αποτελεί το μήνα που τα εργοστάσια υπολειτουργούν λόγω των καλοκαιρινών αδειών. Επίσης, οι πιο καλοί μήνες ως προς την κατανομή των παραγγελιών λογίζονται οι μήνες από Σεπτέμβριο μέχρι και Δεκέμβριο (Διάγραμμα 3, Διάγραμμα 4 και Διάγραμμα 5). Πρόκειται για την περίοδο πριν το κλείσιμο της χρονιάς και τον υπολογισμό του ετήσιου τζίρου που ταυτίζεται με τις εντατικές προσπάθειες των πωλητών για αύξηση των παραγγελιών.



Διάγραμμα 3: Συνολική αξία παραγγελιών ανά μήνα και ανά έτος.



Διάγραμμα 4: Πλήθος παραγγελιών ανά μήνα και ανά έτος.



Διάγραμμα 5: Συνολική αξία παραγγελιών ανά μήνα για τα έτη αναφοράς 2018 - 2022.

Συνολικά παρατηρείται αύξηση τόσο της συνολικής αξίας των παραγγελιών όσο και του πλήθους των αντίστοιχων παραγγελιών σε όλα τα έτη αναφοράς, ενώ παρουσιάζεται μια σταθερή εποχικότητα με τον Αύγουστο να σημειώνει ετησίως πτώση και από το Σεπτέμβριο μέχρι το Δεκέμβριο να σημειώνεται ετησίως άνοδος (Διάγραμμα 3, Διάγραμμα 4 και Διάγραμμα 5).

4.1. Συνολική αξία παραγγελιών ανά πελάτη

Στην ανάλυση της παρούσας εργασίας, ένας πελάτης θεωρείται ότι έχει αποχωρήσει (churn) από το πελατολόγιο αν δεν έχει κάποια νέα παραγγελία σε ένα ημερολογιακό έτος. Ο κανόνας αυτός ακολουθείται και από την επιχείρηση για τη στατιστική ανάλυση χαμένων πελατών και νέων πελατών ανά έτος.

Για την καλύτερη ανάλυση του customers churn του υπό εξέταση δείγματος, παρατίθεται ο πίνακας 4, στον οποίο φαίνονται τα υψηλά επίπεδα απώλειας πελατών με ποσοστά άνω του 43% ανά περίοδο αναφοράς. Το 2020 ήταν η καλύτερη χρονιά σύμφωνα με τα στοιχεία, καθώς ήταν η χρονιά με το μεγαλύτερο αριθμό πελατών, ήτοι 734 και

ταυτοχρόνως η χρονιά με το μεγαλύτερο πλήθος νέων πελατών σε ποσοστό 59% του συνόλου. Επίσης, σε σχέση με τη προηγούμενη χρονιά (2019) αποχώρησαν μόνο 261 με ποσοστό απώλειας 46%. Σε αντίθεση με το 2021 που ήταν η χειρότερη χρονιά με το μεγαλύτερο ποσοστό απώλειας πελατών 52%, μείωση του συνολικού πλήθους των πελατών από 734 (το 2020) στους 675 πελάτες (Πίνακας 4 και Διάγραμμα 6).

Στη γενική εικόνα των ετών του δείγματος, σημειώνεται ανοδική η πορεία του συνολικού πλήθους των πελατών ανά έτος με το 2020 να σημειώνεται η μέγιστη τιμή. Το 2021 αναμφισβήτητα το πλήθος των πελατών παρουσιάζει μείωση με αυτό του 2020 ωστόσο, σε σχέση με τα προηγούμενα έτη (2018 και 2019), ομοίως σημειώνεται αύξηση του συνολικού πλήθους των πελατών. Ακριβώς ανάλογη εικόνα με αυτή του συνολικού πλήθους έχει και η πορεία - εξέλιξη των νέων πελατών (Πίνακας 4 και Διάγραμμα 6).

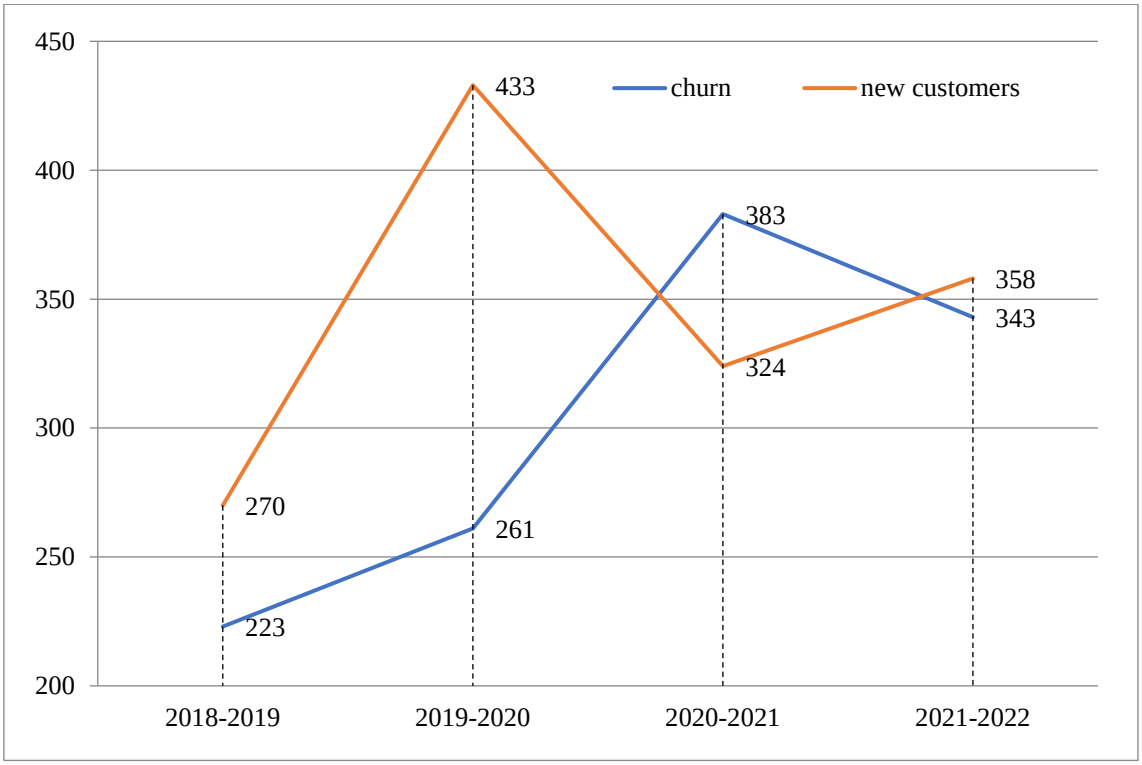
Πίνακας 4: Ανάλυση του churn και των νέων πελατών για τα έτη 2018-2022.

Έτος	**Αριθμός πελατών**	**Νέοι πελάτες**	**Ποσοστό νέων πελατών**	**Περίοδος αναφοράς**	**churn**	**κοινοί πελάτες**	**Ποσοστό απώλειας πελατών**
2018	515			**2018-2019**	223	292	43%
2019	562	270	48%				
				2019-2020	261	301	46%
2020	734	433	59%				
				2020-2021	383	351	52%
2021	675	324	48%				
				2021-2022	343	332	51%
2022	690	358	52%				
Μέσος όρος		**346**	**52%**		**303**	**319**	**48%**

Επίσης, ο μέσος όρος απώλειας πελατών του δείγματος υπολογίζεται στο 48%, δηλαδή σχεδόν ένας στους δύο πελάτες είτε αποχωρεί είτε δεν είναι σταθερός στη συνεργασία με τον όμιλο των εν λόγω επιχειρήσεων. Βέβαια, στα ίδια επίπεδα κυμαίνεται και ο μέσος όρος του ποσοστού των νέων πελατών, 52% για την ακρίβεια, άρα ο ένας από τους δύο

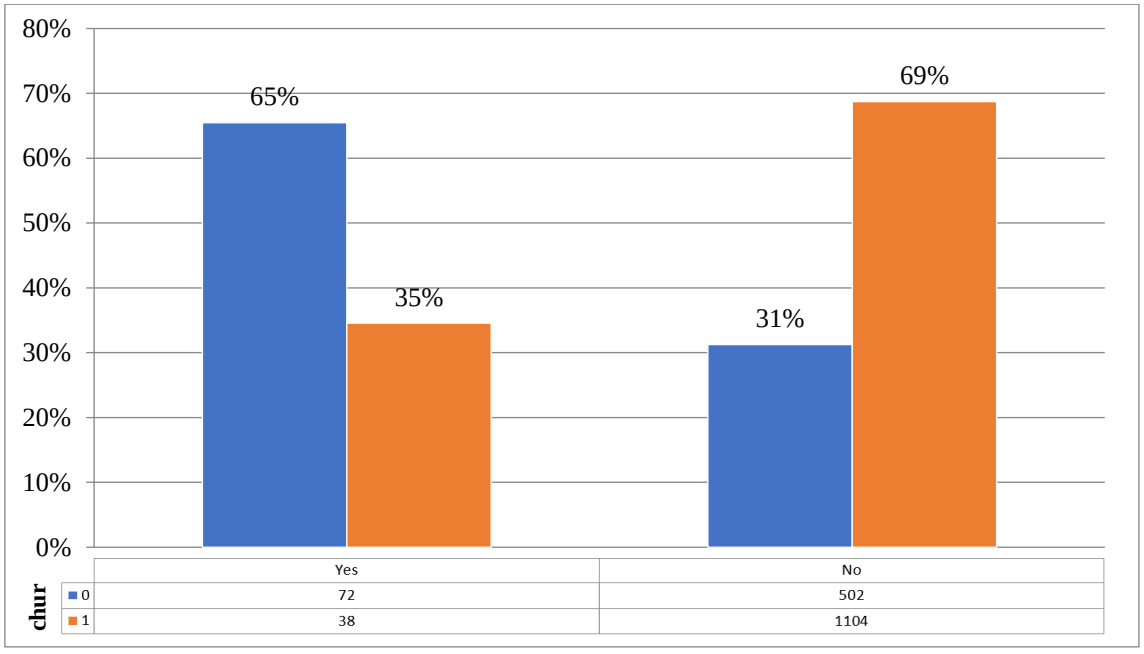
πελάτες που αποχωρεί ή διακόπτει προσωρινά τη συνεργασία αναπληρώνεται μέσα στην ίδια χρονιά με ένα νέο ή προηγούμενων ετών πελάτη (Πίνακας 4).

Στο διάγραμμα 6, αποτυπώνεται το πλήθος των νέων πελατών και αυτών που αποχώρησαν ανά περίοδο αναφοράς. Συγκεκριμένα, η μεγαλύτερη απόσταση μεταξύ των δύο μεταβλητών (νέοι πελάτες και πελάτες που αποχώρησαν) σημειώνεται στην περίοδο αναφοράς 2019-2020, καθώς το 2020, που όπως σημειώθηκε και προηγουμένως αποτελεί την καλύτερη χρονιά του δείγματος με τους περισσότερους νέους πελάτες και ταυτοχρόνως με τη μικρότερη απώλεια πελατών (Διάγραμμα 6). Συγκρίνοντας βέβαια τα τελευταία δύο έτη φαίνεται η αντιστρόφως ανάλογη σχέση του συνολικού πλήθους των πελατών με αυτούς που τελικά αποχωρούν, καθώς όπως σημειώνεται στον πίνακα 4 και στο διάγραμμα 6, ενώ ο αριθμός των πελατών αυξήθηκε από το 2021 στο 2022, το churn μειώθηκε από 57% στο 50% αντίστοιχα.



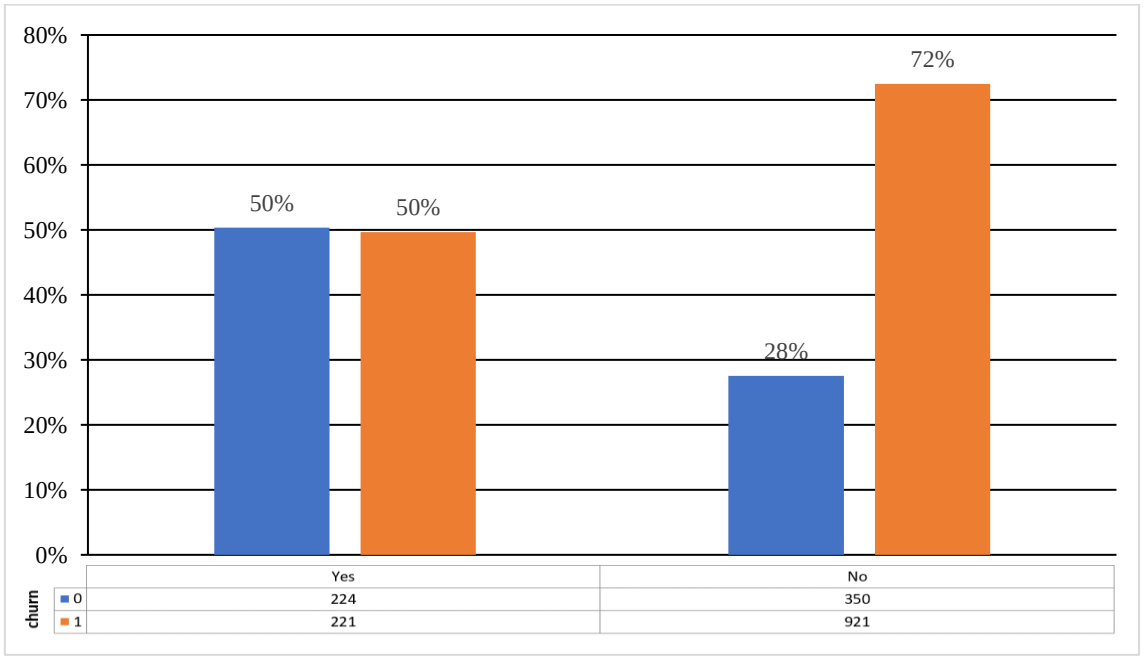
Διάγραμμα 6: Πλήθος πελατών που χάθηκαν (churn) και προστέθηκαν (new customers) ανά χρονική περίοδο.

Από τους 1.716 πελάτες του δείγματος μόλις το 6% έχει πιστωτικό όριο και μόλις ο ένας στους τρεις πελάτες παραμένει ετησίως σταθερός πελάτης των προϊόντων του ομίλου. Επίσης, παρατηρείται ότι οι πελάτες που διαθέτουν πιστωτικό όριο είναι πιο δύσκολο να αποχωρήσουν σε σύγκριση με αυτούς που δε διαθέτουν (Διάγραμμα 7).



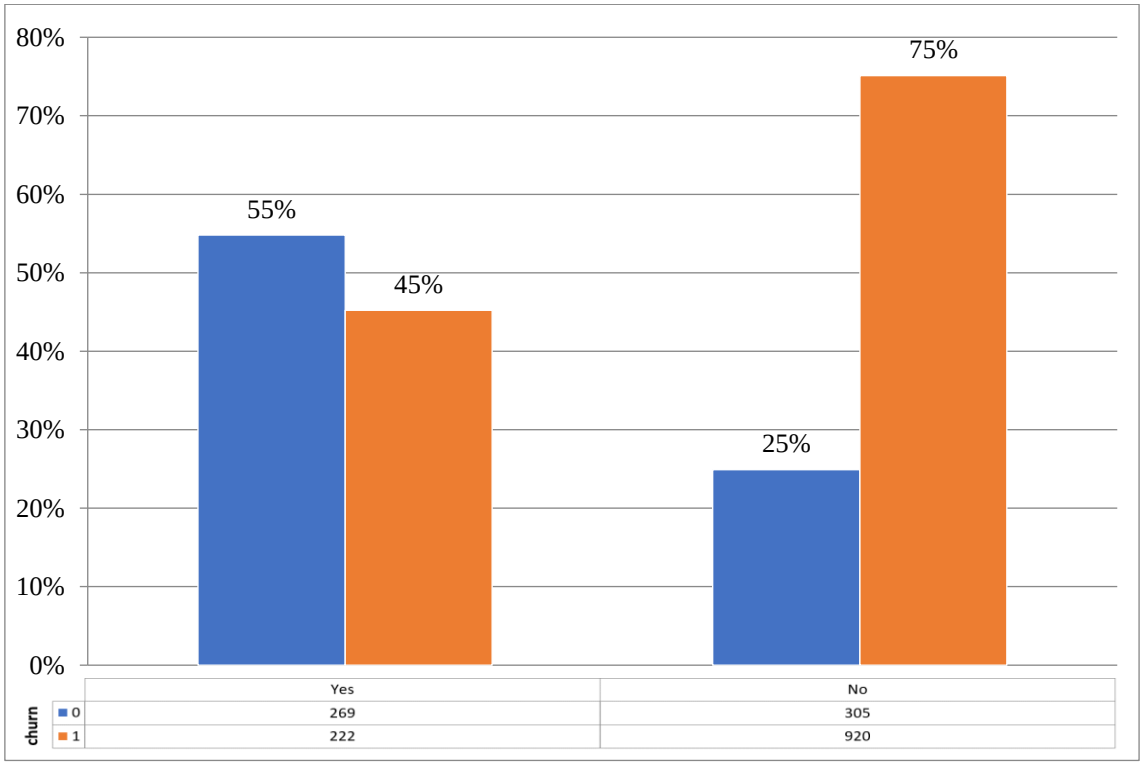
Διάγραμμα 7: Ραβδόγραμμα για τη μεταβλητή του πιστωτικού ορίου σε σχέση με την κατηγορική μεταβλητή churn.

Επίσης, συγκρίνοντας τους πελάτες που είναι χρήστες του ηλεκτρονικού καταστήματος του ομίλου με αυτούς που δεν είναι, φαίνεται πως οι δεύτεροι είναι πολύ πιο πιθανό να αποχωρήσουν. Ειδικότερα, το 54% των πελατών δεν είναι χρήστες του ηλεκτρονικού καταστήματος και είναι καταγεγραμμένοι ως πελάτες που έχουν αποχωρήσει ή δεν έχουν συνεχόμενες συναλλαγές με τον όμιλο των επιχειρήσεων (Διάγραμμα 8).



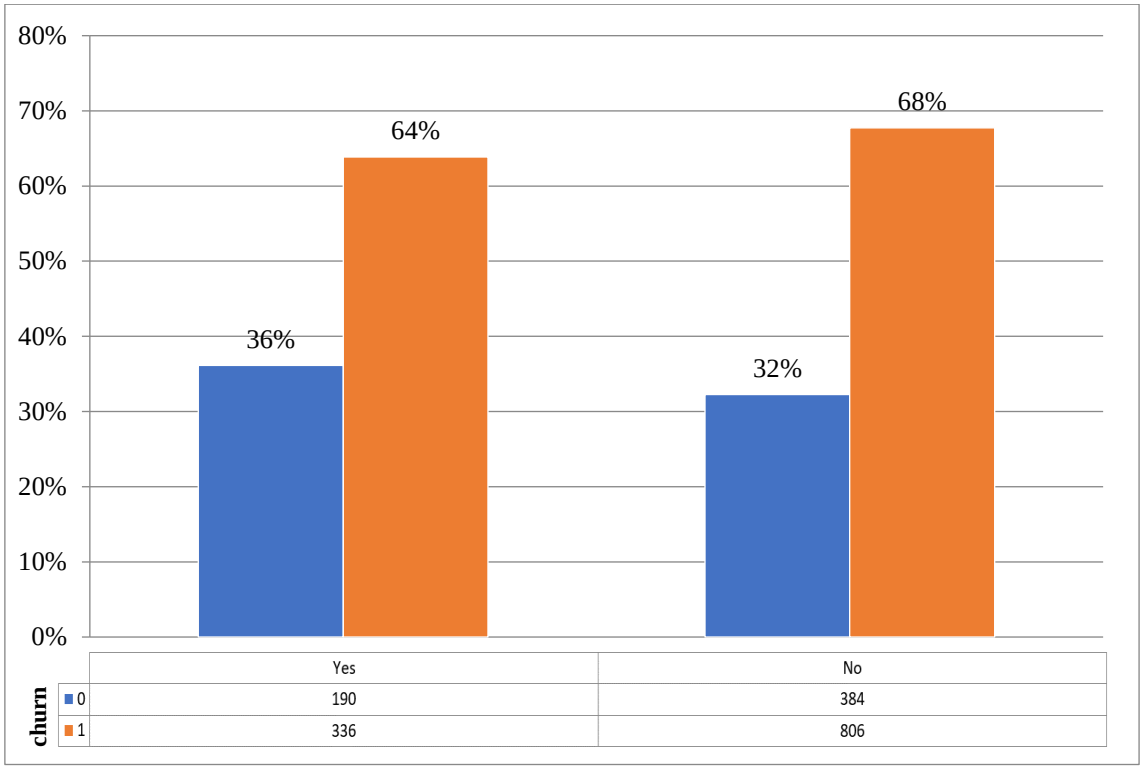
Διάγραμμα 8: Ραβδόγραμμα για τους χρήστες του ηλεκτρονικού καταστήματος σε σχέση με την κατηγορική μεταβλητή churn.

Όμοια εικόνα παρουσιάζεται και στους πελάτες που έχουν πρόσβαση στο portal με αυτούς που δεν έχουν, με τους δεύτερους να σημειώνουν ομοίως ποσοστό 54% ως μη έχοντες πρόσβαση στο portal και συνέχεια στις συναλλαγές με τον όμιλο των επιχειρήσεων (Διάγραμμα 9).



Διάγραμμα 9: Ραβδόγραμμα για τους πελάτες με πρόσβαση στο portal σε σχέση με την κατηγορική μεταβλητή churn.

Αναφορικά στις ελλείψεις των παραγγελιών ή την αποστολή ελαττωματικών εξαρτημάτων, ομοίως και εδώ στους πελάτες που έχουν αποσταλεί παραγγελίες με λάθος υλικά είναι πιο πιθανό να αποχωρήσουν παρά να παραμείνουν πιστοί στην επιχείρηση, καθώς μόλις το 11% των πελατών παραμένει πιστό στον προμηθευτή και ας έχει λάβει τουλάχιστον μία εσφαλμένη παραγγελία. Παρατηρείται επίσης ότι, το 47% των πελατών δεν είχε κάποιο σφάλμα στην παραγγελία του, ωστόσο, έχουν αποχωρήσει ή δεν έχουν συνεχόμενες συναλλαγές με τον όμιλο των επιχειρήσεων (Διάγραμμα 10). Οπότε, για το παραπάνω 47% των πελατών η αιτία αποχώρησης σίγουρα δεν οφείλεται στην αιτία της λαθεμένης παραγγελίας αλλά σε μία από τις άλλες αναλυόμενες αιτίες. Εντούτοις, το 30% του συνόλου των πελατών (526 στους 1.716 συνολικούς πελάτες) έχει λάβει τουλάχιστον μία εσφαλμένη παραγγελία.



Διάγραμμα 10: Ραβδόγραμμα με τις ελλείψεις ή ελαττωματικά εξαρτήματα σε παραγγελία σε σχέση με την κατηγορική μεταβλητή churn.

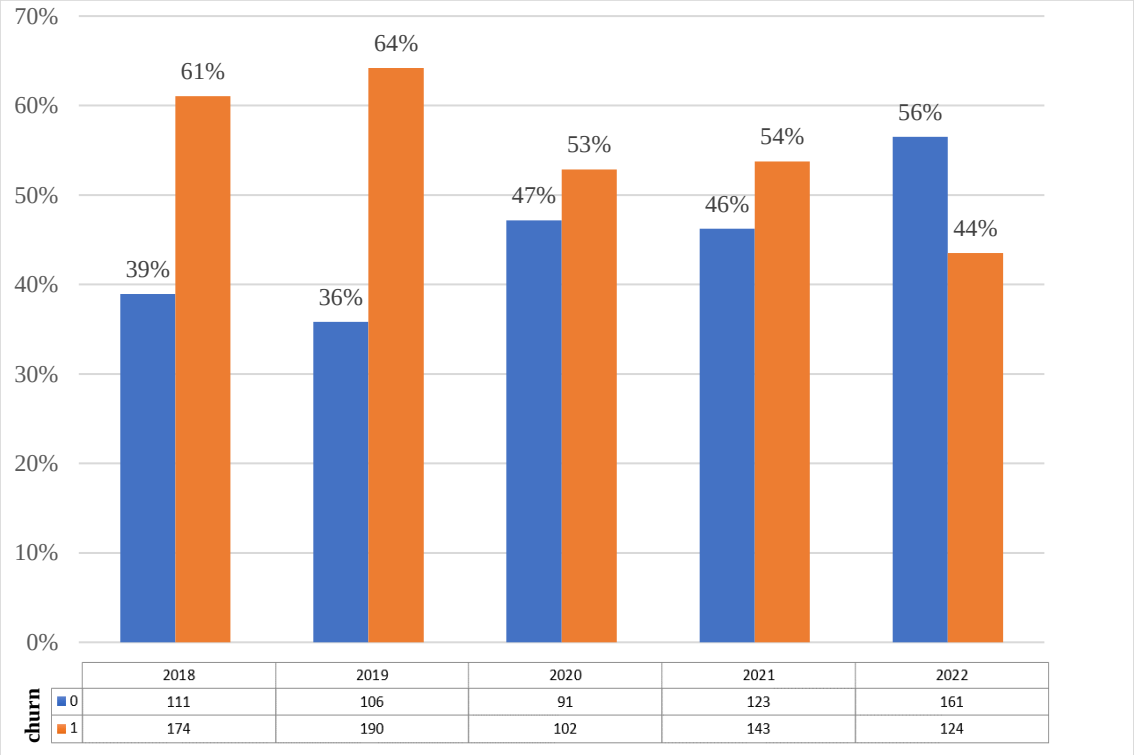
Το 30% των πελατών (526 στους 1.716 συνολικούς πελάτες) έχει λάβει τουλάχιστον μία εσφαλμένη παραγγελία στο εξεταζόμενο διάστημα 2018-2022 (Διάγραμμα 10). Ειδικότερα, στον πίνακα 5 παρατίθεται το πλήθος των πελατών με εσφαλμένες παραγγελίες ανά έτος σε σχέση με την κατηγορική μεταβλητή churn.

Πίνακας 5: Ανάλυση εσφαλμένων παραγγελιών ανά έτος σε σχέση με την κατηγορική μεταβλητή churn.

Churn	2018	2019	2020	2021	2022
0	111	106	91	123	161
1	174	190	102	143	124
Σύνολο πελατών με ελλείψεις ανά έτος	**285**	**296**	**193**	**266**	**285**

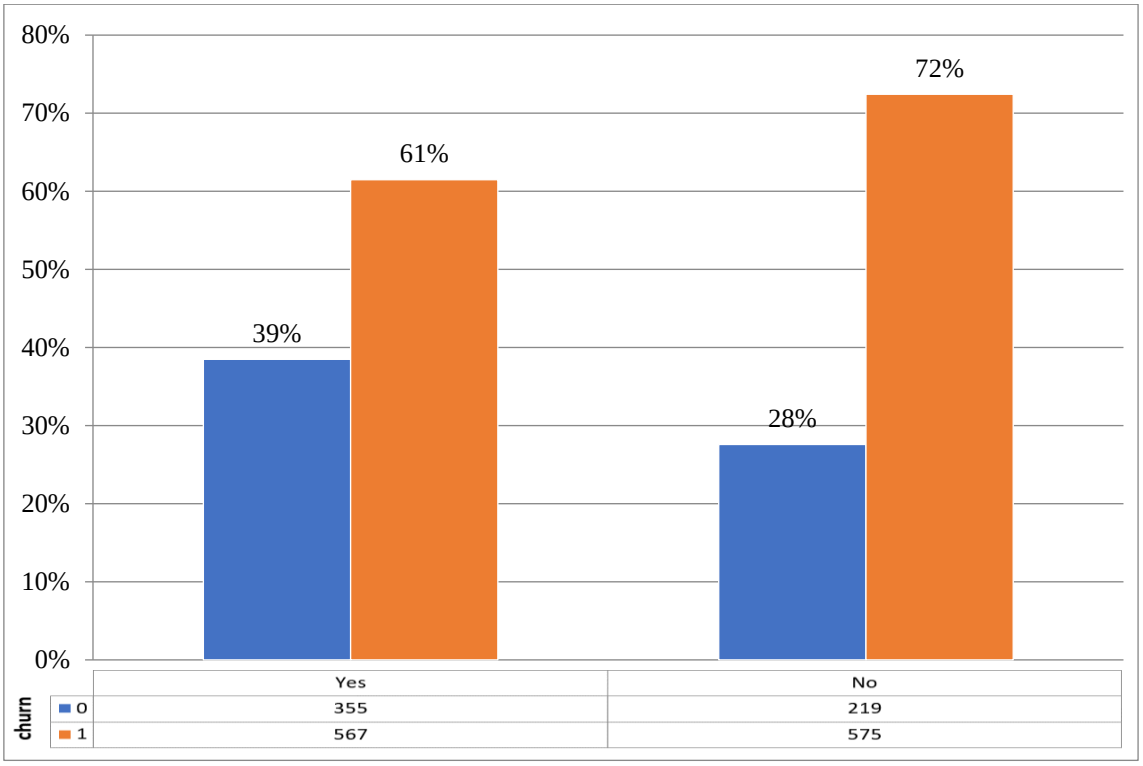
Τόσο από τον πίνακα 5 όσο και από διάγραμμα 11 παρατηρείται ότι οι πελάτες που έχουν λάβει για τον οποιοδήποτε λόγο εσφαλμένη παραγγελία στα έτη 2018 – 2021, έχουν υψηλότερες πιθανότητες να αποχωρήσουν. Ωστόσο, για το έτος 2022 παρατηρείται ότι έχουν αποχωρήσει λιγότεροι πελάτες και ας είχαν εσφαλμένη παραγγελία πράγμα που

ενδέχεται να δικαιολογείται και σε βελτιώσεις – διορθώσεις του ομίλου ώστε να εξαλείψει όσο δύναται τα δικά του λάθη (Διάγραμμα 11).



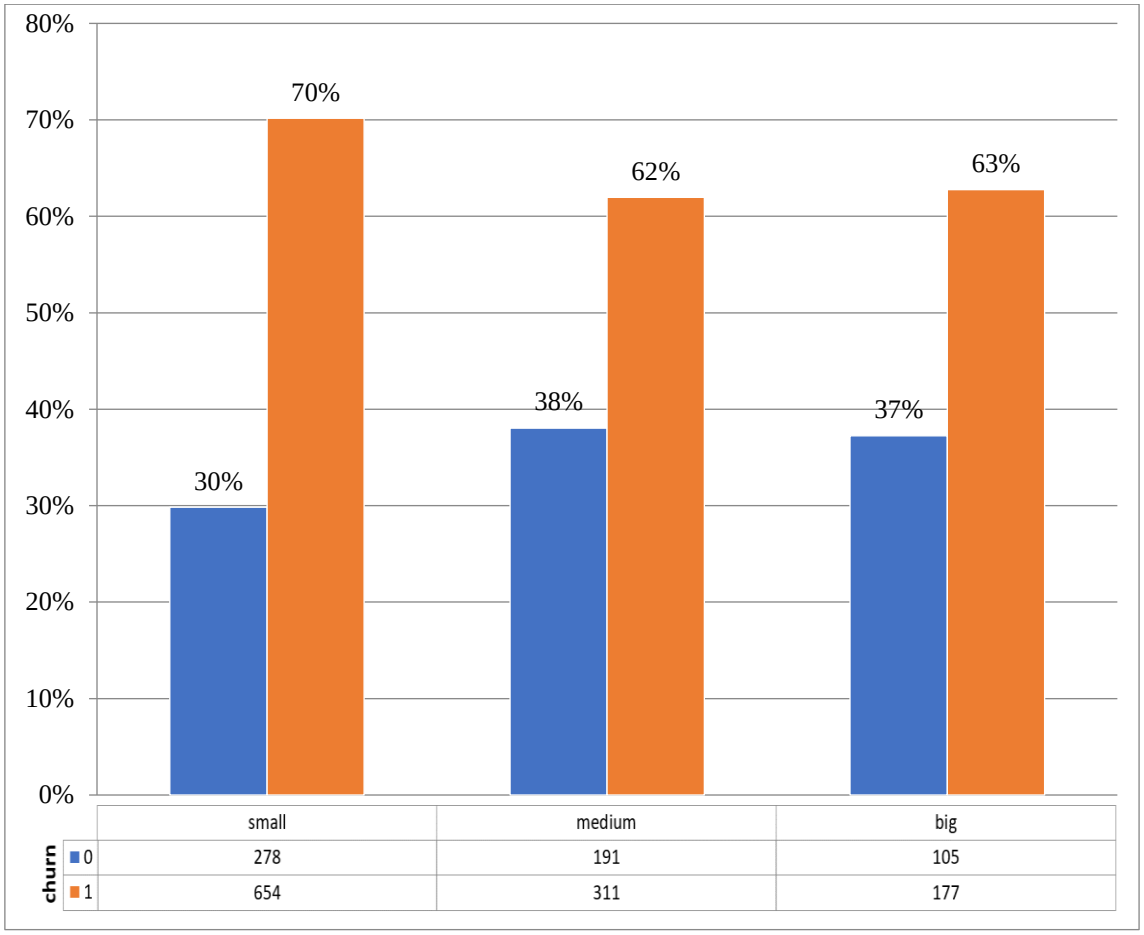
Διάγραμμα 11: Ραβδόγραμμα για τις εσφαλμένες παραγγελίες (ελλείψεις ή ελαττωματικά εξαρτήματα) σε σχέση με την κατηγορική μεταβλητή churn.

Όσον αφορά στους πελάτες που λειτουργούν και ως εγκαταστάτες του προϊόντος που αγοράζουν, φαίνεται ότι έχουν λιγότερες πιθανότητες να αποχωρήσουν σε σχέση με αυτούς που δεν είναι εγκαταστάτες, με ποσοστό 61% των πρώτων έναντι του 72% των δεύτερων (Διάγραμμα 12).



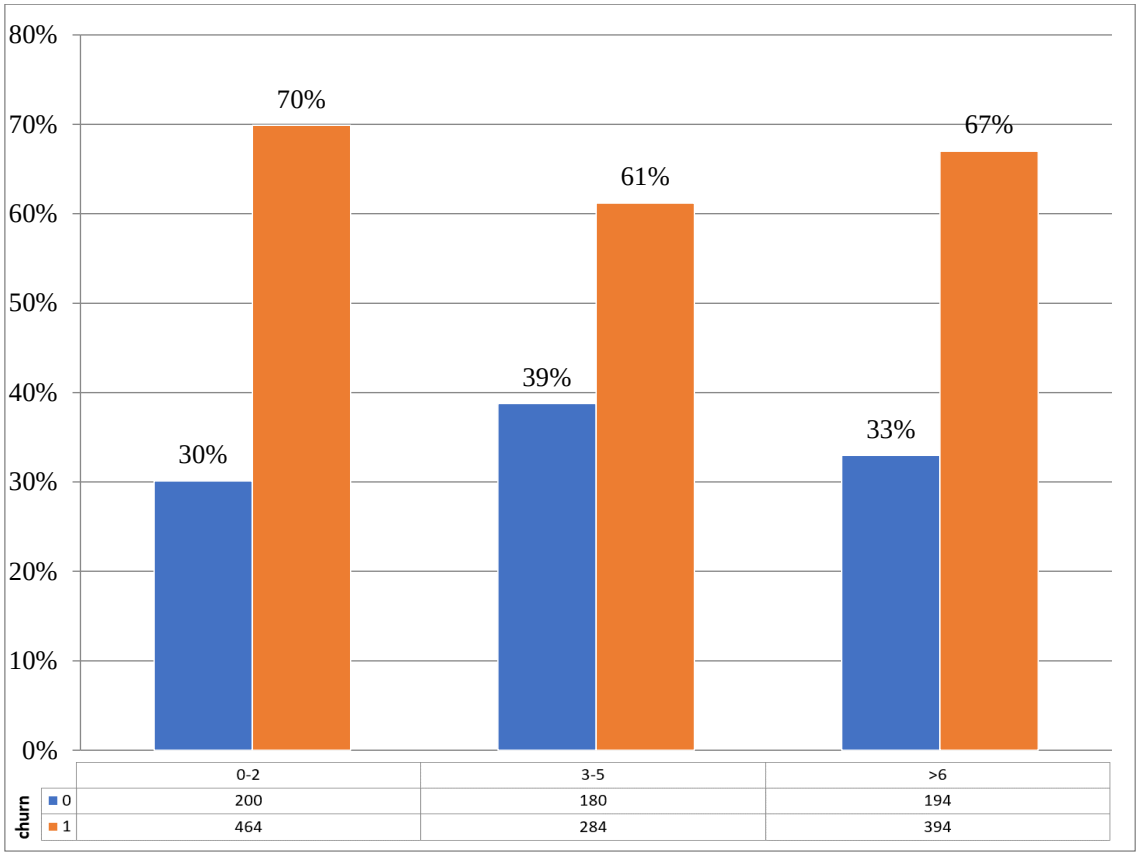
Διάγραμμα 12: Ραβδόγραμμα για τους εγκαταστάτες πελάτες σε σχέση με την κατηγορική μεταβλητή churn.

Για την καλύτερη κατανόηση των δεδομένων, έγινε τμηματοποίηση και κατάταξη των πελατών για το μέγεθος της επιχείρησης τους σε τρεις κατηγορίες small για εταιρείες με 1-9 εργαζομένους, medium για εταιρείες με 10-29 εργαζομένους και big για εταιρείες με πάνω από 30 εργαζομένους. Ακολούθως, στο διάγραμμα 13 παρατίθεται η παραπάνω μεταβλητή σε σχέση με την κατηγορική μεταβλητή churn. Είναι εμφανές ότι το μεγαλύτερο μέρος των πελατών κατατάσσεται στις μικρές επιχειρήσεις σε ποσοστό 54% οι οποίες είναι και πιο πιθανό να αποχωρήσουν σε ποσοστό 70% συγκριτικά με τις άλλες δύο κατηγορίες (medium και big) που παρουσιάζουν παρόμοια πιθανότητα αποχώρησης των πελατών σε ποσοστό 62% και 63% αντίστοιχα.



Διάγραμμα 13: Ραβδόγραμμα για το μέγεθος της επιχείρησης του πελάτη σε σχέση με την κατηγορική μεταβλητή churn.

Η επόμενη μεταβλητή που εξετάστηκε είναι τα χρόνια συνεργασίας με τον πελάτη σε σχέση με την κατηγορική μεταβλητή churn. Οι πελάτες που έχουν συνεργαστεί από 0 έως 2 χρόνια με τον όμιλο είναι πιο πιθανό να αποχωρήσουν σε σχέση με τους υπόλοιπους σε ποσοστό 70% (Διάγραμμα 14). Επιπρόσθετα, σύμφωνα με το διάγραμμα 14, οι περισσότεροι πελάτες ανήκουν στην πρώτη κατηγορία από 0-2 χρόνια συνεργασίας ενώ στην αμέσως επόμενη κατηγορία είναι οι λιγότεροι πελάτες αθροιστικά (464 πελάτες) από τους οποίους σε ποσοστό 61% είτε έχουν παύσει τη συνεργασία είτε δεν έχουν σταθερή ετήσια συνεργασία με τον όμιλο επιχειρήσεων που μελετάτε. Όσον αφορά στην τρίτη κατηγορία, σε αντίθεση με τα έτη συνεργασίας, το 67% των πελατών είτε έχουν σταματήσει τη συνεργασία είτε δεν έχουν σταθερή πελατειακή σχέση με τον όμιλο.



Διάγραμμα 14: Ραβδόγραμμα για τα χρόνια συνεργασίας με τον πελάτη σε σχέση με την κατηγορική μεταβλητή churn.

4.2. Συνολική αξία παραγγελιών ανά πωλητή

Το πλήθος των πωλητών που έχουν φέρει τις 21.099 παραγγελίες ανέρχεται στα 79 άτομα για την περίοδο αναφοράς 2018-2022. Ακολούθως, στον πίνακα 6 παρατίθενται οι top δέκα πωλητές του ομίλου επιχειρήσεων που μελετάτε.

Πίνακας 6: Top 10 πωλητές βάσει του πλήθος των καταχωρημένων παραγγελιών.

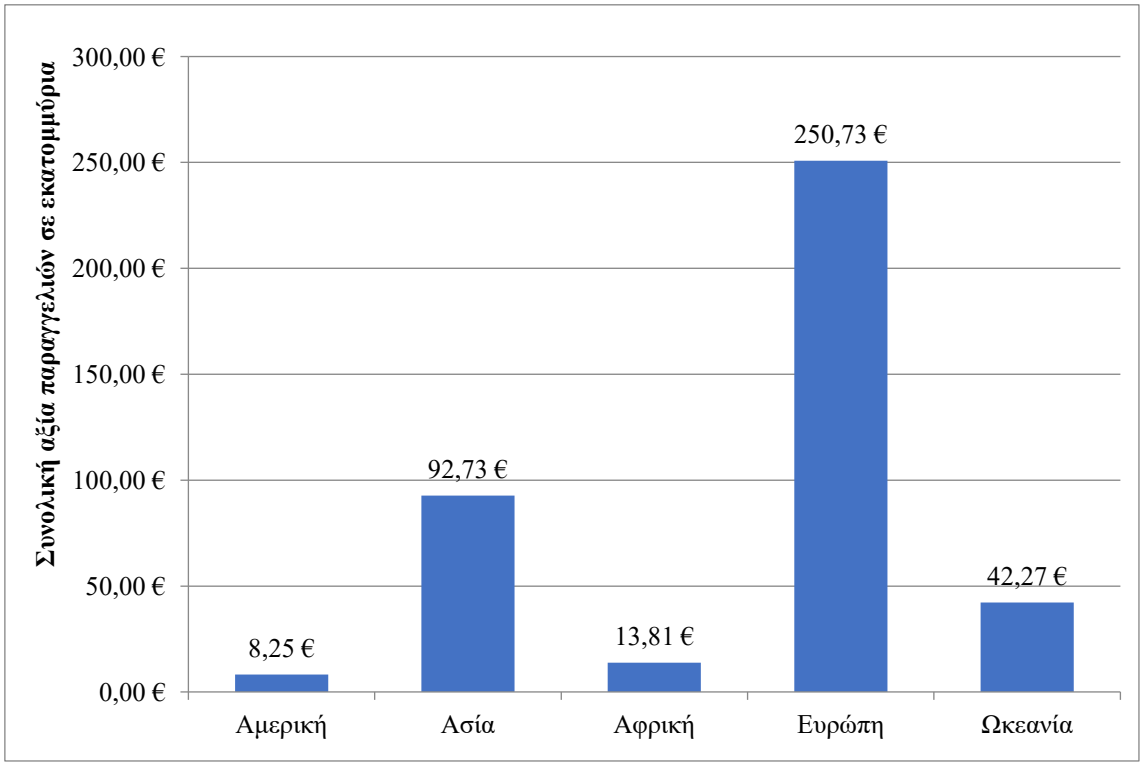
SalesPerson code	Συνολική Αξία παραγγελιών ανά πωλητή	πλήθος παραγγελιών ανά πωλητή	Μέση αξία παραγγελιών ανά πωλητή	Κατανομή πωλητών
SalesPerson 39	40.379.338,72 €	1.502	26.883,71 €	7,12%
SalesPerson 55	19.499.235,00 €	1.074	18.155,71 €	5,09%
SalesPerson 73	17.919.615,69 €	1.005	17.830,46 €	4,76%
SalesPerson 21	15.634.333,20 €	934	16.739,11 €	4,43%
SalesPerson 20	16.728.086,16 €	916	18.262,10 €	4,34%
SalesPerson 63	15.907.840,66 €	828	19.212,37 €	3,92%
SalesPerson 46	6.256.846,70 €	747	8.375,97 €	3,54%
SalesPerson 12	17.262.445,00 €	721	23.942,36 €	3,42%
SalesPerson 44	13.497.379,19 €	627	21.526,92 €	2,97%

Όπως, φαίνεται και από τα ποσοστά κατανομής παραγγελιών των πωλητών στον πίνακα 6 δεν υπάρχουν μεγάλες αποκλίσεις μεταξύ των πωλητών.

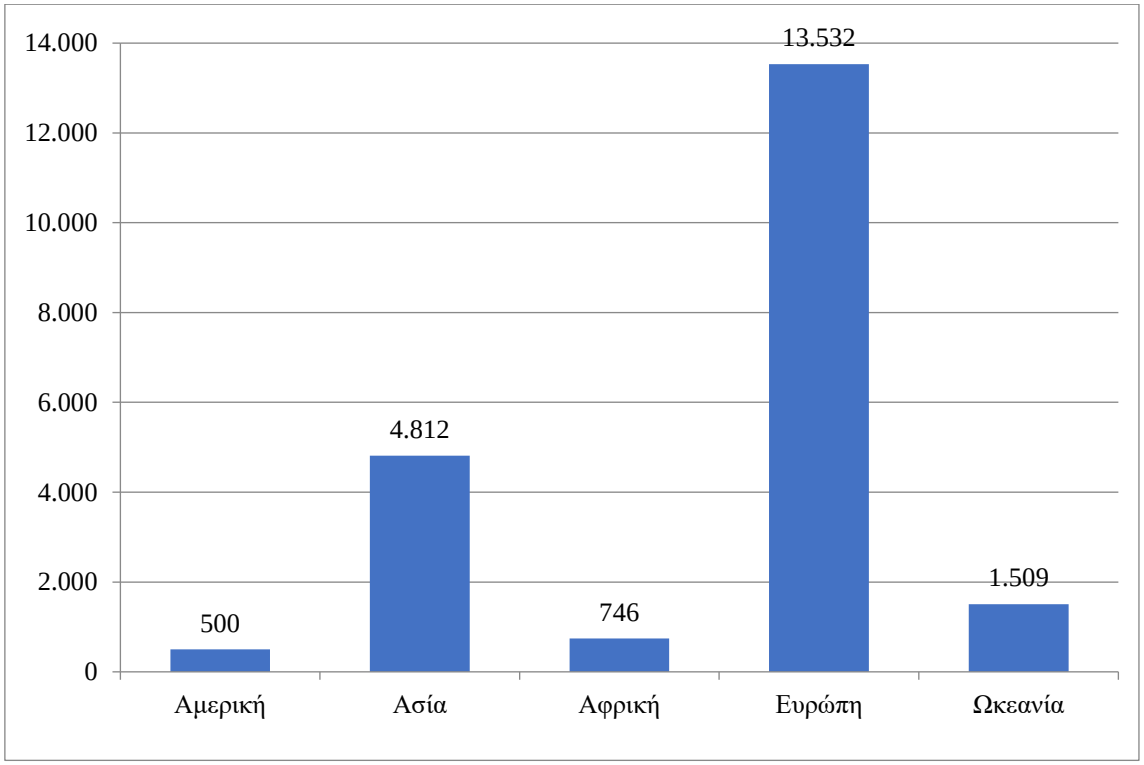
4.3. Συνολική αξία παραγγελιών ανά χώρα

Το πλήθος των χωρών που εξάγει η επιχείρηση συμπεριλαμβανομένης και της Ελλάδας είναι εκατόν δεκατρείς (113) χώρες, με την Αυστραλία να καταλαμβάνει τη νούμερο 1 αγορά με συνολική αξία παραγγελιών στην πενταετία 39.344.168,46 € και την Κολομβία να καταλαμβάνει την 113η θέση (τελευταία) με συνολική αξία παραγγελιών στην πενταετία μόλις 9.491,08 €, καθώς πρόκειται για ένα νέο πελάτη μιας και μόνο παραγγελίας το 2021, ο οποίος έκτοτε αποχώρησε και δε συνέχισε τη συνεργασία.

Ομαδοποιώντας τα παραπάνω δεδομένα ανά ήπειρο (Διάγραμμα 15 και Διάγραμμα 16) παρατηρείται ότι, το πλήθος παραγγελιών είναι πλήρως ανάλογο με αυτό της συνολικής αξίας των παραγγελιών, με την Ευρώπη να καταλαμβάνει πάνω από το 60% του συνόλου των παραγγελιών τόσο σε αξία όσο και σε πλήθος.



Διάγραμμα 15: Συνολική αξία παραγγελιών ανά ήπειρο στην πενταετία (2018-2022).



Διάγραμμα 16: Πλήθος παραγγελιών ανά ήπειρο στην πενταετία (2018-2022).

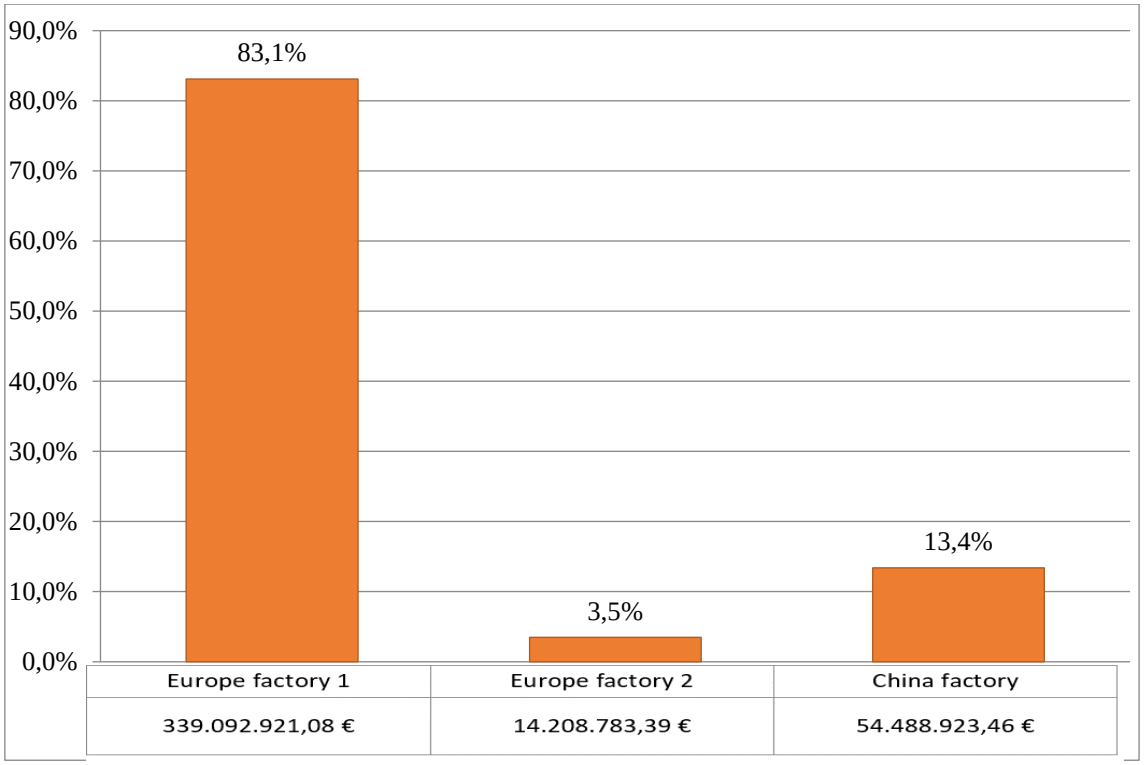
Στον πίνακα 7, παρατίθενται οι δέκα χώρες με το μεγαλύτερο συνολικό αριθμό πελατών και αντίστοιχα πόσοι πελάτες αποχώρησαν και το ποσοστό απώλειας. Εκτός από τη Γαλλία και τη Γερμανία, παρατηρείται ότι οι περισσότεροι πελάτες των υπόλοιπων χωρών σταμάτησαν ή δεν έχουν συνεχόμενη πελατειακή σχέση με την επιχείρηση.

Πίνακας 7: Οι δέκα (10) χώρες με τους περισσότερους πελάτες και το ποσοστό απώλειας (churn rate).

CountryName	Total customers	Churn customers	Churn rate
Greece	276	218	79%
Romania	248	193	78%
Cyprus	51	36	71%
China	110	71	65%
Russian Federation	51	33	65%
Australia	266	171	64%
United Kingdom	59	36	61%
Serbia	98	56	57%
France	27	12	44%
Germany	67	21	31%

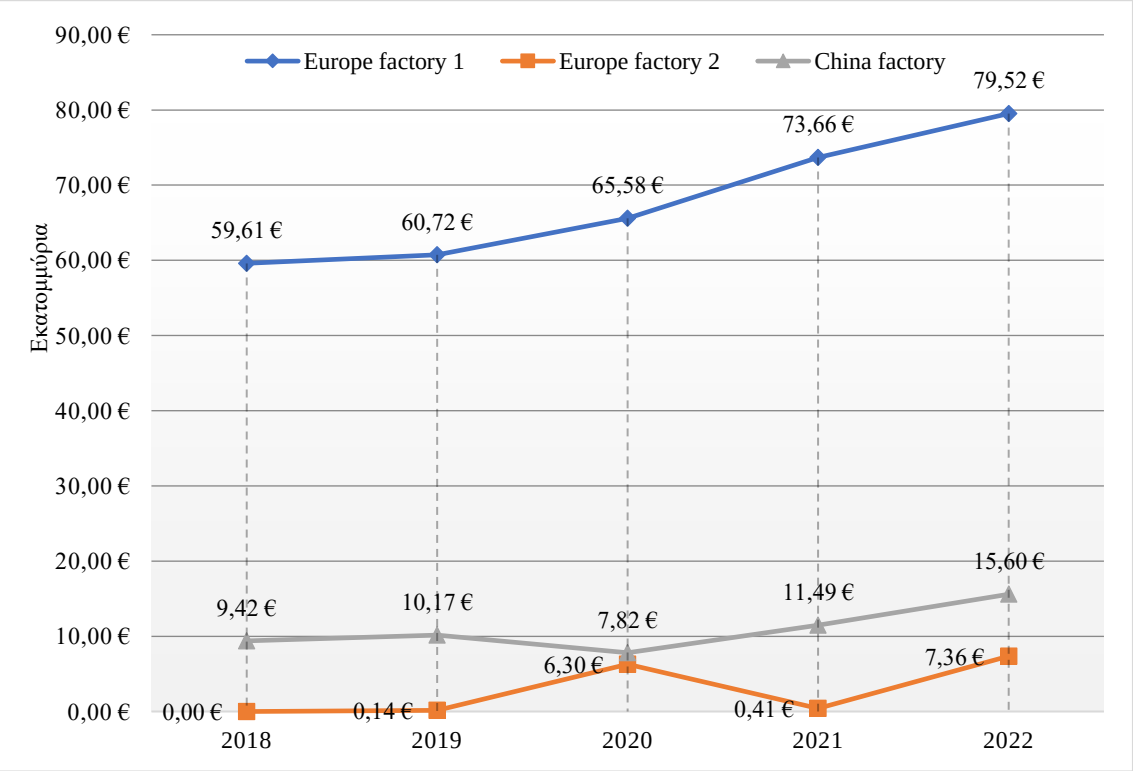
4.4. Συνολική αξία παραγγελιών ανά εργοστάσιο κατασκευής

Το εν λόγω δείγμα περιέχει παραγγελίες που έχουν παραχθεί σε τρία διαφορετικά εργοστάσια τα δύο εκ των οποίων χωροθετούνται στην Ευρώπη και το ένα στην Ασία. Το κύριο εργοστάσιο βάσει των δεδομένων είναι το Europe factory 1 (Διάγραμμα 17).



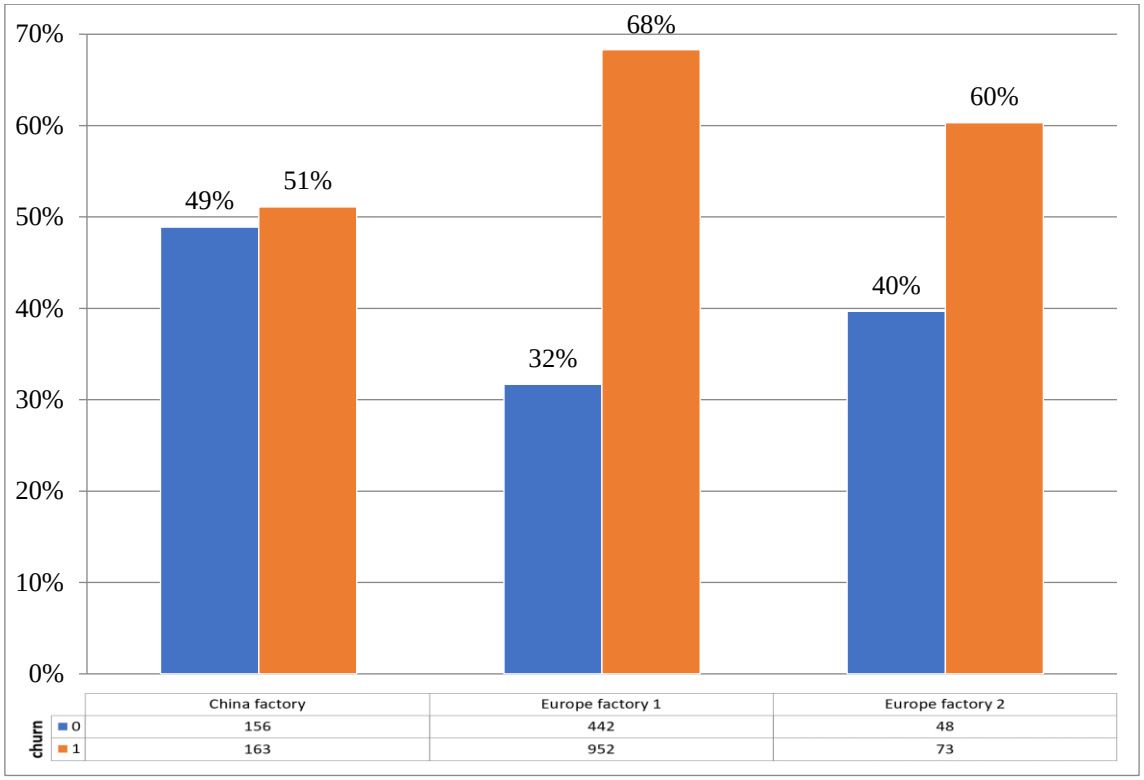
Διάγραμμα 17: Συνολική αξία παραγγελιών ανά εργοστάσιο κατασκευής της παραγγελίας στην πενταετία (2018-2022).

Ωστόσο, αναλύοντας ξεχωριστά τα έτη παρατηρείται ότι το εργοστάσιο Europe factory 2 συμμετέχει στην κατασκευή προϊόντων από το 2019 (Διάγραμμα 18) και έπειτα. Το China factory παρουσιάζει σταθερή άνοδο στη συνολική αξία παραγγελιών, εκτός από το έτος 2020 που συνδέεται με την παγκόσμια πανδημία COVID-19 και για αυτό παρουσίασε πτώση το συγκεκριμένο έτος (Διάγραμμα 18).



Διάγραμμα 18: Συνολική αξία παραγγελιών ανά εργοστάσιο στην πενταετία 2018-2022.

Τέλος, η ίδια εικόνα παρατηρείται και στα τρία εργοστάσια με τους περισσότερους πελάτες να έχουν αποχωρήσει ή να μην έχουν σταθερή πελατειακή σχέση με τον όμιλο. Το Europe factory 1 και πάλι ξεχωρίζει ως το κύριο εργοστάσιο από το συνολικό πλήθος των εξυπηρετούμενων πελατών (Διάγραμμα 19).



Διάγραμμα 19: Ραβδόγραμμα για τα εργοστάσια κατασκευής των παραγγελιών σε σχέση με την κατηγορική μεταβλητή churn.

4.5. Συνολική αξία παραγγελιών ανά τύπο προϊόντος

Το πλήθος των διαφορετικών προϊόντων που συμμετέχουν στο δείγμα για την περίοδο 2018 – 2022 είναι 41, από τα οποία το Product 1 καταλαμβάνει το 26% των πωλήσεων και το Product 2 το 10% των πωλήσεων του δείγματος (Πίνακας 8, Διάγραμμα 20 και Διάγραμμα 21). Πρόκειται για τα βασικά προϊόντα παραγωγής με το μεν πρώτο να παράγεται σχεδόν αποκλειστικά στο Europe factory 1 σε ποσοστό 99,66% και το δε δεύτερο προϊόν αποκλειστικά στο China factory (Πίνακας 9).

Πίνακας 8: Σύνολο προϊόντων για την πενταετία (2018-2022).

Product_Code	Συνολική Αξία παραγγελιών ανά προϊόν	Πλήθος παραγγελιών ανά προϊόν	Μέση αξία παραγγελιών ανά προϊόν	Κατανομή πλήθους προϊόντων	Κατανομή συνολικής αξίας παραγγελιών ανά προϊόν
Product 1	117.089.730,65 €	5.564	21.044,16 €	26,37%	28.71%
Product 2	26.682.888,30 €	2.109	12.651,91 €	10,00%	6.54%
Product 3	20.482.535,09 €	1.314	15.587,93 €	6,23%	5.02%
Product 4	19.223.533,59 €	1.239	15.515,36 €	5,87%	4.71%
Product 5	17.871.750,75 €	1.192	14.993,08 €	5,65%	4.38%
Product 6	44.845.026,33 €	1.158	38.726,27 €	5,49%	11.00%

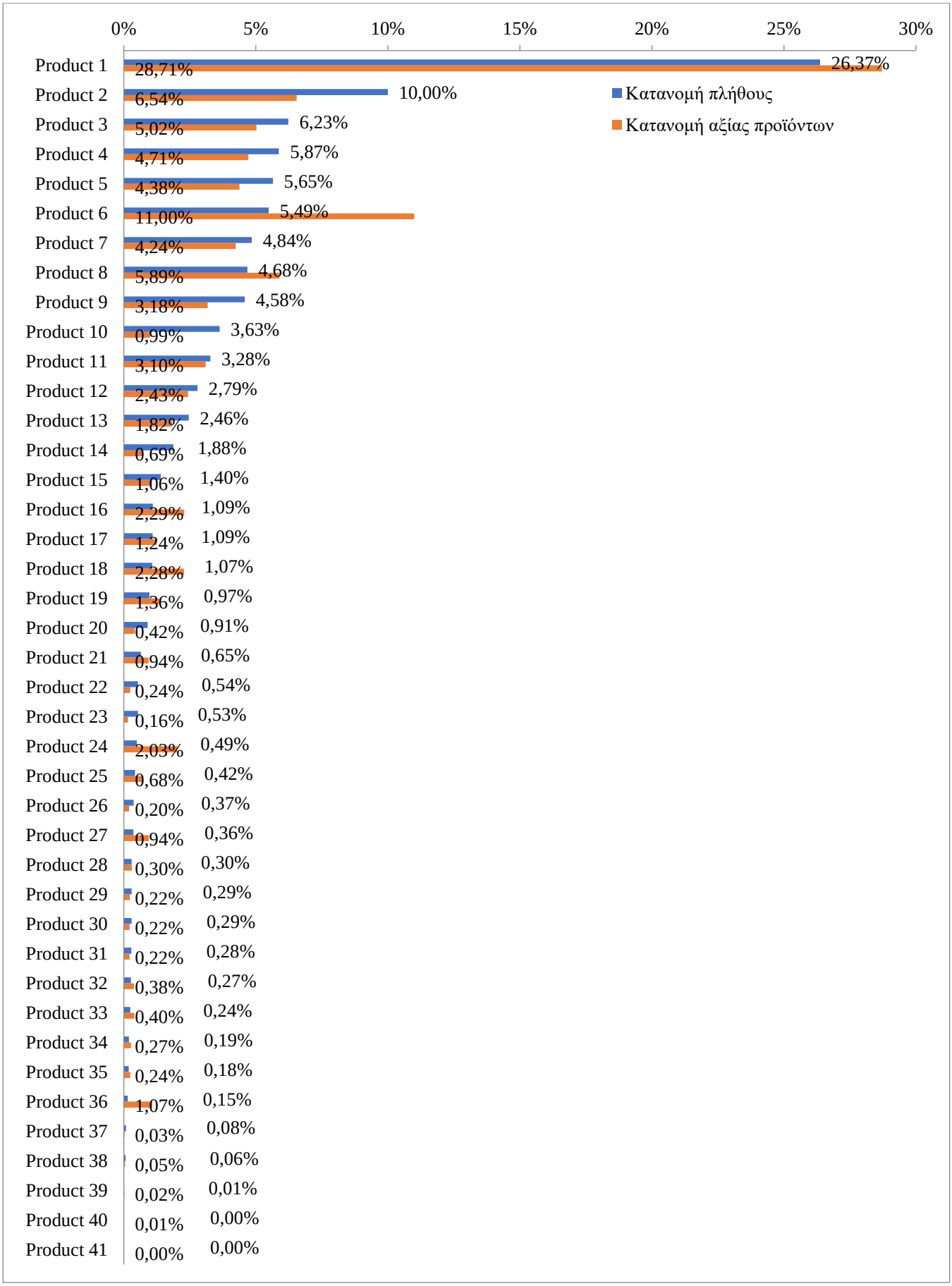
Product_Code	Συνολική Αξία παραγγελιών ανά προϊόν	Πλήθος παραγγελιών ανά προϊόν	Μέση αξία παραγγελιών ανά προϊόν	Κατανομή πλήθους προϊόντων	Κατανομή συνολικής αξίας παραγγελιών ανά προϊόν
Product 7	17.277.453,66 €	1.022	16.905,53 €	4,84%	4.24%
Product 8	24.029.922,10 €	987	24.346,43 €	4,68%	5.89%
Product 9	12.969.003,09 €	966	13.425,47 €	4,58%	3.18%
Product 10	4.036.976,70 €	765	5.277,09 €	3,63%	0.99%
Product 11	12.646.508,55 €	692	18.275,30 €	3,28%	3.10%
Product 12	9.922.556,34 €	589	16.846,45 €	2,79%	2.43%
Product 13	7.434.512,43 €	519	14.324,69 €	2,46%	1.82%
Product 14	2.829.534,16 €	396	7.145,29 €	1,88%	0.69%
Product 15	4.342.342,37 €	295	14.719,80 €	1,40%	1.06%
Product 16	9.324.456,12 €	231	40.365,61 €	1,09%	2.29%
Product 17	5.059.754,64 €	230	21.998,93 €	1,09%	1.24%
Product 18	9.288.951,62 €	226	41.101,56 €	1,07%	2.28%
Product 19	5.560.812,72 €	204	27.258,89 €	0,97%	1.36%
Product 20	1.694.802,41 €	191	8.873,31 €	0,91%	0.42%
Product 21	3.829.338,31 €	137	27.951,37 €	0,65%	0.94%
Product 22	981.640,06 €	113	8.687,08 €	0,54%	0.24%
Product 23	648.400,37 €	112	5.789,29 €	0,53%	0.16%
Product 24	8.292.276,23 €	103	80.507,54 €	0,49%	2.03%
Product 25	2.773.908,57 €	89	31.167,51 €	0,42%	0.68%
Product 26	816.849,71 €	79	10.339,87 €	0,37%	0.20%
Product 27	3.844.625,94 €	76	50.587,18 €	0,36%	0.94%
Product 28	1.243.688,99 €	63	19.741,10 €	0,30%	0.30%
Product 29	909.359,48 €	62	14.667,09 €	0,29%	0.22%
Product 30	896.250,00 €	62	14.455,65 €	0,29%	0.22%
Product 31	895.539,23 €	60	14.925,65 €	0,28%	0.22%
Product 32	1.550.843,32 €	56	27.693,63 €	0,27%	0.38%
Product 33	1.629.770,56 €	51	31.956,29 €	0,24%	0.40%
Product 34	1.098.460,63 €	41	26.791,72 €	0,19%	0.27%
Product 35	990.284,76 €	39	25.391,92 €	0,18%	0.24%
Product 36	4.367.808,85 €	32	136.494,03 €	0,15%	1.07%
Product 37	108.548,62 €	17	6.385,21 €	0,08%	0.03%
Product 38	184.209,83 €	13	14.169,99 €	0,06%	0.05%
Product 39	76.717,00 €	3	25.572,33 €	0,01%	0.02%
Product 40	28.850,00 €	1	28.850,00 €	0,00%	0.01%
Product 41	10.205,86 €	1	10.205,86 €	0,00%	0.00%
	Σ=407.790.627,93 €	**Μ.Ο.=515**	**Μ.Ο.=24.432,03 €**	**Σ=100,00%**	**Σ=100,00%**

Πίνακας 9: Παρουσίαση κύριων προϊόντων παραγωγής βάσει εργοστασίου παραγωγής τους.

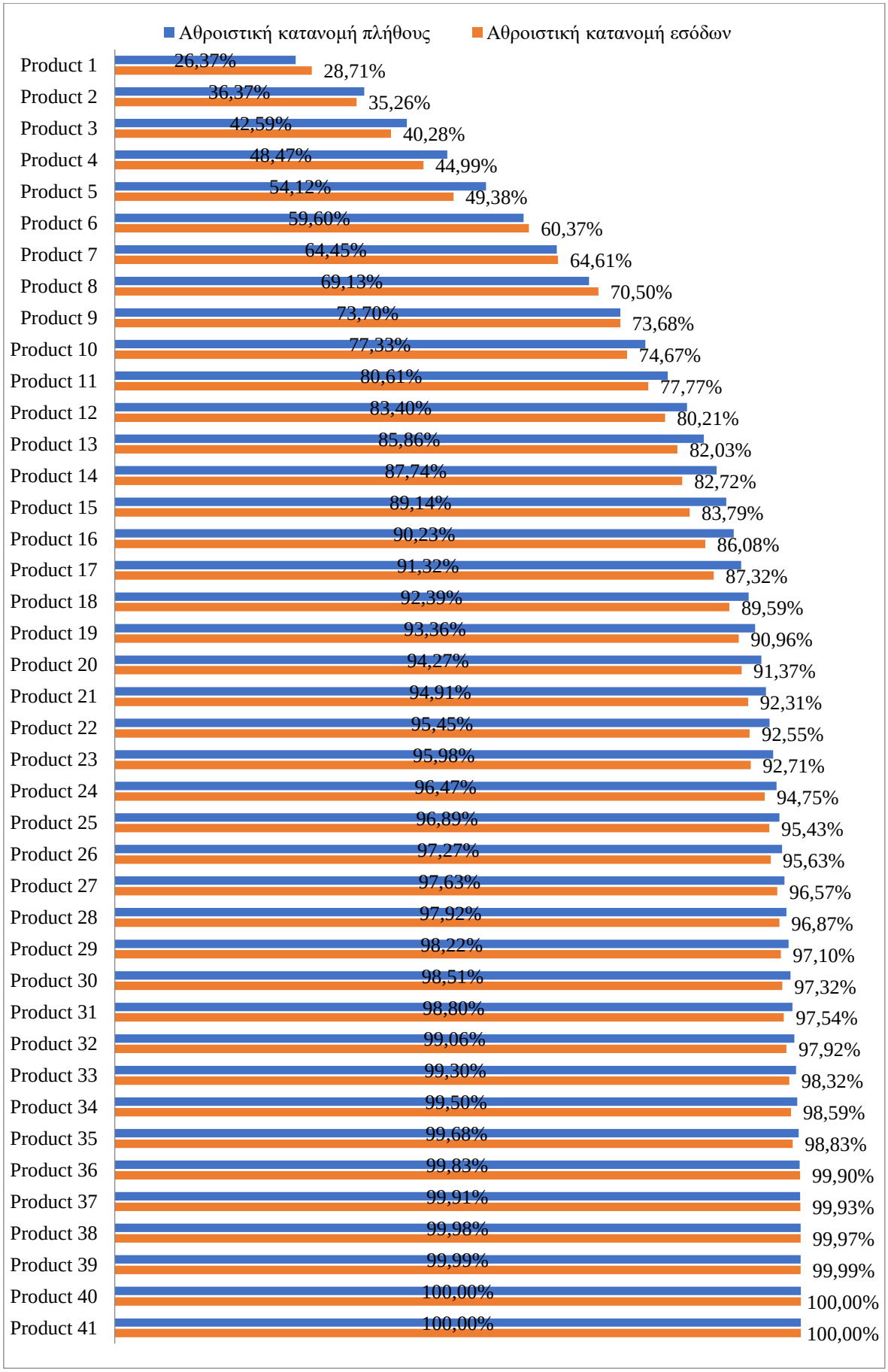
Product_Code	Manufacturer_Code	Πλήθος παραγόμενων προϊόντων	Κατανομή προϊόντος ανά εργοστάσιο
Product 1	China factory	18	0,32%
Product 1	Europe factory 1	5.545	99,66%
Product 1	Europe factory 2	1	0,02%
Product 2	China factory	2.109	100,00%

Επίσης, ένα συμπέρασμα που προκύπτει από τον πίνακα 8 και το διάγραμμα 21 είναι ότι στο 90% των πωληθέντων προϊόντων συμμετέχουν μόλις τα πρώτα 15 προϊόντα από τα 41 συνολικά διαθέσιμα προϊόντα του ομίλου. Ωστόσο, το ποσοστό διαφοροποιείται όταν αθροίζονται τα αντίστοιχα ποσοστά εσόδων που έχουν αποφέρει τα ίδια πρώτα 15 προϊόντα και συγκεκριμένα υπολογίζεται κοντά στο 84% (Πίνακας 8 και Διάγραμμα 21).

Εξετάζοντας περισσότερο τις δύο κατανομές πλήθους και εσόδων για τα 41 διαφορετικά προϊόντα (Διάγραμμα 20 και Διάγραμμα 21) παρατηρούμε ότι το Product 1 να αποφέρει το μεγαλύτερο ποσοστό εσόδων στο σύνολο των προϊόντων σε ποσοστό σχεδόν 29% και ακολουθεί το Product 6 στη δεύτερη θέση όσον αφορά στα έσοδα. Το Product 2 μπορεί σε πλήθος να υπερτερεί κατά πολύ του Product 6 αλλά το τελευταίο με τα μισά πωληθέντα προϊόντα έναντι του πρώτου έχει αποφέρει τα διπλάσια έσοδα. Τα Product 18 και Product 24 μπορεί σε πλήθος προϊόντων να μη φτάνουν ούτε τα μισά του Product 10 που είναι δέκατο στη σειρά (Πίνακας 8) αλλά τα έσοδά τους είναι διπλάσια σε σχέση με τα έσοδα του Product 10. Το Product 27 ενώ το ποσοστό συμμετοχής του πλήθος των παραγγελιών του στην πενταετία 2018 2022 είναι μόλις 0,36% με 76 συνολικά παραγγελίες, το ύψος των εσόδων του αυτό του Product 10 με ποσοστό συμμετοχής στα έσοδα κοντά στο 1%. Παρόμοια περίπτωση είναι και το Product 36 το οποίο σημειώνει μόλις 32 παραγγελίες στο διάστημα αναφοράς 2018 – 2022, αλλά το ύψος των εσόδων αυτών των παραγγελιών ξεπερνά το ύψος των εσόδων 765 παραγγελιών του Product 10.



Διάγραμμα 20: Σχηματική απεικόνιση κατανομής πλήθους και αξίας των 41 διαφορετικών τύπων προϊόντων.



Διάγραμμα 21: Σχηματική απεικόνιση αθροιστικής κατανομής πλήθους και εσόδων για τους 41 διαφορετικούς τύπους προϊόντων.

5. Τμηματοποίηση Πελατών

5.1. Τμηματοποίηση πελατών με την ανάλυση RFM

Για την καλύτερη ανάλυση της συμπεριφοράς των πελατών χρησιμοποιήθηκε η ανάλυση RFM ώστε να κατηγοριοποιηθούν οι πελάτες ανάλογα με τα τρία αυτά χαρακτηριστικά. Συγκεκριμένα, η μεταβλητή Recency αφορά στην πιο πρόσφατη ημερομηνία κατά την οποία ο πελάτης προχώρησε σε αγορά, η μεταβλητή Frequency αφορά στη συχνότητα των παραγγελιών του πελάτη (πλήθος παραγγελιών) και τέλος η συνολική αξία των παραγγελιών του πελάτη αποδίδεται με τη μεταβλητή Monetary. Με τον τρόπο αυτό δημιουργείται για κάθε πελάτη του δείγματος τρεις νέες στήλες που περιγράφουν τα παραπάνω. Για παράδειγμα στον πίνακα 10 που είναι απόσπασμα από το συνολικό πίνακα των δεδομένων, ο πελάτης 100 πραγματοποίησε την τελευταία του συναλλαγή στις 14/4/2022, έχει συνολικά 16 παραγγελίες την περίοδο αναφοράς και η συνολική αξία παραγγελιών είναι 222.905€.

Πίνακας 10: Απόσπασμα πίνακα δεδομένων που απεικονίζει για κάθε πελάτη την ημερομηνία τελευταίας συναλλαγής (Recency) το συνολικό αριθμό παραγγελιών (Frequency) και τη συνολική αξία (Monetary).

CustomerName code	Recency	Frequency	Monetary
Customer 100	14/4/2022	16	222.905,00 €
Customer 1000	10/6/2022	2	30.136,11 €
Customer 1001	16/12/2020	1	15.400,00 €

Εν συνεχεία, ορίζονται τρεις νέες μεταβλητές, οι R-score, F-score και M-score που παίρνουν τις τιμές 1-5 με το 1 να υποδηλώνει τη μικρότερη βαθμολογία και το 5 την υψηλότερη βαθμολογία. Οι τιμές 1-5 αποτελούν ουσιαστικά τις κλάσεις κάθε μεταβλητής βάσει της ιεράρχησης που αναφέρεται στη συνέχεια. Οι κλάσεις που δημιουργήθηκαν για τις παραπάνω μεταβλητές παρατίθενται στον πίνακα 11.

Για τη μεταβλητή Recency ακολουθήθηκε η θεωρία που ορίζει ότι τα δεδομένα χωρίζονται σε πεμπτημόρια, όμως για τις μεταβλητές Frequency και Monetary, οι κλάσεις δημιουργήθηκαν εμπειρικά και βάσει του κλάδου στον οποίο ανήκει η επιχείρηση.

Πιο συγκεκριμένα, για τη μεταβλητή Frequency αν οι πέντε κλάσεις αντιστοιχούσαν στο 20% των δεδομένων, οι δύο πρώτες κλάσεις θα περιλάμβαναν 1.388 παρατηρήσεις από τις συνολικά 1.716 που δε θα βοηθούσε στην ανάλυση. Αυτό συμβαίνει γιατί υπάρχουν 722 πελάτες που έχουν ακριβώς μία παραγγελία και 666 πελάτες που έχουν 2-9 παραγγελίες. Το να υπάρχει πελάτης με μία παραγγελία είναι σύνηθες στον κλάδο καθώς ο πελάτης τιμολογίου μπορεί κάθε φορά να μην είναι η τεχνική εταιρεία (κατασκευαστής) ή ο επενδυτής του έργου αλλά η εταιρεία διαχείρισης του έργου που είναι μοναδική οντότητα. Με αντίστοιχη λογική δημιουργήθηκαν και οι κλάσεις της μεταβλητής Monetary για να αποφευχθεί η συγκέντρωση των μοναδικών παραγγελιών σε μια κλάση που θα μπορούσε να έχει πλήθος άνω των 900 εγγραφών που αντιστοιχεί σε ποσοστό άνω του 50% των παρατηρήσεων.

Πίνακας 11: Οι κλάσεις της ανάλυσης RFM.

R rank		F rank		M rank	
1/1/2018	1	0	1	2.000 €	1
1/1/2019	2	2	2	20.000 €	2
1/1/2020	3	10	3	50.000 €	3
1/1/2021	4	50	4	500.000 €	4
1/1/2022	5	100	5	1.000.000 €	5

Η κατανομή των πελατών με βάση τις μεταβλητές R-score, F-score και M-score φαίνονται στον πίνακα 12.

Πίνακας 12: Κατανομή πελατών R-score, F-score και M-score.

R-score Distribution		F-score Distribution		M-score Distribution	
1	158	1	722	1	487
2	206	2	666	2	472
3	319	3	228	3	582
4	343	4	56	4	88
5	690	5	44	5	87
Total	1.716	**Total**	1.716	**Total**	1.716

Αθροίζοντας τα επιμέρους σκορ του κάθε πελάτη, προκύπτει το συνολικό σκορ που αποτελεί μια νέα μεταβλητή το RFM score. Εφόσον το σκορ των μεταβλητών R-score, F-score και M-score κυμαίνονται από 1 ως 5, σημαίνει ότι το RFM score κάθε πελάτη μπορεί να πάρει τις τιμές 3-15. Με βάση το RFM score μπορούν να βγουν κάποια πρώτα

συμπεράσματα για το πελατολόγιο και να τους ταξινομήσουμε σε κατηγορίες. Πελάτες με χαμηλό σκορ σημαίνει ότι δεν είναι αρκετά σημαντικοί για την επιχείρηση καθώς δεν έχουν αγοράσει πρόσφατα, δεν έχουν μεγάλο αριθμό παραγγελιών και ο συνολικός τζίρος τους είναι χαμηλός ενώ αντίθετα πελάτες με μεγάλο RFM score είναι πολύ σημαντικοί. Για να γίνει μια ομαδοποίηση πελατών ανάλογα με τη σημαντικότητά τους, δημιουργήθηκαν 5 κλάσεις ανάλογα με το RFM score. Όπως αναφέρθηκε και στη θεωρία οι κατηγορίες πελατών που προκύπτουν από την ανάλυση RFM με αύξουσα σημαντικότητα είναι 1-Lost, 2-At risk, 3-Can't loose them, 4-Promising, 5-Potential. Με βάση το άθροισμα του RFM score μπορεί να γίνει η αντιστοίχιση που φαίνεται στον πίνακα 13.

Πίνακας 13: Τμηματοποίηση πελατών με βάση το RFM score και της αντίστοιχης σημαντικότητάς τους.

RFM score	Category	Importance
3-4	1	Lost
5-7	2	At risk
8-10	3	Can't loose them
11-13	4	Promising
14-15	5	Potential

Η κατηγορία 1 περιλαμβάνει τους χαμένους πελάτες με RFM score 3 και 4, δηλαδή πελάτες που δεν έχουν κάνει κάποια συναλλαγή πρόσφατα, έχουν πολύ μικρό αριθμό παραγγελιών και μικρό συνολικό τζίρο πενταετίας. Για τους συγκεκριμένους πελάτες δεν είναι ανάγκη να γίνει κάποια στρατηγική διατήρησης (retention) καθώς δε θα βοηθήσουν καθόλου την κερδοφορία της επιχείρησης. Η κατηγορία 2 περιλαμβάνει πελάτες με RFM score 5, 6 και 7 και είναι πελάτες με παρόμοιο προφίλ με την κατηγορία 1 και ομοίως δεν κρίνεται απαραίτητη η διατήρηση των πελατών. Η κατηγορία 3 είναι μια κρίσιμη μάζα πελατών με RFM score 8, 9 και 10 που τους τοποθετεί στο μέσο της κλίμακας. Οι πελάτες αυτοί στο προσεχές μέλλον ή θα γίνουν πελάτες At risk ή θα ανέβουν και θα πάνε στην κατηγορία 4, τους υποσχόμενους πελάτες. Για τους συγκεκριμένους πελάτες η επιχείρηση πρέπει να αναγνωρίσει τη δυναμική τους και αντίστοιχα να προβεί σε στρατηγικής διατήρησης ή όχι. Στην κατηγορία 4 με RFM score 11, 12 και 13 ανήκουν οι πελάτες που έχουν κάποια δυναμική και θα μπορούσε η επιχείρηση να εστιάσει με πρόσθετα κίνητρα που θα οδηγούσε σε περισσότερες παραγγελίες ή/και υψηλότερο τζίρο. Τέλος, η κατηγορία 5 περιλαμβάνει τους καλύτερους πελάτες ή όπως συναντάται στη βιβλιογραφία τους πελάτες διαμάντια (diamond), αυτούς με τις καλύτερες

προοπτικές και τους πιο κερδοφόρους. Είναι οι πελάτες που είχαν συναλλαγή με την επιχείρηση πρόσφατα, έχουν μεγάλο αριθμό παραγγελιών και μεγάλο τζίρο.

Η κατανομή των πελατών ως προς το συνολικό RFM score φαίνεται στον πίνακα 14. Όπως φαίνεται υπάρχουν 201 πελάτες στην κατηγορία 1-Lost, στις κατηγορίες 2 και 3 περιλαμβάνονται οι περισσότεροι πελάτες, η κατηγορία 4 περιλαμβάνει 197 πελάτες και στην κατηγορία 5 με τους καλύτερους πελάτες ανήκουν 74 πελάτες.

Πίνακας 14: Κατανομή πελατών με βάση το RFM score.

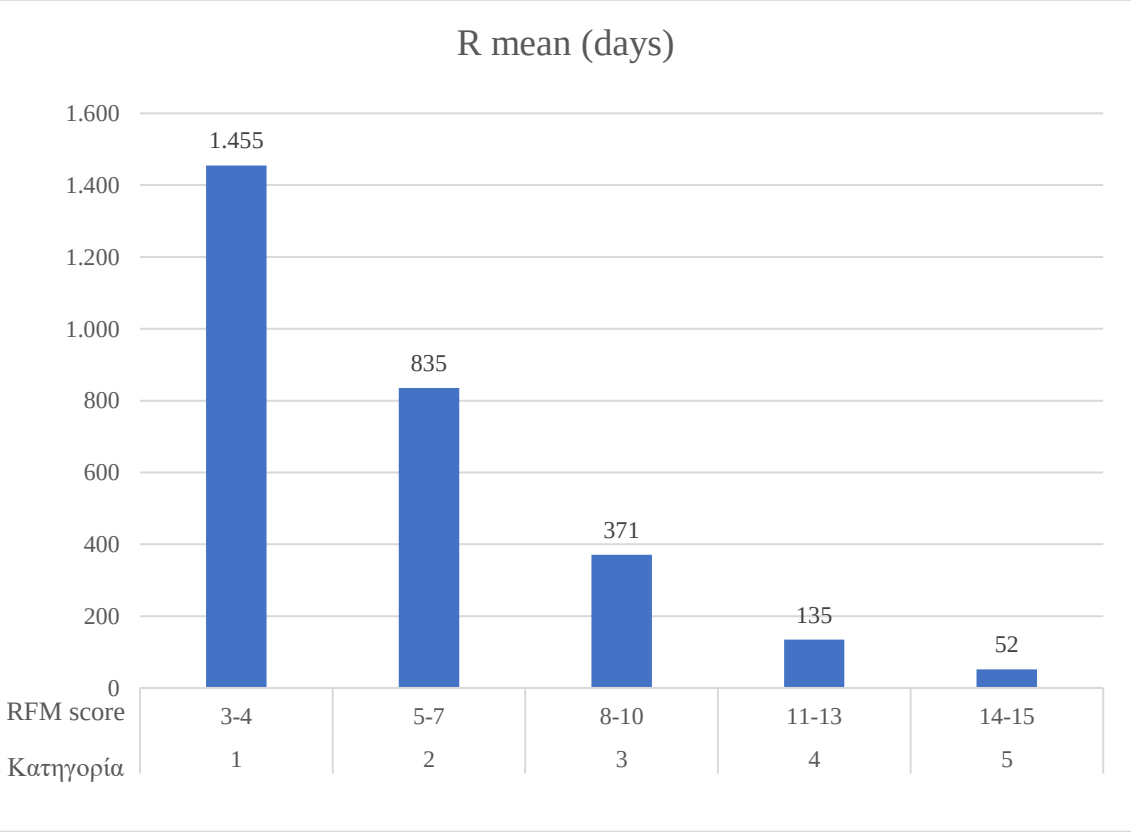
Category	RFM score	Distribution
1	3-4	201
2	5-7	621
3	8-10	623
4	11-13	197
5	14-15	74
	Total	**1.716**

Μεγάλο ενδιαφέρον παρουσιάζει η ανάλυση των μέσων τιμών για τις μεταβλητές Recency, Frequency και Monetary από την παραπάνω κατηγοριοποίηση για να γίνει κατανοητό το ενδεικτικό προφίλ πελάτη ανάλογα με την κατηγορία στην οποία ανήκει. Ο πίνακας 15 παρουσιάζει τις μέσες τιμές για τις 5 κατηγορίες. Οι πελάτες της κατηγορίας 1 είχαν κατά μέσο όρο τελευταία συναλλαγή με την επιχείρηση 1.455 μέρες πριν από την περίοδο αναφοράς (31/12/2022), πραγματοποίησαν 1,03 παραγγελία μέσης αξίας 13.276,42€. Στην κατηγορία 2 οι πελάτες είχαν κατά μέσο όρο συναλλαγή 835 μέρες πριν, 1,72 παραγγελίες και συνολικό τζίρο 30.991,94€ ενώ για την κατηγορία 3 οι μέσες τιμές γίνονται 371 μέρες, 4,69 παραγγελίες και 109.003,70€ τζίρος. Η κατηγορία 4 περιλαμβάνει πελάτες που είχαν συναλλαγή κατά μέσο όρο πριν 135 μέρες ενώ ο αριθμός των παραγγελιών και ο τζίρος είναι σχεδόν πενταπλάσιος συγκριτικά με αυτούς της κατηγορίας 3 με τιμές 27,49 και 518.248,23€ αντίστοιχα. Τέλος, η κατηγορία 5 περιλαμβάνει τους 74 καλύτερους πελάτες της επιχείρησης που κατά μέσο όρο είχαν συναλλαγή πριν 52 μέρες, πραγματοποίησαν 155,26 παραγγελίες και συνολικό μέσο τζίρο για την περίοδο αναφοράς 2.917.187,34€. Είναι προφανές ότι η αρχή του Pareto ισχύει απόλυτα στην περίπτωση της επιχείρησης καθώς αν πολλαπλασιάσουμε τον αριθμό των πελατών κάθε κατηγορίας με το μέσο τζίρο προκύπτει ότι οι πελάτες των κατηγοριών 4 και 5 αποτελούν το 78% του συνολικού τζίρου.

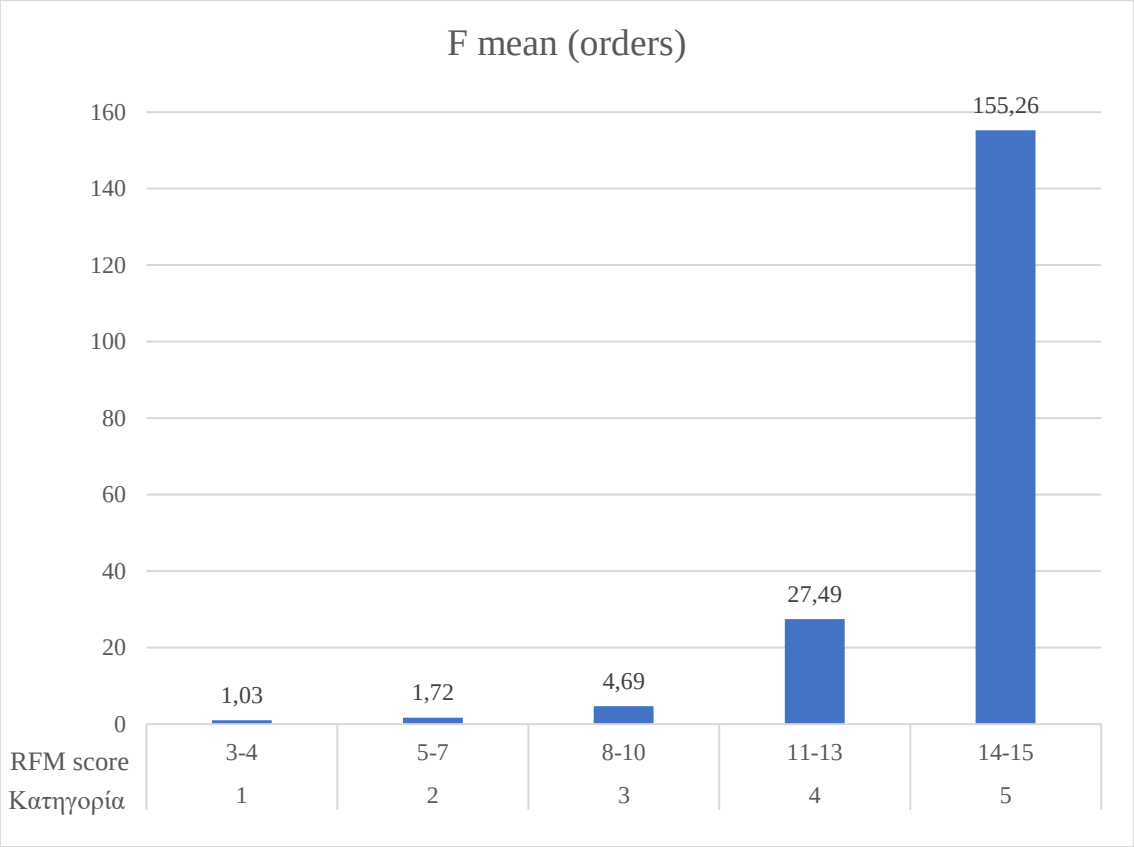
Πίνακας 15: Μέσες τιμές των μεταβλητών R, F και M ανά κατηγορία πελατών.

Category	RFM score	R mean (days)	F mean (orders)	M mean (value)
1	3-4	1455	1,03	13.276,42 €
2	5-7	835	1,72	30.991,94 €
3	8-10	371	4,69	109.003,70 €
4	11-13	135	27,49	518.248,23 €
5	14-15	52	155,26	2.917.187,34 €

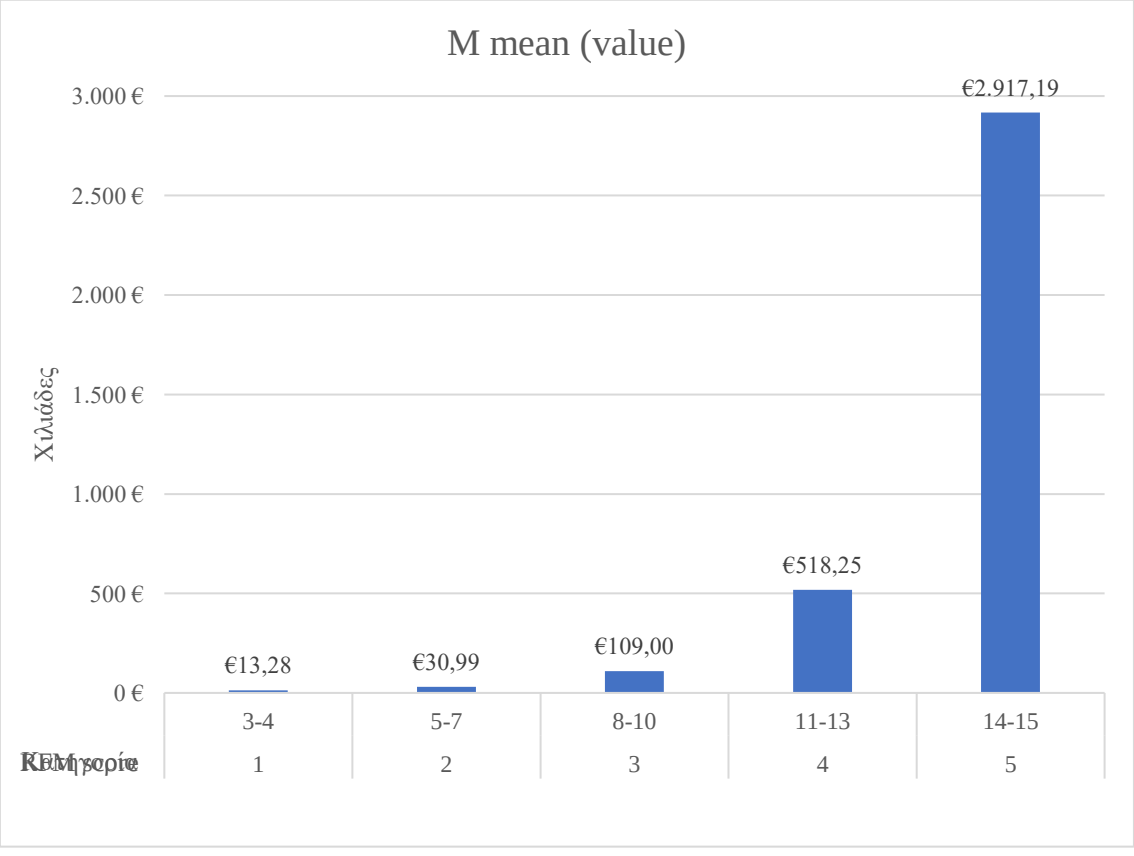
Τα στοιχεία του πίνακα 15 απεικονίζονται γραφικά ανά μεταβλητή στα ακόλουθα διαγράμματα (Διάγραμμα 22, Διάγραμμα 23 και Διάγραμμα 24) όπου οπτικοποιούνται οι διαφορές μεταξύ των 5 κατηγοριών και είναι ξεκάθαρο ποιοι πελάτες είναι σημαντικοί για την επιχείρηση και ποιοι όχι. Όπως αναφέρθηκε, οι πελάτες της κατηγορίας 3 οι οποίοι είναι συνολικά 623 από τους 1.716, είναι κρίσιμοι για την επιχείρηση καθώς μπορούν είτε να φθίνουν και να χαθούν είτε να ανεβάσουν τον τζίρο τους και να γίνουν σταθεροί πελάτες με προοπτικές ανάπτυξης.



Διάγραμμα 22: Μέση τιμή της μεταβλητής R (Recency) ανά κατηγορία πελατών.



Διάγραμμα 23: Μέση τιμή της μεταβλητής F (Frequency) ανά κατηγορία πελατών.



Διάγραμμα 24: Μέση τιμή της μεταβλητής M (Monetary) ανά κατηγορία πελατών.

Στο τελευταίο στάδιο της ανάλυσης RFM ερευνάται η απώλεια πελατών ως προς την τμηματοποίηση που έγινε με βάση το RFM score. Για το λόγο αυτόν εξετάζεται πόσοι πελάτες πήραν την ετικέτα churn σε κάθε κατηγορία και το αντίστοιχο ποσοστό τους σε σχέση με το συνολικό αριθμό πελατών στην κατηγορία. Από την ανάλυση των αποτελεσμάτων που φαίνονται στον πίνακα 16 και το αντίστοιχο διάγραμμα 25 προκύπτει ότι υπάρχει συσχέτιση μεταξύ του χαμηλού RFM score και του churn.

Πιο συγκεκριμένα, χαμηλό RFM score συνεπάγεται υψηλό ποσοστό churn. Μελετώντας τον πίνακα 16 και το διάγραμμα 25 συμπεραίνεται ότι όλοι οι πελάτες της κατηγορίας 1 που έχουν το χαμηλότερο RFM score, τελικά αποχωρούν από το πελατολόγιο της επιχείρησης και η τμηματοποίηση ως “Lost” είναι σωστή. Αν πολλαπλασιάσουμε τη μέση αξία συναλλαγών της κατηγορίας με τον αριθμό των πελατών που είναι 201, προκύπτει μια μέση απώλεια σε τζίρο 2.668.560,42€. Στην κατηγορία 2, το ποσοστό churn ανέρχεται σε 87,76% και επίσης συνάδει με την τμηματοποίηση των πελατών ως “At risk”. Με βάση τη μέση αξία συναλλαγών της κατηγορίας 2 για τους 545 πελάτες που αποχωρούν, η μέση αξία του χαμένου τζίρου ανέρχεται σε 16.890.607,30€ για την περίοδο αναφοράς.

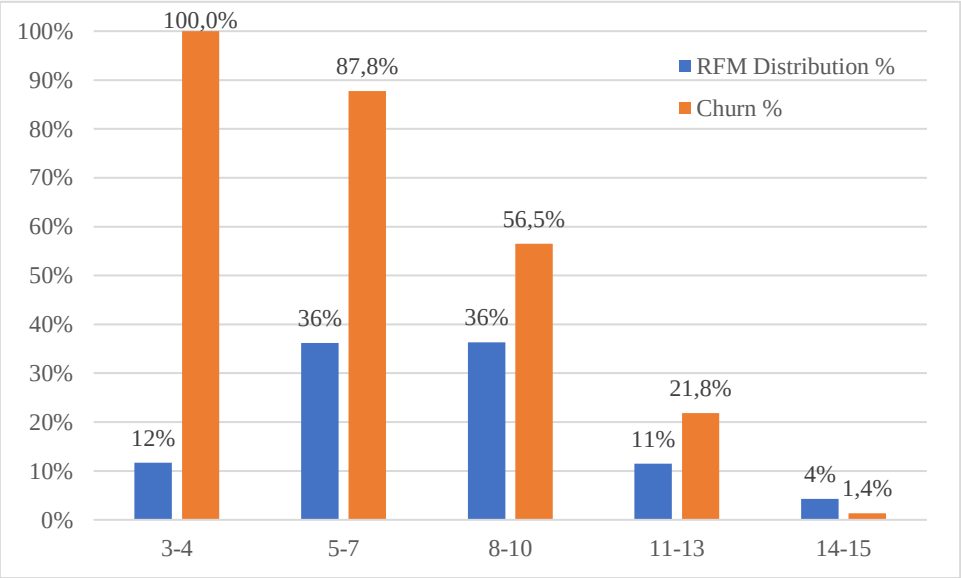
Στην κατηγορία 3, λίγο πάνω από τους μισούς πελάτες παίρνουν το χαρακτηρισμό churn και πιο συγκεκριμένα το ποσοστό ανέρχεται σε 56,5%. Οι πελάτες αυτοί βρίσκονται στη μέση της κλίμακας του RFM score που σημαίνει ότι γενικώς έχουν ένα μέτριο τζίρο και αριθμό παραγγελιών και πιθανώς έχουν αρκετό καιρό να πραγματοποιήσουν συναλλαγή με την επιχείρηση οπότε θα πρέπει να επιλεγούν ποιοι από αυτούς θα πρέπει να παραμείνουν στο πελατολόγιο με την κατάλληλη στρατηγική μάρκετινγκ. Καθώς βρίσκονται στο μεταίχμιο μεταξύ μη κερδοφόρων πελατών και καλών πελατών, σωστά χαρακτηρίζονται ως “Can’t loose them”. Η μέση αξία του χαμένου τζίρου λόγω απώλειας των 352 συγκεκριμένων πελατών ανέρχεται σε 38.369.302,40€.

Στην κατηγορία 4 που ανήκουν οι καλοί πελάτες της επιχείρησης, παρατηρείται ότι 21,83% παίρνει την ετικέτα churn γεγονός που πρέπει να προβληματίσει την επιχείρηση καθώς σχεδόν 1 στους 5 πελάτες της κατηγορίας αυτής, τελικά αποχωρούν. Αν μεταφράσουμε την απώλεια του πελάτη σε αξία συναλλαγών, αυτό σημαίνει κατά μέσο όρο 518.248,23€ ανά πελάτη, δηλαδή για τους 43 πελάτες που αποχώρησαν η μέση αξία χαμένου τζίρου είναι 22.284.673,89€. Τέλος, στην κατηγορία 5 που βρίσκονται οι

καλύτεροι πελάτες, το ποσοστό απώλειας είναι 1,35% και τελικά αποχωρεί 1 πελάτης με μέση αξία συναλλαγών 2.917.187,34€.

Πίνακας 16: Απώλεια πελατών (churn) ανά κατηγορία πελατών βάσει RFM-score.

Category	RFM score	Distribution	Churn	Churn %
1	3-4	201	201	100,0%
2	5-7	621	545	87,76%
3	8-10	623	352	56,50%
4	11-13	197	43	21,83%
5	14-15	74	1	1,35%
Σύνολο		**1.716**	**1.142**	



Διάγραμμα 25: Απώλεια πελατών (churn) ανά κατηγορία πελατών βάσει RFM-score.

Ως γενικό συμπέρασμα μπορεί να αναφερθεί ότι υπάρχει αρνητική συσχέτιση μεταξύ RFM score και churn, ειδικά στις κατηγορίες 1 και 2 παρατηρείται ότι οι περισσότεροι πελάτες τελικά αποχωρούν από την επιχείρηση. Αν αθροίσουμε τις μέσες αξίες του χαμένου τζίρου προκύπτει το ποσό των 19.559.167,72€. Ωστόσο ο χαμένος τζίρος της απώλειας πελατών από την κατηγορία 3 είναι διπλάσιος σε σχέση με τις κατηγορίες 1 και 2. Αρκετά υψηλός χαμένος τζίρος προκύπτει και από την απώλεια των πελατών της κατηγορίας 4 που είναι μεγαλύτερος από τον αντίστοιχο χαμένο τζίρο των κατηγοριών 1 και 2. Η απώλεια τζίρου από την απώλεια πελατών της κατηγορίας 5 είναι 2.917.187,34€ και παρότι είναι ο χαμηλότερος από όλες τις κατηγορίες, έχει το μεγαλύτερο αντίκτυπο για την επιχείρηση καθώς ο τζίρος αυτός ανήκει σε έναν και μόνο πελάτη οπότε κάθε απώλεια πελάτη από την κατηγορία 5 σημαίνει πολύ μεγάλη μείωση του τζίρου.

5.2. Τμηματοποίηση πελατών με τη μέθοδο K-means clustering

Η τμηματοποίηση πελατών με τη μέθοδο RFM έδωσε κάποια συμπεράσματα για το ποιοι πελάτες έχουν παρόμοιο προφίλ και κοινά χαρακτηριστικά ώστε να γίνει κατανοητό ποιοι τελικά πελάτες αποχωρούν από την επιχείρηση. Καθώς όμως ο διαχωρισμός των κλάσεων έγινε χειροκίνητα και επιλέχτηκαν 5 κλάσεις, θα εξεταστεί και η τμηματοποίηση πελατών με τη μέθοδο μηχανικής μάθησης K-means clustering για να δούμε αν θα προκύψουν παρόμοια αποτελέσματα.

Συνοπτικά, η μέθοδος έχει στόχο να ομαδοποιήσει ένα σύνολο δεδομένων σε κλάσεις που έχουν παρόμοια χαρακτηριστικά. Η διαδικασία έχει ως εξής:

1. Επιλογή του αριθμού των κλάσεων
2. Τυχαία επιλογή των κέντρων των K κλάσεων
3. Κατηγοριοποίηση των δεδομένων στο πιο κοντινό κέντρο της κλάσης, συνήθως με βάση την Ευκλείδεια απόσταση και υπολογισμός των νέων κέντρων των κλάσεων που είναι ο μέσος όρος των δεδομένων της κάθε κλάσης
4. Επανάληψη της διαδικασίας για ένα συγκεκριμένο αριθμό επαναλήψεων ή μέχρι οι κλάσεις να μην τροποποιούνται περεταίρω (Bagul, Surana, Berad, & Khachane, 2021).

Το πρόγραμμα που χρησιμοποιήθηκε για την ομαδοποίηση K-means clustering είναι το JMP 17.2.0 της εταιρείας SAS. Τα δεδομένα που χρησιμοποιήθηκαν είναι τα ίδια με αυτά της RFM ανάλυσης δηλαδή οι μεταβλητές R-score, F-score, M-score, RFM score και εξαρτημένη μεταβλητή το churn. Έγινε δοκιμή του αλγορίθμου με βάση τις παραπάνω μεταβλητές και η βέλτιστη λύση ήταν 5 κλάσεις, ίδιες δηλαδή σε αριθμό με την ανάλυση RFM. Ωστόσο έγινε δοκιμή του αλγορίθμου δίνοντας βαρύτητα στη μεταβλητή M-score καθώς ως εμπορική επιχείρηση αυτό που είναι σημαντικό για τη βιωσιμότητα της επιχείρησης είναι ο τζίρος. Αξίζει να σημειωθεί ότι έγινε και μία ακόμη δοκιμή δίνοντας βαρύτητα στο RFM-score και οι διαφορές στις κλάσεις και τους μέσους όρους είναι αμελητέα.

Ο αλγόριθμος δοκιμάστηκε για εύρος 2-10 κλάσεις και αποδείχτηκε ότι ο βέλτιστος αριθμός κλάσεων είναι 5. Το πρόγραμμα χρησιμοποιεί ως μέθοδο επαλήθευσης για κάθε δοκιμή αριθμού κλάσεων, το CCC score (Cubic Cluster Criterion) και προτείνει το βέλτιστο αριθμό που ομαδοποιεί τα δεδομένα καλύτερα. Στην εικόνα 16 φαίνεται το CCC

score για τις δοκιμές κλάσεων 2-10 και ότι ο ιδανικός αριθμός κλάσεων είναι 5, με CCC score 14,1389.

Εικόνα 16: CCC score της μεθόδου K-means clustering για αριθμό clusters 2-10.

Cluster Comparison

Method	NCluster	CCC	Best
K Means Cluster	5	14.1389	Optimal CCC
K Means Cluster	2	-7.4702	
K Means Cluster	3	-4.3734	
K Means Cluster	4	11.7793	
K Means Cluster	5	14.1389	
K Means Cluster	6	13.5605	
K Means Cluster	7	25.6195	
K Means Cluster	8	30.7763	
K Means Cluster	9	33.0442	
K Means Cluster	10	26.1456	

Όπως φαίνεται στην εικόνα 17, η κλάση 1 περιλαμβάνει 746 πελάτες, 220 πελάτες η κατηγορία 2, 252 πελάτες η κατηγορία 3, 104 πελάτες η κατηγορία 4 και 396 πελάτες η κατηγορία 5. Οι κλάσεις δεν είναι σε αύξουσα ή φθίνουσα σειρά σημαντικότητας πελατών, αυτό σημαίνει ότι π.χ. η κατηγορία 1 δε μπορεί να χαρακτηριστεί καλύτερη ή χειρότερη σε σχέση με την κατηγορία 2. Το αν η κλάση περιλαμβάνει καλούς ή μέτριους πελάτες ή αν περιλαμβάνει πελάτες που αποχωρούν από την επιχείρηση φαίνεται στην εικόνα 18 που φαίνονται οι μέσες τιμές των μεταβλητών ανά κλάση.

Εικόνα 17: Κατανομή πελατών σε κλάσεις με τη μέθοδο K-means clustering.

Cluster Summary

Cluster	Count	Step
1	746	11
2	220	
3	252	
4	104	
5	394	

Εικόνα 18: Μέσες τιμές μεταβλητών που χρησιμοποιήθηκαν στη μέθοδο K-means clustering ανά κλάση.

Cluster Means

Cluster	Churn	R-score	F-score	M-score	RFM-score
1	1	2.4701815	1.38115817	1.82627485	5.67761452
2	0	5	1.24864865	1.86486486	8.11351351
3	0	5	2.52552927	3.23412204	10.7596513
4	0.01814516	4.98991935	4.38306452	4.80645161	14.1794355
5	1	3.94639719	2.26362039	2.99824253	9.20826011

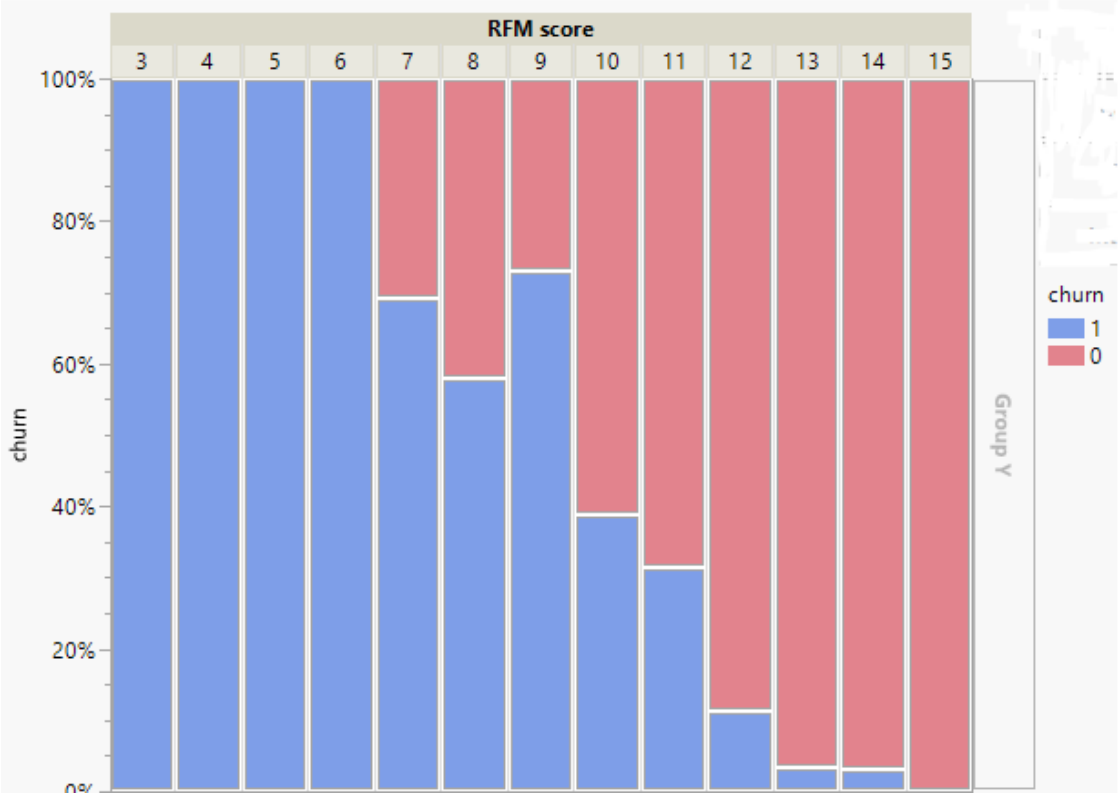
Παρατηρώντας τις κλάσεις φαίνεται ότι η κλάση 1 περιλαμβάνει πελάτες που αποχωρούν από την επιχείρηση (churn=1) ενώ έχουν επίσης τα χαμηλότερα κέντρα για όλες τις μεταβλητές, εκτός από την F-score που είναι 1,38 και είναι η δεύτερη χαμηλότερη μεταξύ των κλάσεων. Οι πελάτες αυτοί δηλαδή έχουν αποχωρήσει, έχουν καιρό να πραγματοποιήσουν συναλλαγή με την επιχείρηση, έχουν μικρό αριθμό παραγγελιών και μικρό τζίρο επομένως μπορούν να χαρακτηριστούν κακοί ή Lost πελάτες όπως και στην ανάλυση RFM. Συγκρίνοντας το συνολικό αριθμό των πελατών που είναι 746, παρατηρείται ότι ακριβώς ο ίδιος αριθμός προκύπτει στην ανάλυση RFM αν προσθέσουμε τους πελάτες που αποχώρησαν (churn) της κλάσης 1 και 2. Όπως είδαμε παραπάνω, οι πελάτες της κλάσης 1 έχουν 100% ποσοστό αποχώρησης ενώ η κατηγορία 2 έχει ποσοστό 87,8%, δηλαδή αν πάρουμε το σταθμισμένο μέσο όρο προκύπτει ένα ποσοστό 91,1%. Το ίδιο συμπέρασμα προκύπτει και από την κλάση 1 της μεθόδου K-means clustering καθώς η μέση τιμή του churn είναι 1 δηλαδή 100% των πελατών αποχωρούν.

Η κλάση 2 περιλαμβάνει 220 πελάτες που δεν αποχωρούν από την επιχείρηση (churn=0) έχουν μέσο R-score 5, δηλαδή πολύ πρόσφατη συναλλαγή, έχουν το χαμηλότερο μέσο F-score 1,25, δηλαδή μικρό αριθμό παραγγελιών, μέτριο M-score δηλαδή όχι πολύ μεγάλο τζίρο και σχετικά μέτριο RFM-score 8,11. Οι πελάτες δηλαδή χαρακτηρίζονται ότι παραμένουν στο πελατολόγιο βάσει του υψηλού R-score. Παρόμοια χαρακτηριστικά έχει η κλάση 3 με μέσο R-score 5, F-score 2,53, M-score 3,23 και RFM-score 10,76, δηλαδή 252 πελάτες που είχαν πρόσφατη συναλλαγή με μικρό σχετικά αριθμό παραγγελιών, μέτριο προς υψηλό τζίρο και το δεύτερο καλύτερο RFM-score δηλαδή θεωρούνται καλοί πελάτες.

Στην κλάση 4 ανήκουν οι 104 καλύτεροι πελάτες με τα υψηλοτέρα score σε όλες τις μεταβλητές, δηλαδή R-score 4,99, F-score 4,38, M-score 4,81 και RFM-score 14,18. Η μεταβλητή churn έχει μέση τιμή 0,018, δηλαδή όχι 0 όπως οι κλάσεις 2 και 3 οπότε αν πολλαπλασιάσουμε το 0,018 με το σύνολο των πελατών της κλάσης προκύπτει 1,89, δηλαδή περίπου 2 πελάτες. Αντίστοιχα στη μέθοδο RFM προέκυψε 1 πελάτης που αποχώρησε που είναι πολύ κοντά στον αριθμό 2 της K-means clustering.

Τέλος, η κλάση 5 περιλαμβάνει επίσης πελάτες με churn=1 με συνολικό αριθμό 394. Το προφίλ των πελατών αυτών είναι ότι έχουν τη δεύτερη χαμηλότερη μέση τιμή για το R-score, ενώ έχουν την τρίτη χαμηλότερη μέση τιμή για το F-score, M-score και RFM-score αντίστοιχα. Όπως φαίνεται ο αλγόριθμος K-means clustering δίνει μεγαλύτερη βαρύτητα στο R-score που μετράει τις μέρες από την τελευταία συναλλαγή με την επιχείρηση κάτι που είναι λογικό καθώς όσο περισσότερες οι μέρες τόσο πιο πιθανός είναι ο χαρακτηρισμός του πελάτη ως χαμένος (churn). Οι πελάτες αυτοί έχουν μέτριο RFM score 9,21 και είναι αντίστοιχοι της κλάσης 3 στην RFM ανάλυση με το χαρακτηρισμό Can’t loose them όπου το ποσοστό αποχώρησης ήταν 56,5%, δηλαδή αρκετά πιθανό. Ο αριθμός πελατών της κλάσης αυτής όπως είχαμε δει είναι 352 χαμένοι πελάτες και το νούμερο είναι πολύ κοντά στο 394 της τμηματοποίησης που έκανε ο αλγόριθμος K-means clustering. Αν προσθέσουμε τους χαμένους πελάτες της ανάλυσης RFM που περιλαμβάνονται στις κατηγορίες 3 και 4 θα δούμε ότι το σύνολο των πελατών που αποχωρούν είναι 395 και ο αριθμός αυτός είναι σχεδόν ίσος με το 394 που προκύπτει από την τμηματοποίηση K-means clustering. Ο πίνακας 17 δίνει καλύτερη εικόνα των κλάσεων για τις δύο μεθόδους.

Στην εικόνα 19 φαίνονται τα ραβδογράμματα με το ποσοστό απώλειας πελατών ανά RFM-score. Οι πελάτες με RFM-score 3-6 έχουν ποσοστό churn 100%, δηλαδή όλοι αποχωρούν από την επιχείρηση ενώ οι πελάτες με RFM-score 15 έχουν ποσοστό churn 0%, δηλαδή όλοι παραμένουν στο πελατολόγιο.



Εικόνα 19: Ποσοστό churn ανά RFM-score πελατών.

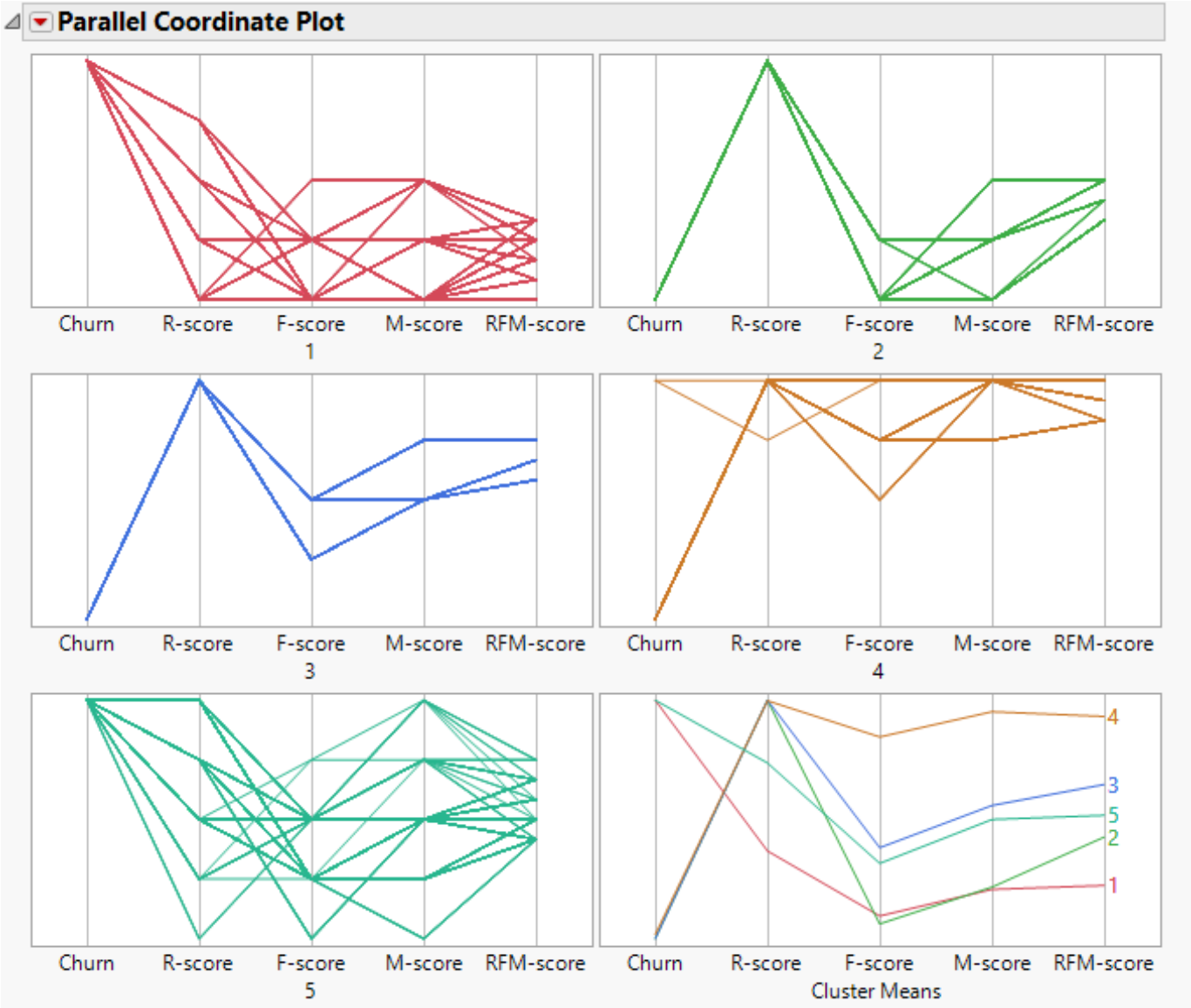
Η εικόνα 20 αναπαριστά γραφικά την κατανομή των πελατών σε κάθε κλάση ανάλογα με τις τιμές των μεταβλητών churn R-score, F-score, M-score και RFM-score. Όπως φαίνεται η κλάση 1 έχει μέτρια ομοιογένεια καθώς οι τιμές churn είναι όλες 1 και τα RFM-score είναι αρκετά κάτω από τη μέση δηλαδή κάτω από 7,5. Από την άλλη, οι τιμές R-score, F-score και M-score έχουν διασπορά τιμών ειδικά στη μεταβλητή R-score που παίρνει όλες τις τιμές εκτός από 5. Παρόλα αυτά, η γραφική απεικόνιση μας δείχνει τα βασικά χαρακτηριστικά της κλάσης που είναι churn=1, δηλαδή απώλεια των πελατών και χαμηλά RFM-score.

Οι κλάσεις 2 και 3 έχουν μεγαλύτερη ομοιογένεια και παρόμοια χαρακτηριστικά πελατών. Όπως φαίνεται, όλοι οι πελάτες της κατηγορίας 2 έχουν churn=0, R-score =5 και χαμηλά F-score 1 ή 2. Το M-score κυμαίνεται 1-3 δηλαδή χαμηλή και μέση αξία αγορών ενώ το RFM-score παίρνει τιμές 7, 8 και 9 δηλαδή αποκλειστικά στο μέσο της κλίμακας. Η κλάση 3 έχει πελάτες με churn=0, R-score=5 και μεγάλη ομοιογένεια στις υπόλοιπες μεταβλητές όπου οι τιμές κυμαίνονται 2 ως 3 για το F-score, 3 ως 4 για το M-score και 10, 11 και 12 για το RFM-score.

Η κλάση 4 έχει τη μεγαλύτερη ομοιογένεια με σχεδόν όλους τους πελάτες εκτός από δύο να έχουν churn=1 και όλοι εκτός από έναν να έχουν τιμή R-score=5. Οι πελάτες της κλάσης αυτής έχουν F-score 3, 4 ή 5 ενώ το M-score είναι 4ή 5 δηλαδή μόνο μεγάλες αξίες αγορών. Τέλος, το RFM-score κυμαίνεται παίρνει τιμές 13, 14 ή 15 δηλαδή μόνο υψηλές τιμές που τονίζει τη σπουδαιότητα των πελατών αυτών.

Η κλάση 5 έχει τη μεγαλύτερη ανομοιογένεια στις τιμές πέρα από το churn και το RFM-score. Το churn είναι 1 για όλους τους πελάτες, δηλαδή αποχωρούν από την επιχείρηση ενώ το RFM-score παίρνει τιμές 8-12 δηλαδή τιμές περίπου στο μέσο της κλίμακας και αυτή είναι η μεγαλύτερη διαφορά της κλάσης αυτής με την κλάση 1 που έχει πελάτες με χαμηλό RFM-score. Η μεταβλητές R-score και M-score παίρνουν όλες τις δυνατές τιμές 1-5 και η μεταβλητή F-score παίρνει τιμές 1-4. Όπως παρατηρείται και στο γράφημα, πέρα από το churn και ίσως το RFM-score, όλες οι άλλες μεταβλητές έχουν μεγάλη διασπορά και οι πελάτες παρότι χαρακτηρίζονται ως churn, τα υπόλοιπα χαρακτηριστικά τους δεν είναι ιδιαίτερα ομοιογενή.

Στο τελευταίο γράφημα της εικόνας 20, φαίνονται όλες οι κλάσεις με τις μέσες τιμές τους ανά μεταβλητή και είναι μια καλή απεικόνιση των διαφορών μεταξύ των 5 κλάσεων.



Εικόνα 20: Γραφήματα των 5 κλάσεων και των τιμών όλων των μεταβλητών.

5.3. Συνδυαστικό συμπέρασμα ανάλυσης RFM και μεθόδου K-means clustering

Από την τμηματοποίηση με ανάλυση RFM και K-means clustering προκύπτουν δύο βασικές ενδείξεις για την επιχείρηση. Η πρώτη αφορά στους χαμένους κακούς πελάτες, δηλαδή τους πελάτες που ναι μεν αποχώρησαν από το πελατολόγιο αλλά η επιχείρηση δεν πρέπει απαραίτητα να κάνει κάποια ενέργεια για να τους κρατήσει. Είναι οι 746 πελάτες που δεν έχουν μεγάλο αριθμό παραγγελιών ούτε μεγάλο τζίρο και κατανέμονται στις κλάσεις 1 και 2 των δύο μεθόδων.

Η δεύτερη και ίσως βασικότερη ένδειξη αφορά στους 396 καλούς ή πολύ καλούς πελάτες που αποχώρησαν και η επιχείρηση πρέπει να προβληματιστεί για τους λόγους των αποχωρήσεων. Οι πελάτες αυτοί έχουν μέσους έως μεγάλους τζίρους και θα πρέπει να εξεταστούν ένας προς έναν για το αν η επιχείρηση αξίζει να κάνει κάποια ενέργεια ώστε να τους διατηρήσει στο πελατολόγιό της. Ειδικότερα οι 394 πελάτες που περιλαμβάνονται στην κλάση 5 της τμηματοποίησης K-means clustering όπως φαίνεται και στην εικόνα 20 παραπάνω μπορεί να έχουν υψηλούς ως πολύ υψηλούς τζίρους που μπορεί να έχουν μεγάλο αντίκτυπο στα έσοδα και τα αποτελέσματα της επιχείρησης.

Αν αθροιστούν οι πελάτες που αποχώρησαν με τη μέθοδο RFM το σύνολο είναι 1.142 πελάτες, ενώ με τη μέθοδο K-means clustering 1.141,89, ουσιαστικά δηλαδή δίνουν ακριβώς ίδια αποτελέσματα. Ο πίνακας 17 μας δίνει τη συνολική εικόνα των δεδομένων για τις δύο μεθόδους και αντιστοιχίζονται χρωματικά οι χαμένοι πελάτες (churn).

Πίνακας 17: Κατανομή συνολικών πελατών και χαμένων πελατών (churn) ανά κλάση και ανά μέθοδο τμηματοποίησης.

	Category	Distribution	Churn rate	Churn
RFM analysis	**1**	**201**	**100,00%**	**201**
	2	**621**	**87,76%**	**545**
	3	**623**	**56,50%**	**352**
	4	**197**	**21,83%**	**43**
	5	**74**	**1,35%**	**1**
	Total	1.716		1.142

	Category	Distribution	Churn rate	Churn
K-means clustering	**1**	**746**	**100,00%**	**746**
	2	**220**	**0,00%**	**0**
	3	**252**	**0,00%**	**0**
	4	**104**	**1,81%**	**1,89**
	5	**394**	**100,00%**	**394**
	Total	1.716		1.141,89

6. Παράδειγμα πρόβλεψης απώλειας πελατών

6.1. Προ-επεξεργασία δεδομένων

Για τη δόμηση των μοντέλων στο πρόγραμμα JMP, χρειάστηκε κάποιες μεταβλητές να μετασχηματιστούν σε αριθμητικές ώστε να μπορέσουν να τρέξουν τα μοντέλα. Αρχικά, η μεταβλητή CustomerName code είναι η περιγραφική μεταβλητή που χαρακτηρίζει κάθε

πελάτη και είναι μοναδική για τον καθένα (ID). Όσες μεταβλητές απαντούν σε ερωτήματα ναι/όχι ορίστηκαν ως κατηγορικές με τιμή 1/0, αυτές είναι η μεταβλητή hasCreditlimit που ορίζει αν ο πελάτης έχει πιστωτικό όριο ή όχι, η μεταβλητή Portal access που ορίζει αν ένας πελάτης έχει πρόσβαση ή όχι στο σύστημα παραγγελιοληψίας Portal, η μεταβλητή Installer που ορίζει αν ο πελάτης είναι εγκαταστάτης μηχανολογικού εξοπλισμού ή όχι, η μεταβλητή eshop's users που ορίζει αν ο πελάτης έχει πρόσβαση στο ηλεκτρονικό κατάστημα ανταλλακτικών, η μεταβλητή churn που ορίζει την απώλεια πελάτη και η μεταβλητή missing parts που ορίζει αν ο πελάτης είχε κάποια παραγγελία με ελλείψεις ή παραλείψεις υλικών.

Στα δεδομένα υπάρχουν μεταβλητές που δεν είναι αριθμητικές και πρέπει με κάποιο τρόπο να κωδικοποιηθούν ώστε το πρόγραμμα να καταλάβει στη συνέχεια αν είναι κατηγορική (categorical), διάταξης (ordinal) ή συνεχής (continuous). Η κατηγορική μεταβλητή SalesPerson code περιγράφει τον πωλητή που διαχειρίζεται το συγκεκριμένο πελάτη και προφανώς οι πωλητές έχουν παραπάνω του ενός πελάτη. Η μεταβλητή Country code είναι κατηγορική με διαφορετικές τιμές ανά χώρα που έχουν τιμές 1-113. Η μεταβλητή Manufacturer code περιγράφει το εργοστάσιο κατασκευής της παραγγελίας που όπως είδαμε μπορεί να είναι Europe 1, Europe 2 ή China. όμως υπάρχουν πελάτες που έχουν προμηθευτεί παραγγελίες από δύο διαφορετικά εργοστάσια και για το λόγο αυτό δημιουργήθηκαν οι εξής κατηγορίες με τις αντίστοιχες τιμές: Europe 1 με τιμή 1, Europe 2 με τιμή 2, China με τιμή 3, Europe 1 και Europe 2 με τιμή 4 και τέλος Europe 1 και China με τιμή 5.

Όταν οι τιμές μια κατηγορικής μεταβλητής παίρνει διακριτές τιμές και ορίζει τη σειρά ή το μέγεθος τότε ονομάζονται τακτικές. Η μεταβλητή Product code που χαρακτηρίζει τα διαφορετικά προϊόντα της επιχείρησης αν παραμείνει ως product 1, product 2 κτλ. δεν έχει να προσφέρει κάτι στην ανάλυση και αποφασίστηκε να δημιουργηθεί η νέα μεταβλητή Product types per customer που ορίζει πόσα διαφορετικά προϊόντα έχει προμηθευτεί ένας πελάτης . Καθώς ο αριθμός προϊόντων ανά πελάτη ποικίλλει από 1 ως 22 και η τιμή 1 αφορά 1017 πελάτες, αποφασίστηκε να δημιουργηθούν 3 κατηγορίες: 1 προϊόν με τιμή 1, 2-5 προϊόντα με τιμή 2 και πάνω από 6 προϊόντα με τιμή 3. Η μεταβλητή Company size εμπίπτει στην κατηγορία αυτή και όπως είδαμε στο κεφάλαιο 5 χαρακτηρίζει το μέγεθος των πελατών ως small για πελάτες με 1-9 εργαζομένους, medium για πελάτες με 10-29 εργαζομένους και big για πελάτες με πάνω από 30 εργαζομένους. Για να γίνει η μεταβλητή διάταξης επεξεργάσιμη στο πρόγραμμα JMP,

δόθηκε η τιμή 1 για την κατηγορία small, 2 για την κατηγορία medium και 3 για την κατηγορία big. Παρόμοια λογική ακολουθήθηκε για τη μεταβλητή years of collaboration που ορίζει τα χρόνια συνεργασίας των πελατών με την επιχείρηση που μπορεί να είναι 0-2, 3-5 και πάνω από 6. Η κωδικοποίηση της μεταβλητής διάταξης έγινε δίνοντας την τιμή 1 για 0-2 χρόνια συνεργασίας, την τιμή 2 για 3-5 χρόνια συνεργασίας και την τιμή 3 για πάνω από 6 χρόνια συνεργασίας. Για τη μεταβλητή Recency που ορίζει το χρόνο από την πιο πρόσφατη παραγγελία του πελάτη, έπρεπε για κάθε πελάτη να βρεθεί η ημερομηνία της πιο πρόσφατης παραγγελίας και να υπολογιστούν οι μέρες από την ημερομηνία αναφοράς που είναι η 31/12/2022 και να οριστεί σε ημέρες. Για παράδειγμα μια παραγγελία με ημερομηνία 20/10/2022 απέχει 72 μέρες από την ημερομηνία 31/12/2022. Η μεταβλητή αυτή ονομάστηκε date of last transaction.

Η μεταβλητή churn times, δηλαδή πόσες φορές μέσα στη περίοδο αναφοράς (2018-2022) έχει αποχωρήσει ο πελάτης θεωρείται και αυτή διάταξης. Όπως είχε οριστεί, ένας πελάτης χαρακτηρίζεται χαμένος (churn) αν δεν προβεί σε παραγγελία για ένα ημερολογιακό έτος οπότε αυτό σημαίνει ότι στη διάρκεια των πέντε ετών της περιόδου αναφοράς μπορεί να αποχωρήσει και να επιστρέψει περισσότερες από μία φορά. Το ίδιο ισχύει και για τη μεταβλητή times of missing parts που ορίζει πόσες φορές, δηλαδή σε πόσες διαφορετικές παραγγελίες υπήρχαν ελλείψεις ή παραλείψεις υλικών. Μεταβλητή διάταξης είναι επίσης η total number of missing parts που περιγράφει το συνολικό αριθμό των υλικών σε έλλειψη για όλες τις παραγγελίες του πελάτη όπως επίσης και η μεταβλητή Frequency που περιγράφει το συνολικό αριθμό παραγγελιών του πελάτη.

Μεταβλητές διάταξης είναι επίσης οι R-score, F-score και M-score που παίρνουν τιμές 1-5 με το 1 να χαρακτηρίζει το χειρότερο σκορ και το 5 το καλύτερο σκορ. Κατ' αντιστοιχία η μεταβλητή διάταξης RFM-score παίρνει τιμές 3-15 καθώς προκύπτει από το άθροισμα των επιμέρους σκορ των μεταβλητών R-score, F-score και M-score.

Τέλος, η μοναδική συνεχής μεταβλητή είναι η Total Value που είναι το άθροισμα της αξίας των παραγγελιών ανά πελάτη σε ευρώ. Στον πίνακα 18 φαίνονται όλες οι μεταβλητές και ο τύπος τους όπως ορίστηκαν στο πρόγραμμα JMP.

Πίνακας 18: Τύπος μεταβλητών.

Μεταβλητές	Τύπος
SalesPerson code	Κατηγορική
CustomerName code	Κατηγορική
Product type per customer	Διάταξης
date of last transaction	Διάταξης
Total Value	Συνεχής
Country code	Κατηγορική
Manufacturer code	Κατηγορική
hasCreditLimit	Κατηγορική
Portal access	Κατηγορική
Company size	Διάταξης
Installer	Κατηγορική
years of collaboration	Διάταξης
eshop's users	Κατηγορική
churn	Κατηγορική
churn times	Διάταξης
missing parts	Κατηγορική
times of missing parts	Διάταξης
total number of missing parts	Διάταξης
Frequency	Διάταξης
RFM score	Διάταξης
R-score	Διάταξης
F-score	Διάταξης
M-score	Διάταξης

Σε όλες τις μεταβλητές δεν υπάρχουν ελλείπουσες τιμές ενώ δεν υπάρχουν και outliers για τον αριθμό παραγγελιών και τη συνολική αξία καθώς από τις παραγγελίες εξαιρούνται αυτές των ανταλλακτικών. Η επιχείρηση διασφαλίζει την πληρότητα των στοιχείων των πελατών και παραγγελιών καθώς δε μπορούν να προωθηθούν στην παραγωγή αν υπάρχουν κενά πεδία στη βάση δεδομένων της.

Από τις παραπάνω 23 μεταβλητές, η μεταβλητή CustomerName code ουσιαστικά είναι το αναγνωριστικό (ID) του κάθε πελάτη οπότε οι μεταβλητές που μπορούν να

χρησιμοποιηθούν είναι 22 συνολικά. Για τις μεταβλητές αυτές, έγινε ανάλυση συσχέτισης στο πρόγραμμα JMP για να ελεγχθεί ποιες ανεξάρτητες μεταβλητές έχουν μεγάλη συσχέτιση μεταξύ τους και αν θα πρέπει κάποια από αυτές να απαλειφθούν.

Η ανάλυση συσχέτισης εξετάζει τις μεταβλητές σε ζεύγη και υπολογίζει το συντελεστή συσχέτισης Pearson που κυμαίνεται μεταξύ -1 και 1, δηλαδή πλήρη αρνητική ως πλήρη θετική συσχέτιση. Ο συντελεστής Pearson είναι ο πιο διαδεδομένος και εξετάζει την πιθανή γραμμική σχέση μεταξύ δύο μεταβλητών είναι όμως ευαίσθητος στις ακραίες τιμές. Για το λόγο αυτόν, υπολογίστηκε και ο συντελεστής συσχέτισης Spearman που είναι η μη παραμετρική εκδοχή του συντελεστή συσχέτισης Pearson και χρησιμοποιείται όταν η σχέση μεταξύ δύο μεταβλητών είναι μη γραμμική. Στην περίπτωση της παρούσας εργασίας οι συντελεστές Pearson και Spearman δίνουν παρόμοιες ενδείξεις και δε φαίνεται να επηρεάζονται από ακραίες τιμές.

Η εικόνα 21 δείχνει σε μορφή πίνακα το συντελεστή Pearson μεταξύ όλων των μεταβλητών σε ζεύγη από όπου μπορούν να βγουν κάποια συμπεράσματα για το ποιες μεταβλητές συσχετίζονται και το αν θα πρέπει να γίνουν απαλοιφές. Ως κρίσιμο κατώφλι ορίστηκε η τιμή 0,7, δηλαδή 70% που υποδεικνύει ισχυρή συσχέτιση.

Οι μεταβλητές με υψηλές συσχετίσεις είναι κυρίως αυτές που έχουν να κάνουν με τον αριθμό παραγγελιών, την ημερομηνία της πιο πρόσφατης παραγγελίας και τη συνολική αξία παραγγελιών. Μια μεταβλητή που αξίζει σχολιασμού είναι η Product types per customer όπου εμφανίζει υψηλή συσχέτιση 0,7225 με το RFM-score, 0,7820 με το F-score και με το M-score 0,7061. Όπως προκύπτει, όσο περισσότερα προϊόντα προμηθεύεται ένας πελάτης τόσο υψηλότερο RFM-score, F-score και M-score έχει χωρίς όμως να υπάρχει κάποια ξεκάθαρη αιτιότητα οπότε η μεταβλητή Product types per customer διατηρείται στη λίστα. Υψηλή αρνητική συσχέτιση -0,7833 εμφανίζεται μεταξύ της μεταβλητής date of last transaction και RFM-score που εξηγείται από το γεγονός ότι η ημερομηνία τελευταίας συναλλαγής εμπεριέχεται στο RFM-score καθώς ορίζει ουσιαστικά το R-score. Αυτό ακριβώς υποδηλώνει και η πολύ υψηλή αρνητική συσχέτιση μεταξύ date of last transaction και R-score καθώς περιγράφουν το ίδιο χαρακτηριστικό από την αντίθετη σκοπιά. Όσο μικρότερος ο αριθμός ημερών από την τελευταία συναλλαγή τόσο υψηλότερο το R-score, δηλαδή πιο πρόσφατη η συναλλαγή. Στην περίπτωση αυτή θα πρέπει να εξεταστεί αν κάποια από τις δύο μεταβλητές θα πρέπει να απαλειφθεί που όπως αναφέρεται στη συνέχεια η μεταβλητή αυτή θα είναι η R-score.

Η μεταβλητή Total value παρουσιάζει υψηλή θετική συσχέτιση 0,9314 με τη μεταβλητή Frequency και είναι πολύ φυσιολογικό καθώς όσο αυξάνει ο αριθμός παραγγελιών τόσο αυξάνει και ο συνολικός τζίρος των παραγγελιών. Υψηλή θετική συσχέτιση 0,9329 εμφανίζουν και οι μεταβλητές churn και churn times καθώς το churn times μετράει πόσες φορές έχει αποχωρήσει ένας πελάτης οπότε αν πάρει την τιμή 1 ή παραπάνω, υπονοεί ότι και η μεταβλητή churn. Το churn times είναι παραπλανητική μεταβλητή και αποφασίστηκε να απαλειφθεί από την ανάλυση. Παρόμοια περίπτωση αποτελεί και το ζεύγος μεταβλητών missing parts και times of missing parts που ουσιαστικά περιγράφουν το ίδιο χαρακτηριστικό και μπορεί μία εκ των δύο να παραληφθεί οπότε απαλείφεται η μεταβλητή times of missing parts.

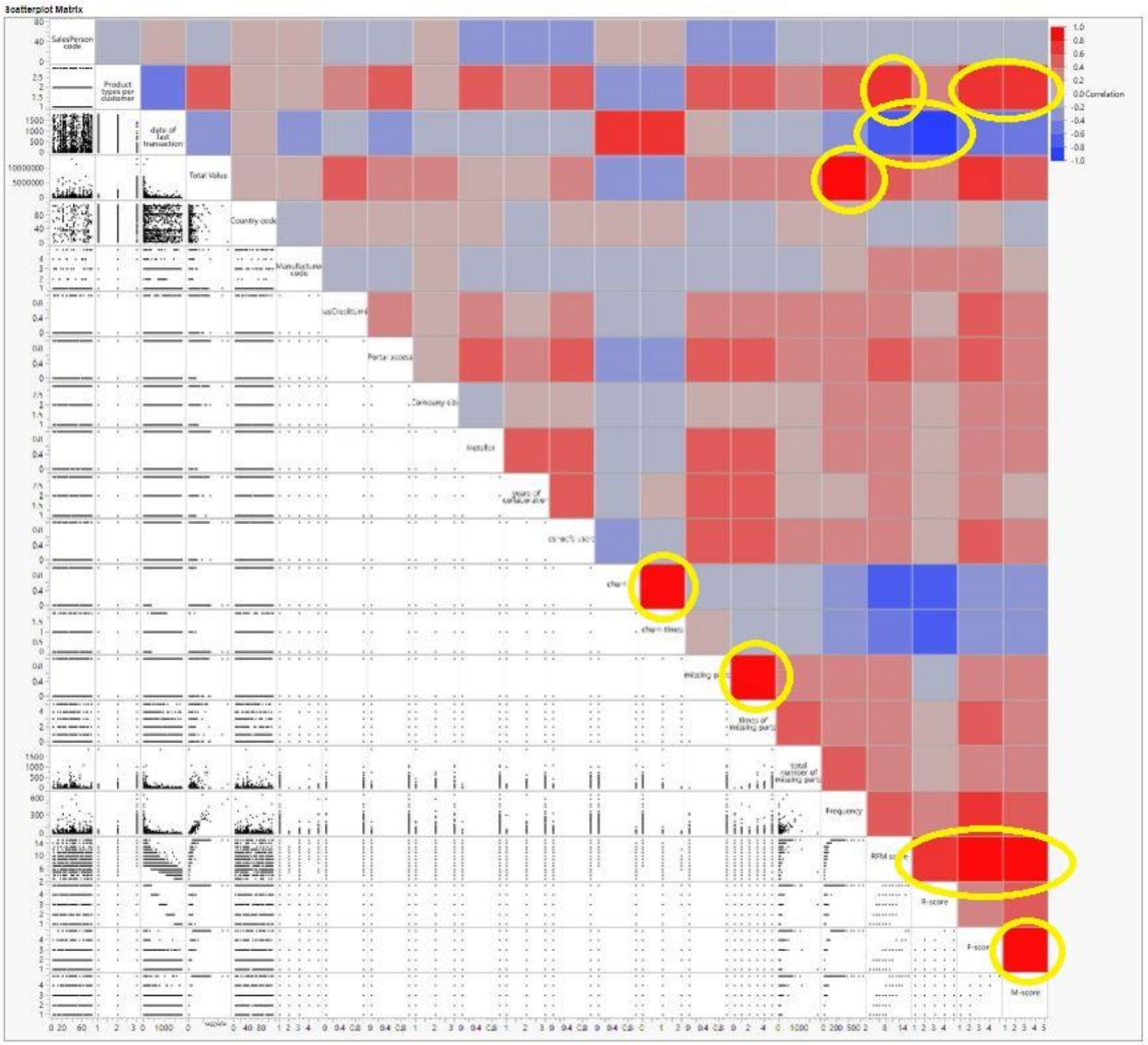
Τέλος, υψηλές συσχετίσεις προκύπτουν μεταξύ RFM-score και των μεταβλητών R-score, F-score και M-score με συντελεστές 0,7677, 0,8477 και 0,8732 αντίστοιχα. Καθώς το συνολικό RFM-score είναι ουσιαστικά το άθροισμα των R-score, F-score και M-score, θα πρέπει είτε να απαλειφθεί το RFM-score είτε οι άλλες τρεις μεταβλητές. Τελικά κρίθηκε να απαλειφθούν οι μεταβλητές R-score, F-score και M-score καθώς ουσιαστικά αποτελούν προεκτάσεις των μεταβλητών date of last transaction (Recency), Frequency και Total value (Monetary) που περιλαμβάνονται ήδη στη λίστα μεταβλητών.

Correlations

	SalesPerson code	Product types per customer	date of last transaction	Total Value	Country code	Manufacturer code	hasCreditLimit	Portal access	Company size	Installer	years of collaboration	eshop's users	churn	churn times	missing parts	times of missing parts	total number of missing parts	Frequency	RFM score	R-score	F-score	M-score
SalesPerson code	1.0000	-0.1258	0.0623	-0.0451	0.1861	0.0518	-0.0959	-0.1445	0.0855	-0.2306	-0.2705	-0.2290	0.0379	0.0207	-0.2492	-0.2229	-0.0581	-0.0346	-0.0705	-0.0668	-0.0539	-0.0511
Product types per customer	-0.1258	1.0000	-0.4046	0.4785	0.0312	0.1358	0.3866	0.4966	0.1974	0.4178	0.3400	0.4738	-0.3287	-0.2579	0.4149	0.4969	0.2759	0.4679	0.7225	0.3745	0.7820	0.7061
date of last transaction	0.0623	-0.4046	1.0000	-0.2479	0.0377	-0.2211	-0.1374	-0.2751	-0.0533	-0.0750	-0.0511	-0.1999	0.6920	0.6107	0.0123	-0.1260	-0.1388	-0.2525	-0.7833	-0.9797	-0.4114	-0.4365
Total Value	-0.0451	0.4785	-0.2479	1.0000	0.0441	0.1228	0.4084	0.3102	0.2451	0.1930	0.1720	0.2679	-0.2875	-0.2648	0.2108	0.3029	0.3876	0.9314	0.5384	0.2182	0.6403	0.5512
Country code	0.1861	0.0312	0.0377	0.0441	1.0000	-0.0837	0.0566	0.0307	0.0128	0.0720	-0.0021	0.0288	0.0203	0.0327	-0.0427	-0.0575	0.0014	0.0284	-0.0437	-0.0464	0.0063	-0.0598
Manufacturer code	0.0518	0.1358	-0.2211	0.1228	-0.0837	1.0000	-0.0881	-0.0125	0.1085	-0.0219	-0.0533	-0.1401	-0.1920	-0.1843	-0.1676	-0.1154	-0.0386	0.1190	0.2433	0.2011	0.2083	0.1930
hasCreditLimit	-0.0959	0.3866	-0.1374	0.4084	0.0566	-0.0881	1.0000	0.2765	0.1669	0.2238	0.1530	0.3554	-0.1775	-0.1446	0.2646	0.3287	0.2579	0.3965	0.3408	0.1204	0.4128	0.3643
Portal access	-0.1445	0.4966	-0.2751	0.3102	0.0307	-0.0125	0.2765	1.0000	0.0584	0.5410	0.3276	0.5963	-0.2863	-0.2435	0.4069	0.4469	0.2147	0.3260	0.4299	0.2555	0.4713	0.3747
Company size	0.0855	0.1974	-0.0533	0.2451	0.0128	0.1085	0.1669	0.0584	1.0000	-0.0261	0.0466	0.0506	-0.0731	-0.0676	0.0526	0.0929	0.1201	0.2259	0.2451	0.0456	0.2810	0.3264
Installer	-0.2306	0.4178	-0.0750	0.1930	0.0720	-0.0219	0.2238	0.5410	-0.0261	1.0000	0.5711	0.5144	-0.1154	-0.0657	0.4826	0.4366	0.1690	0.2130	0.2326	0.0537	0.3701	0.2071
years of collaboration	-0.2705	0.3400	-0.0511	0.1720	-0.0021	-0.0533	0.1530	0.3276	0.0466	0.5711	1.0000	0.4235	-0.0281	0.0275	0.4390	0.4360	0.1458	0.1855	0.2018	0.0438	0.3273	0.1777
eshop's users	-0.2290	0.4738	-0.1999	0.2679	0.0288	-0.1401	0.3554	0.5963	0.0506	0.5144	0.4235	1.0000	-0.2118	-0.1384	0.5526	0.5982	0.2282	0.2739	0.3463	0.1845	0.4020	0.3087
churn	0.0379	-0.3287	0.6920	-0.2875	0.0203	-0.1920	-0.1775	-0.2863	-0.0731	-0.1154	-0.0281	-0.2118	1.0000	0.9329	-0.0376	-0.1295	-0.1401	-0.2898	-0.6022	-0.6857	-0.3627	-0.3785
churn times	0.0207	-0.2579	0.6107	-0.2648	0.0327	-0.1843	-0.1446	-0.2435	-0.0676	-0.0657	0.0275	-0.1384	0.9329	1.0000	0.0120	-0.0753	-0.1306	-0.2678	-0.5192	-0.6010	-0.2999	-0.3253
missing parts	-0.2492	0.4149	0.0123	0.2108	-0.0427	-0.1676	0.2646	0.4069	0.0526	0.4826	0.4390	0.5526	-0.0376	0.0120	1.0000	0.8184	0.2913	0.2167	0.2008	-0.0161	0.3431	0.2353
times of missing parts	-0.2229	0.4969	-0.1260	0.3029	-0.0575	-0.1154	0.3287	0.4469	0.0929	0.4366	0.4360	0.5982	-0.1295	-0.0753	0.8184	1.0000	0.4708	0.3164	0.3487	0.1154	0.4587	0.3508
total number of missing parts	-0.0581	0.2759	-0.1388	0.3876	0.0014	-0.0386	0.2579	0.2147	0.1201	0.1690	0.1458	0.2282	-0.1401	-0.1306	0.2913	0.4708	1.0000	0.4047	0.2856	0.1199	0.3515	0.2771
Frequency	-0.0346	0.4679	-0.2525	0.9314	0.0284	0.1190	0.3965	0.3260	0.2259	0.2130	0.1855	0.2739	-0.2898	-0.2678	0.2167	0.3164	0.4047	1.0000	0.5315	0.2208	0.6471	0.5244
RFM score	-0.0705	0.7225	-0.7833	0.5384	-0.0437	0.2433	0.3408	0.4299	0.2451	0.2326	0.2018	0.3463	-0.6022	-0.5192	0.2008	0.3487	0.2856	0.5315	1.0000	0.7677	0.8477	0.8732
R-score	-0.0668	0.3745	-0.9797	0.2182	-0.0464	0.2011	0.1204	0.2555	0.0456	0.0537	0.0438	0.1845	-0.6857	-0.6010	-0.0161	0.1154	0.1199	0.2208	0.7677	1.0000	0.3710	0.4067
F-score	-0.0539	0.7820	-0.4114	0.6403	0.0063	0.2083	0.4128	0.4713	0.2810	0.3701	0.3273	0.4020	-0.3627	-0.2999	0.3431	0.4587	0.3515	0.6471	0.8477	0.3710	1.0000	0.8412
M-score	-0.0511	0.7061	-0.4365	0.5512	-0.0598	0.1930	0.3643	0.3747	0.3264	0.2071	0.1777	0.3087	-0.3785	-0.3253	0.2353	0.3508	0.2771	0.5244	0.8732	0.4067	0.8412	1.0000

Εικόνα 21: Πίνακας συντελεστή συσχέτισης Pearson ανά ζεύγος μεταβλητών.

Τα ίδια συμπεράσματα μπορούν να εξαχθούν μελετώντας τα διαγράμματα διασποράς και το heatmap των μεταβλητών της εικόνας 22. Ειδικότερα στο heatmap φαίνεται με έντονο κόκκινο χρώμα η ισχυρή θετική συσχέτιση όπως για παράδειγμα Total value με Frequency ή το RFM-score με το R-score, F-score και M-score, ενώ με έντονο μπλε χρώμα φαίνονται οι ισχυρές αρνητικές συσχετίσεις όπως για παράδειγμα date of last transaction με RFM-score και R-score. Στην εικόνα έχουν μαρκαριστεί με κίτρινο κύκλο οι περιπτώσεις που αναφέρθηκαν παραπάνω.



Εικόνα 22: Διαγράμματα διασποράς και heatmap συσχετίσεων μεταξύ των μεταβλητών.

Για τον συντελεστή συσχέτισης Spearman στον πίνακα 19 φαίνονται ενδεικτικά οι συσχετίσεις μεταξύ των μεταβλητών που αναλύθηκαν παραπάνω όπου επίσης φαίνονται οι αντίστοιχοι συντελεστές Pearson. Όπως προκύπτει δεν υπάρχουν ακραίες τιμές που να επηρεάζουν την τιμή του συντελεστή Pearson καθώς για όλα τα ζεύγη του πίνακα οι τιμές είναι παρόμοιες με το συντελεστή Spearman.

Πίνακας 19: Σύγκριση συντελεστών Pearson και Spearman για τα ζεύγη μεταβλητών με τις υψηλότερες συσχετίσεις.

Ζεύγος μεταβλητών	**Συντελεστής Pearson**	**Συντελεστής Spearman**
Products types per customer – RFM-score	0,7225	0,6886
Products types per customer – F-score	0,7820	0,7722
Products types per customer – M-score	0,7061	0,6699
Date of last transaction – RFM-score	-0,7833	0,8137
Date of last transaction – R-score	-0,9797	-0,9582
Total value – Frequency	0,9314	0,8688
Churn – churn times	0,9329	0,9649
missing parts – times of missing parts	0,8184	0,9795
RFM-score – R-score	0,7677	0,8061
RFM-score – F-score	0,8477	0,8056
RFM-score – M-score	0,8732	0,8478

Μετά την αφαίρεση των μεταβλητών churn times, times of missing parts, R-score, F-score και M-score, οι συνολικές μεταβλητές που μπορούν να χρησιμοποιηθούν στους αλγόριθμους πρόβλεψης της απώλειας πελατών είναι 16. Τα δεδομένα για τους 1.716 πελάτες εισήχθησαν στο πρόγραμμα JMP για να δοκιμαστούν τρεις αλγόριθμοι ταξινόμησης: Decision Trees, Neural Networks και Logistic Regression. Στη συνέχεια ακολουθεί η αξιολόγηση των αποτελεσμάτων και η σύγκριση των τριών μεθόδων για την εξαρτημένη μεταβλητή churn.

6.2. Εφαρμογή Μοντέλων Πρόβλεψης

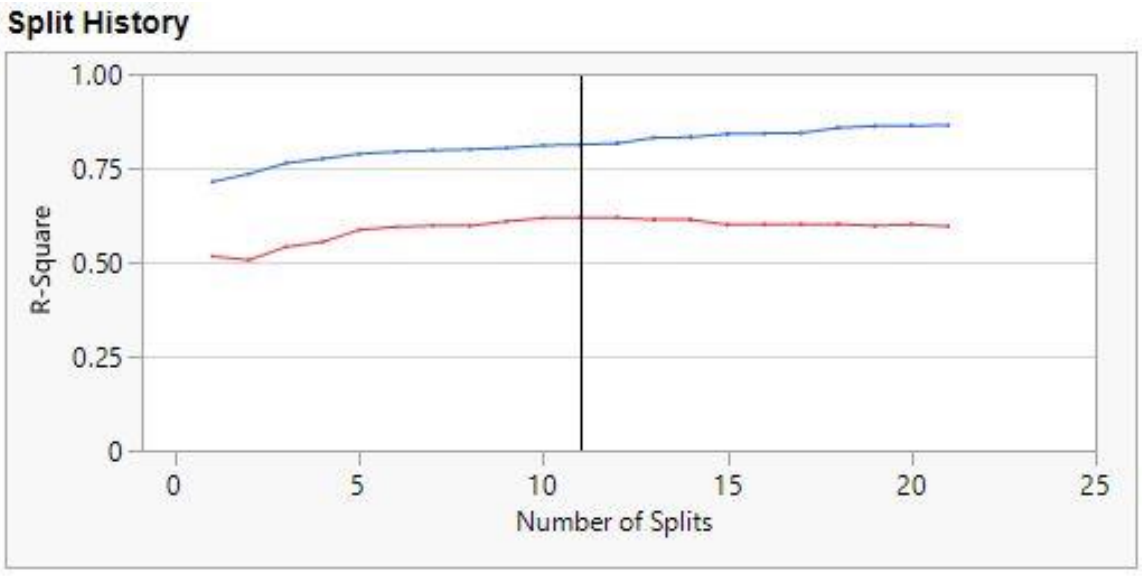
Η Trial έκδοση του προγράμματος JMP 17.2.0 έχει περιορισμό στη χρήση όλων των διαθέσιμων αλγορίθμων και για το λόγο αυτό έγιναν δοκιμές μόνο για τους διαθέσιμους αλγόριθμους Decision Trees, Neural Networks και Logistic Regression.

6.2.1. Μοντέλο Δέντρου Αποφάσεων (Decision tree)

Για τη δόμηση του μοντέλου πρόβλεψης επιλέχτηκε η μεταβλητή churn ως εξαρτημένη μεταβλητή και οι παραπάνω 16 μεταβλητές ως ανεξάρτητες μεταβλητές. Έγιναν διάφορες δοκιμές για το διαχωρισμό των δεδομένων σε δεδομένα εκπαίδευσης (training dataset) και δεδομένα επαλήθευσης (validation dataset) και τα καλύτερα αποτελέσματα προέκυψαν για διαχωρισμό 75% training και 25% validation.

Πιο συγκεκριμένα, για τον αλγόριθμο Decision Tree, ο δείκτης αξιολόγησης του μοντέλου R2 (Entropy R2) είναι 81,27% για τα δεδομένα εκπαίδευσης και 61,87% για τα δεδομένα επαλήθευσης, αποτελέσματα που προέκυψαν μετά από 11 split tests. Ο αλγόριθμος μετά το 11ο split σταματάει καθώς αν και συνεχίζει να βελτιώνει το R^2 για τα δεδομένα εκπαίδευσης, δε συμβαίνει το ίδιο και με τα δεδομένα επαλήθευσης όπως φαίνεται και στην εικόνα 23. Το ποσοστό εσφαλμένης ταξινόμησης (misclassification rate) είναι 5,85% για τα δεδομένα εκπαίδευσης και 8,52% για τα δεδομένα επαλήθευσης. Οι δείκτες προσαρμογής του μοντέλου φαίνονται στην εικόνα 24.

Εικόνα 23: Τιμή R^2 για κάθε split του μοντέλου Decision Tree.



Εικόνα 24: Δείκτες προσαρμογής του μοντέλου Decision Tree.

Fit Details

Measure	Training	Validation
Entropy RSquare	0.8127	0.6187
Generalized RSquare	0.8958	0.7546
Mean -Log p	0.1202	0.2379
RASE	0.1966	0.2733
Mean Abs Dev	0.0799	0.1737
Misclassification Rate	0.0583	0.0852
N	1270	446

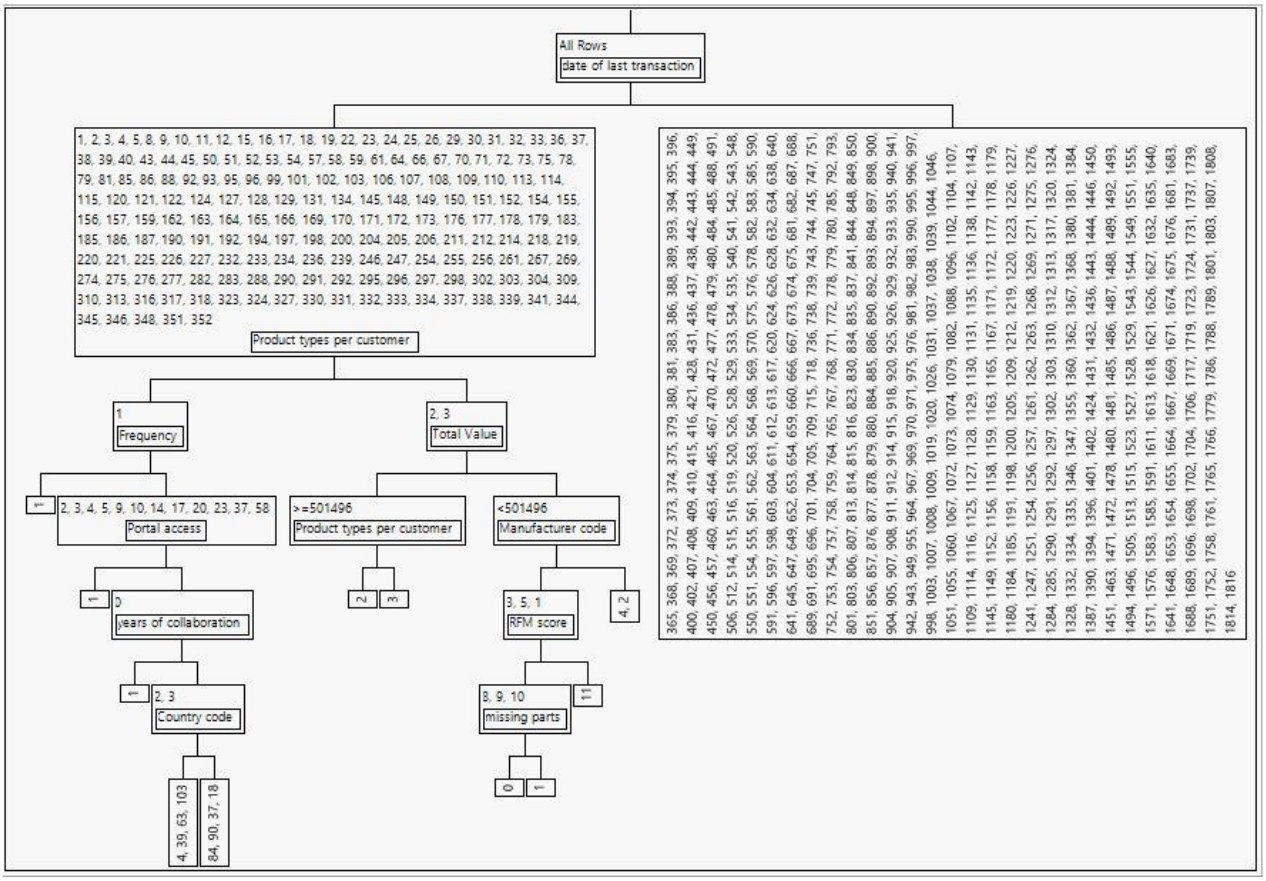
Ο αλγόριθμος προχωράει τη διαδικασία των διαχωρισμών (splits) επιλέγοντας την ανεξάρτητη μεταβλητή που μεγιστοποιεί την τιμή G2, έναν στατιστικό δείκτη που αξιολογεί τη σημαντικότητα μεταξύ παρατηρούμενης και αναμενόμενης συχνότητας της εξαρτημένης μεταβλητής. Όσο μεγαλύτερο το G2 τόσο μεγαλύτερη βαρύτητα έχει η ανεξάρτητη μεταβλητή στην επεξήγηση της εξαρτημένης. Όπως φαίνεται και στην εικόνα 25 από τα αποτελέσματα του αλγορίθμου, η μεταβλητή date of last transaction έχει το μεγαλύτερο G2 και συμμετέχει κατά 87,59% στην εξήγηση της μεταβλητής churn, δηλαδή το αν ένας πελάτης θα αποχωρήσει από την επιχείρηση. Το ποσοστό είναι μεγάλο καθώς εξ ορισμού το churn συνδέεται με το πόσες μέρες έχει να πραγματοποιήσει παραγγελία ο πελάτης και στα δικά μας δεδομένα όπως έχει αναφερθεί ορίζεται ως churn ο ένας χρόνος. Ακολουθούν σε σημαντικότητα οι μεταβλητές Total Value, product types per customer και Frequency, δηλαδή στις καλύτερες τέσσερις μεταβλητές ανήκουν ουσιαστικά το Recency (date of last transaction), Frequency και Monetary (Total value) όπως σχολιάστηκε και στο κεφάλαιο 6. Στην εικόνα 26 φαίνεται συνοπτική απεικόνιση του Δέντρου Αποφάσεων.

Εικόνα 25: Συμμετοχή των ανεξάρτητων μεταβλητών στην επεξήγηση της εξαρτημένης μεταβλητής churn.

Column Contributions

Term	Number of Splits	G^2		Portion
date of last transaction	1	1163.74735		0.8759
Total Value	1	46.3213346		0.0349
Product types per customer	2	38.4255762		0.0289
Frequency	1	22.2015251		0.0167
Manufacturer code	1	19.472239		0.0147
Country code	1	11.6985596		0.0088
RFM score	1	8.92059767		0.0067
years of collaboration	1	6.45730692		0.0049
missing parts	1	5.92589729		0.0045
Portal access	1	5.45392579		0.0041
SalesPerson code	0	0		0.0000
Company size	0	0		0.0000
hasCreditLimit	0	0		0.0000
Installer	0	0		0.0000
eshop's users	0	0		0.0000
total number of missing parts	0	0		0.0000

Εικόνα 26: Συνοπτική απεικόνιση μοντέλου Decision Tree.



Στην εικόνα 27 φαίνονται οι πίνακες σύγχυσης (confusion matrices) για τα δεδομένα εκπαίδευσης και επαλήθευσης και οι αντίστοιχες επαληθεύσεις σε απόλυτο αριθμό και ποσοστιαία. Για τα δεδομένα επαλήθευσης που έχουν μεγαλύτερη αξία, προκύπτει ότι ο αλγόριθμος ταξινομεί σωστά τους πελάτες που όντως αποχωρούν (TP) σε ποσοστό 94,8% και τους πελάτες που δεν αποχωρούν σε ποσοστό 84,4% (TN).

Εικόνα 27: Πίνακας σύγχυσης του μοντέλου Decision Tree.

Confusion Matrix

Training

Actual	Predicted Count	
churn	1	0
1	790	47
0	27	406

Actual	Predicted Rate	
churn	1	0
1	0.944	0.056
0	0.062	0.938

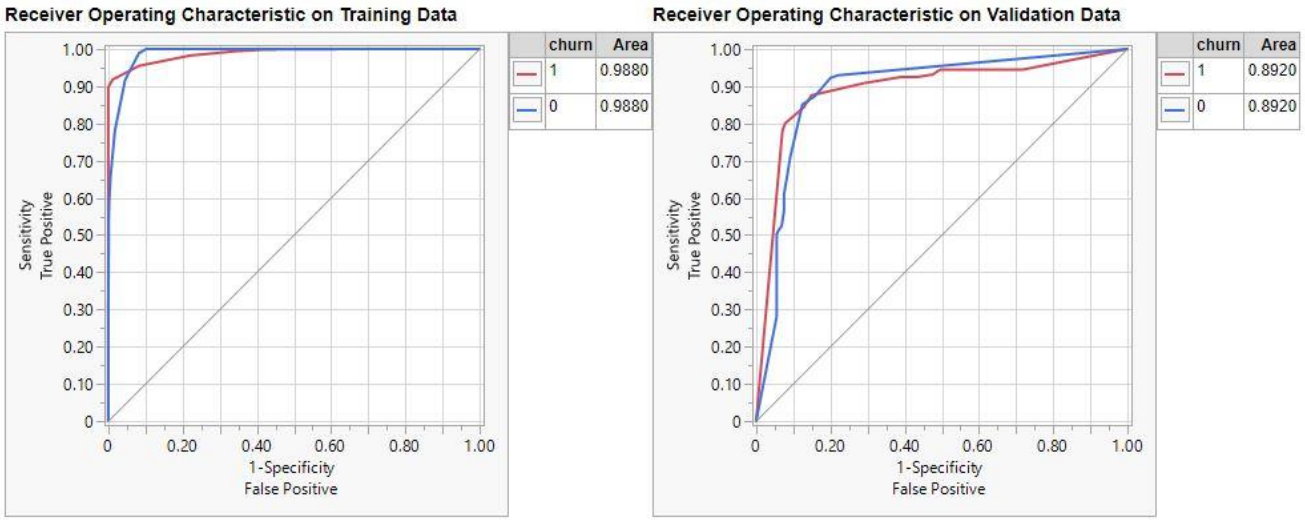
Validation

Actual	Predicted Count	
churn	1	0
1	289	16
0	22	119

Actual	Predicted Rate	
churn	1	0
1	0.948	0.052
0	0.156	0.844

Η εικόνα 28 δείχνει τις καμπύλες ROC (Receiver Operating Characteristics) για τα δεδομένα εκπαίδευσης και επαλήθευσης. Το ιδανικό σημείο στην καμπύλη είναι στην πάνω αριστερή γωνία όπου είναι υψηλό το sensitivity ενώ ταυτόχρονα χαμηλά το specificity. Για τα δεδομένα εκπαίδευσης το AUC (Area Under Curve) είναι 0,988 και για τα δεδομένα επαλήθευσης 0,892.

Εικόνα 28: Καμπύλες ROC του μοντέλου Decision Tree.



Οι μετρικές αξιολόγησης Sensitivity, Specificity, Precision, Accuracy, F1 score και MCC τόσο για τα δεδομένα εκπαίδευσης όσο και επαλήθευσης φαίνονται στην εικόνα 29 και οι τιμές τους αποδεικνύουν ότι το μοντέλο έχει πολύ καλή προβλεπτική ικανότητα.

Εικόνα 29: Σύνοψη μετρικών αξιολόγησης μοντέλου Decision Tree.

Metrics Training

Method	TP	FN	FP	TN	Sensitivity	Specificity	Precision	Accuracy	F1	MCC
Partition	406	27	47	790	0.9376	0.9438	0.8962	0.9417	0.9165	0.8723

Metrics Validate

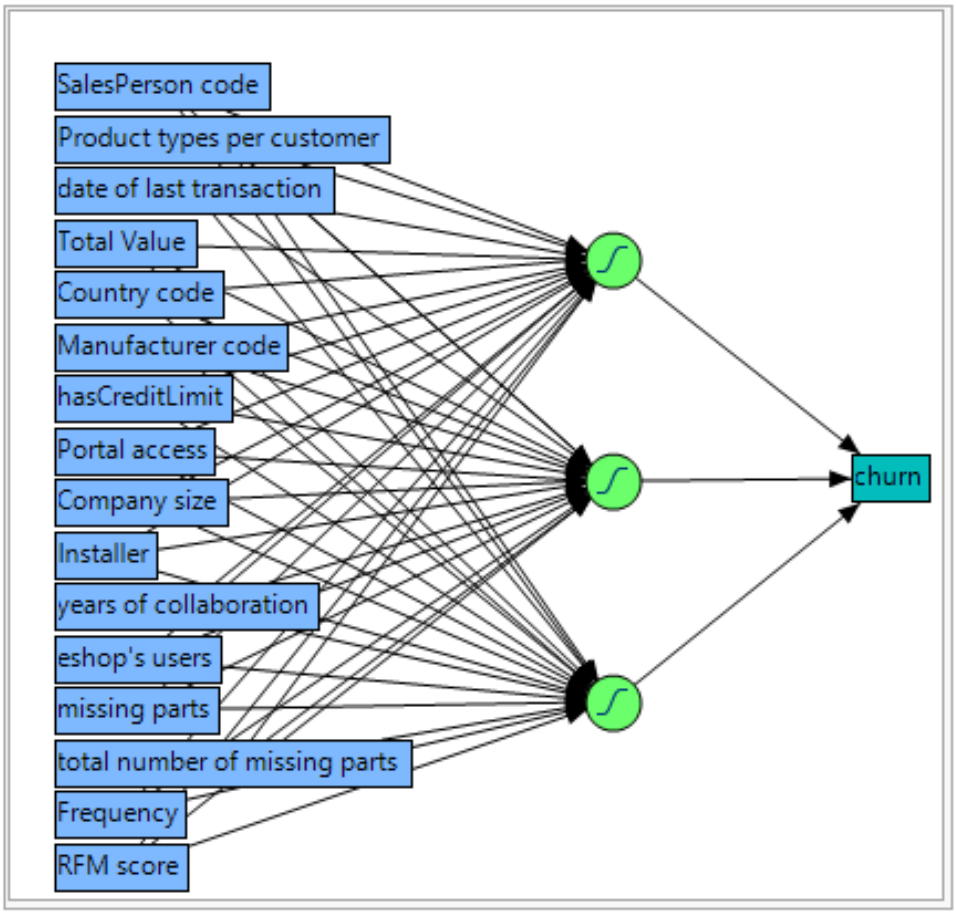
Method	TP	FN	FP	TN	Sensitivity	Specificity	Precision	Accuracy	F1	MCC
Partition	119	22	16	289	0.8440	0.9475	0.8815	0.9148	0.8623	0.8011

6.2.2. Μοντέλο Νευρωνικού Δικτύου (Neural Network)

Για το μοντέλο Νευρωνικού Δικτύου επιλέχτηκε 1 κρυφό επίπεδο με 3 κόμβους και μέθοδος επαλήθευσης η Holdback με ποσοστό 30%. Αυτό σημαίνει ότι το Νευρωνικό Δίκτυο χρησιμοποίησε 70% των δεδομένων για την εκπαίδευση του μοντέλου και το 30% των δεδομένων το κράτησε για επαλήθευση. Ο διαχωρισμός 70-30 έγινε μετά από δοκιμές που έδωσαν τα καλύτερα αποτελέσματα ενώ αξίζει να σημειωθεί ότι τα αποτελέσματα στα Νευρωνικά Δίκτυα δεν είναι ποτέ ακριβώς τα ίδια καθώς σε κάθε δοκιμή η επιλογή των αρχικών τιμών βαρύτητας αλλά και η επιλογή των δεδομένων εκπαίδευσης είναι τυχαία. Η δομή του Νευρωνικού Δικτύου με τις 16 ανεξάρτητες μεταβλητές, την εξαρτημένη μεταβλητή churn και τους 3 κόμβους φαίνεται στην εικόνα 30.

Εικόνα 30: Διαγραμματική απεικόνιση του μοντέλου Neural Network.

Diagram



Ο δείκτης αξιολόγησης του μοντέλου R^2 (Entropy R^2) για τα δεδομένα εκπαίδευσης είναι 88,04% και για τα δεδομένα επαλήθευσης είναι 52,81% ενώ το ποσοστό εσφαλμένης ταξινόμησης είναι 0,83% και 13,18% αντίστοιχα. Οι δείκτες προσαρμογής του μοντέλου φαίνονται στην εικόνα 31.

Εικόνα 31: Δείκτες προσαρμογής του μοντέλου Neural Network.

Training		Validation	
churn		churn	
Measures	**Value**	**Measures**	**Value**
Generalized RSquare	0.9360849	Generalized RSquare	0.6800612
Entropy RSquare	0.8803906	Entropy RSquare	0.528076
RASE	0.1037649	RASE	0.300918
Mean Abs Dev	0.0594881	Mean Abs Dev	0.187725
Misclassification Rate	0.0083333	Misclassification Rate	0.1317829
-LogLikelihood	91.442369	-LogLikelihood	155.32439
Sum Freq	1200	Sum Freq	516

Το Νευρωνικό Δίκτυο υπολογίζει για κάθε τιμή της κάθε μεταβλητής τη συμμετοχή τους στην περιγραφή της ανεξάρτητης μεταβλητής churn. Οι μεταβλητές ταξινομήθηκαν με φθίνουσα σειρά σημαντικότητας και προέκυψε ότι οι μεταβλητές με τη μεγαλύτερη βαρύτητα είναι η date of last transaction, Frequency, total number of missing parts, SalesPerson και Country.

Καθώς ο αλγόριθμος λαμβάνει υπόψιν τη βαρύτητα κάθε τιμής της κάθε μεταβλητής, δημιουργούνται 3876 υπομεταβλητές που επηρεάζουν είτε θετικά είτε αρνητικά το churn. Μια ταξινόμηση 3876 μεταβλητών σε φθίνουσα σειρά θα ήταν πολύ δύσκολο να οπτικοποιηθεί και για τον λόγο αυτόν αθροίστηκαν οι απόλυτες τιμές τους ώστε να γίνει και πάλι η σύνθεση των αρχικών 16 μεταβλητών. Η βαρύτητα των μεταβλητών στην επεξήγηση της εξαρτημένης μεταβλητής churn φαίνονται στον πίνακα 20 με φθίνουσα σειρά σημαντικότητας.

Όπως και στο μοντέλο Decision Tree έτσι και στο Neural Network η μεταβλητή με τη μεγαλύτερη σημαντικότητα είναι η date of last transaction και ακολουθείται από τις μεταβλητές SalesPerson, Frequency, total number of missing parts, Total Value και RFM score.

Πίνακας 20: Βαρύτητα των ανεξάρτητων μεταβλητών του μοντέλου Neural Network στην επεξήγηση της εξαρτημένης μεταβλητής churn.

Parameter	Estimate
date of last transaction	0,162
SalesPerson	0,097
Frequency	0,095
total number of missing parts	0,090
total value	0,090
RFM score	0,081
Country code	0,079
product types per customer	0,050
Manufacturer code	0,035
eshop's user	0,024
Portal access	0,018

Parameter	Estimate
hasCreditLimit	0,012
years of collaboration	0,009
missing parts	0,008
Company size	0,008
Installer	0,004

Στην εικόνα 32 φαίνονται οι πίνακες σύγχυσης για τα δεδομένα εκπαίδευσης και επαλήθευσης και οι αντίστοιχες επαληθεύσεις σε απόλυτο αριθμό και ποσοστιαία. Για τα δεδομένα επαλήθευσης προκύπτει ότι ο αλγόριθμος Neural Network ταξινομεί σωστά τους πελάτες που αποχωρούν (TP) σε ποσοστό 88,9% και τους πελάτες που δεν αποχωρούν σε ποσοστό 82,7% (TN).

Εικόνα 32: Πίνακας σύγχυσης του μοντέλου Neural Network.

Training

Confusion Matrix

Actual	Predicted Count	
churn	1	0
1	792	7
0	3	398

Confusion Rates

Actual	Predicted Rate	
churn	1	0
1	0.991	0.009
0	0.007	0.993

Validation

Confusion Matrix

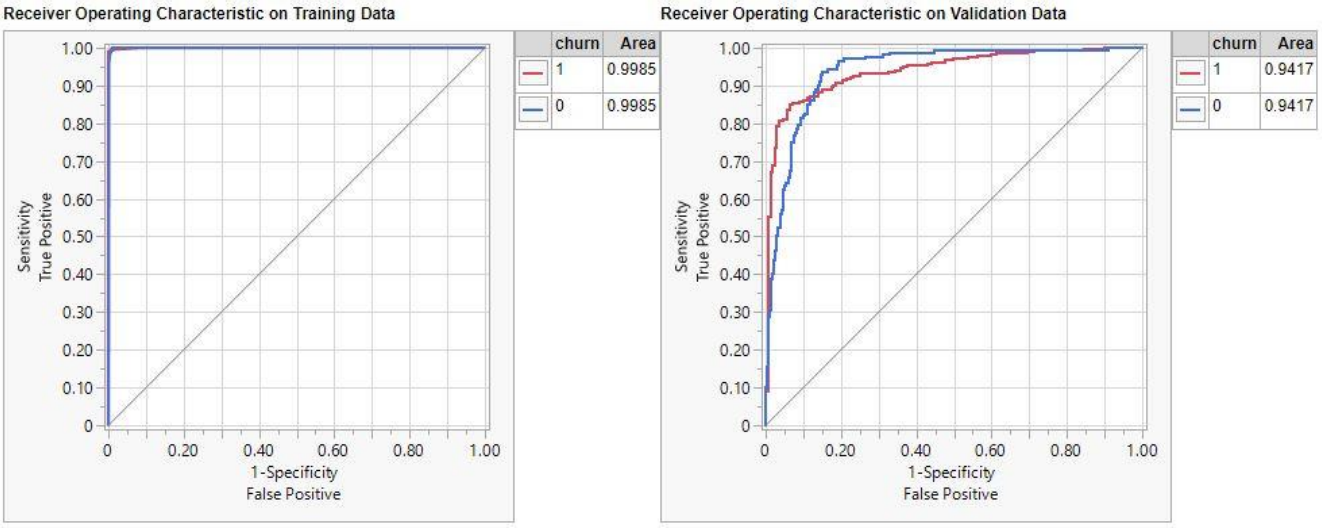
Actual	Predicted Count	
churn	1	0
1	305	38
0	30	143

Confusion Rates

Actual	Predicted Rate	
churn	1	0
1	0.889	0.111
0	0.173	0.827

Στην εικόνα 33 φαίνονται οι καμπύλες ROC για τα δεδομένα εκπαίδευσης και επαλήθευσης και οι αντίστοιχες επιτεύξεις AUC. Για τα δεδομένα εκπαίδευσης το AUC είναι πολύ υψηλό, πρακτικά 1 (0,999) και για τα δεδομένα επαλήθευσης 0,942.

Εικόνα 33: Καμπύλες ROC του μοντέλου Neural Network.



Οι μετρικές αξιολόγησης Sensitivity, Specificity, Precision, Accuracy, F1 score και MCC τόσο για τα δεδομένα εκπαίδευσης όσο και επαλήθευσης φαίνονται στην εικόνα 34 και οι τιμές τους αποδεικνύουν ότι το μοντέλο έχει αρκετά καλή προβλεπτική ικανότητα.

Εικόνα 34: Σύνοψη μετρικών αξιολόγησης μοντέλου Neural Network.

Metrics Training

Method	TP	FN	FP	TN	Sensitivity	Specificity	Precision	Accuracy	F1	MCC
Neural	398	3	7	792	0.9925	0.9912	0.9827	0.9917	0.9876	0.9813

Metrics Validate

Method	TP	FN	FP	TN	Sensitivity	Specificity	Precision	Accuracy	F1	MCC
Neural	143	30	38	305	0.8266	0.8892	0.7901	0.8682	0.8079	0.7081

6.2.3. Μοντέλο Λογιστικής Παλινδρόμησης (Logistic Regression)

Για το μοντέλο της Λογιστικής Παλινδρόμησης και την επιλογή των μεταβλητών που είναι σημαντικές (feature importance) για την πρόβλεψη της εξαρτημένης μεταβλητής churn, πραγματοποιήθηκε πρώτα η μέθοδος Stepwise Regression. Η διαδικασία αυτή εκτελεί μια ανάλυση παλινδρόμησης με όλες τις διαθέσιμες μεταβλητές και στη συνέχεια εκτελεί την ανάλυση προσθέτοντας κάθε φορά μια μεταβλητή (forward) υπολογίζοντας τα κριτήρια σημαντικότητας. Αν η προσθήκη της μεταβλητής αυξάνει την προβλεπτική ισχύ του μοντέλου, η μεταβλητή διατηρείται αλλιώς αφαιρείται. Όταν πλέον δεν υπάρχει

περεταίρω βελτίωση, η διαδικασία σταματά και προτείνεται την τελική λίστα προτεινόμενων μεταβλητών.

Η διαδικασία μπορεί να χωρίσει τις τιμές μια μεταβλητής σε κλάσεις που περιγράφουν καλύτερα τη βαρύτητα τους στην επεξήγηση του churn ενώ υπάρχει πιθανότητα διαφορετικές κλάσεις της ίδιας μεταβλητής να έχουν θετική ή αρνητική βαρύτητα. Η διαδικασία τελικά πρότεινε 10 μεταβλητές που είναι υποκατηγορίες (κλάσεις) των αρχικών μεταβλητών date of last transaction, Frequency, Country code και product types per customer. Αξίζει να σημειωθεί ότι η διαδικασία Stepwise Regression είναι χρονοβόρα και απαιτεί μεγάλη υπολογιστική ισχύ καθώς συνυπολογίζει ταυτόχρονα πολλές μεταβλητές με πολλές επαναληπτικές διαδικασίες.

Στην εικόνα 35 φαίνονται οι δείκτες προσαρμογής του μοντέλου Logistic Regression και πιο συγκεκριμένα το Entropy R^2 για τα δεδομένα εκπαίδευσης είναι 81,11% και για τα δεδομένα επαλήθευσης 78,64% ενώ το ποσοστό εσφαλμένης ταξινόμησης είναι 5,91% και 6,99% αντίστοιχα.

Εικόνα 35: Δείκτες προσαρμογής του μοντέλου Logistic Regression.

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0,8111	0,7864	1-Loglike(model)/Loglike(0)
Generalized RSquare	0,8944	0,8787	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0,1203	0,1363	∑ -Log(ρ[j])/n
RASE	0,1981	0,2152	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0,0783	0,0895	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0,0591	0,0699	∑ (ρ[j]≠ρMax)/n
N	1287	429	n

Η βαρύτητα των μεταβλητών φαίνεται στον πίνακα 21 και όπως αναφέρθηκε παραπάνω, το μοντέλο της Λογιστικής Παλινδρόμησης μπορεί να χωρίσει τις αρχικές μεταβλητές σε κλάσεις ανάλογα με τις τιμές που λαμβάνουν για τη βέλτιστη επεξήγηση του churn. Έτσι για παράδειγμα η μεταβλητή Frequency 1-3 επηρεάζει θετικά το churn με βαρύτητα 11,54, η μεταβλητή product types per customer 1-2 επηρεάζει θετικά με βαρύτητα 0,37 και η μεταβλητή product types per customer1-3 επηρεάζει αρνητικά με βαρύτητα 0,33. Η βαρύτητα των 10 μεταβλητών του μοντέλου φαίνεται στον πίνακα 21. Ομοίως με τα μοντέλα Decision Tree και Neural Network η μεταβλητή με τη μεγαλύτερη βαρύτητα

είναι η date of last transaction ενώ ακολουθούν Frequency, Country Code και Product types per customer.

Πίνακας 21: Βαρύτητα των ανεξάρτητων μεταβλητών του μοντέλου Logistic Regression στην επεξήγηση της εξαρτημένης μεταβλητής churn.

Parameter	Value	Estimate
Product types per customer	1-3	-0,325
Product types per customer	1-2	0,374
date of last transaction	1-1816	22,717
Country Code	var1	0,615
Country Code	var2	0,411
Country Code	var3	2,108
Frequency	1-15	1,124
Frequency	1-7	3,465
Frequency	1-4	5,772
Frequency	1-3	11,541

Η εικόνα 36 δείχνει τον πίνακα σύγχυσης του μοντέλου Logistic Regression για τα δεδομένα εκπαίδευσης και επαλήθευσης σε απόλυτο αριθμό και ποσοστιαία. Στα δεδομένα επαλήθευσης, το ποσοστό ορθής πρόβλεψης ταξινομημένων παρατηρήσεων ως churn είναι 91,6% (TP) και ορθής πρόβλεψης διατήρησης (TN) 95,8%.

Εικόνα 36: Πίνακας σύγχυσης του μοντέλου Logistic Regression.

Confusion Matrix

Training

Actual	Predicted Count	
churn	1	0
1	800	57
0	19	411

Validation

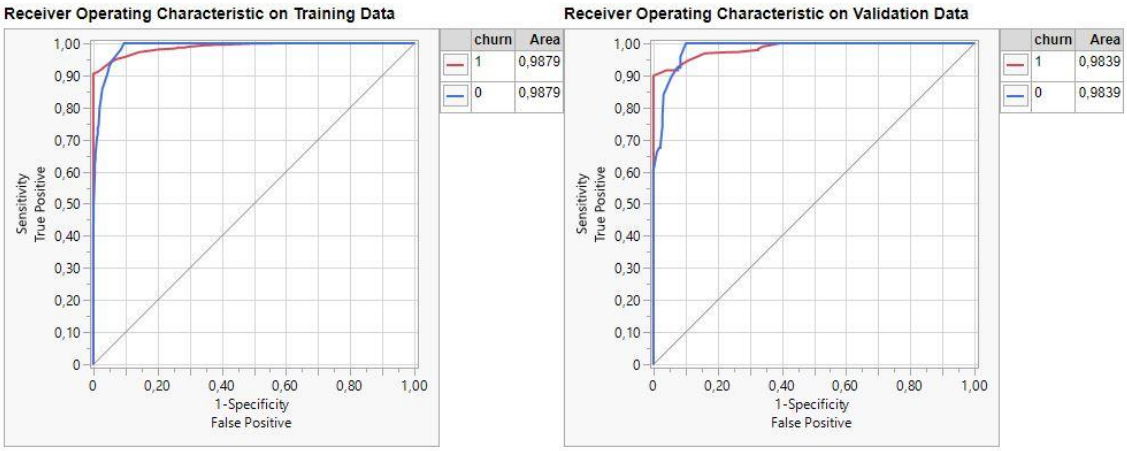
Actual	Predicted Count	
churn	1	0
1	261	24
0	6	138

Actual	Predicted Rate	
churn	1	0
1	0,933	0,067
0	0,044	0,956

Actual	Predicted Rate	
churn	1	0
1	0,916	0,084
0	0,042	0,958

Η εικόνα 37 δείχνει την καμπύλη ROC για τα δεδομένα εκπαίδευσης και επαλήθευσης με AUC 0.988 και 0,984 αντίστοιχα ενώ οι μετρικής αξιολόγησης του μοντέλου φαίνονται στην εικόνα 38. Το μοντέλο έχει πολύ καλή προσαρμογή τόσο στα δεδομένα εκπαίδευσης όσο και στα δεδομένα επαλήθευσης αποδεικνύοντας ότι μπορεί να γενικεύσει το μοτίβο σε νέα δεδομένα.

Εικόνα 37: Καμπύλη ROC του μοντέλου Logistic Regression.



Εικόνα 38: Σύνοψη μετρικών αξιολόγησης μοντέλου Logistic Regression για τα δεδομένα εκπαίδευσης.

Metrics Training

Method	TP	FN	FP	TN	Sensitivity	Specificity	Precision	Accuracy	F1	MCC
Fit Nominal Logistic	800	57	19	411	0,9335	0,9558	0,9768	0,9409	0,9547	0,872

Metrics Validation

Method	TP	FN	FP	TN	Sensitivity	Specificity	Precision	Accuracy	F1	MCC
Fit Nominal Logistic	261	24	6	138	0,9158	0,9583	0,9775	0,9301	0,9457	0,8515

7. Συμπεράσματα και προτάσεις

7.1. Σύγκριση αποτελεσμάτων

Για τη σύγκριση των αποτελεσμάτων δημιουργήθηκε ο πίνακας 22 που συνοψίζει τις μετρικές αξιολόγησης για τα δεδομένα εκπαίδευσης και επαλήθευσης των τριών μοντέλων.

Πίνακας 22: Συγκεντρωτικός πίνακας μετρικών αξιολόγησης.

Training

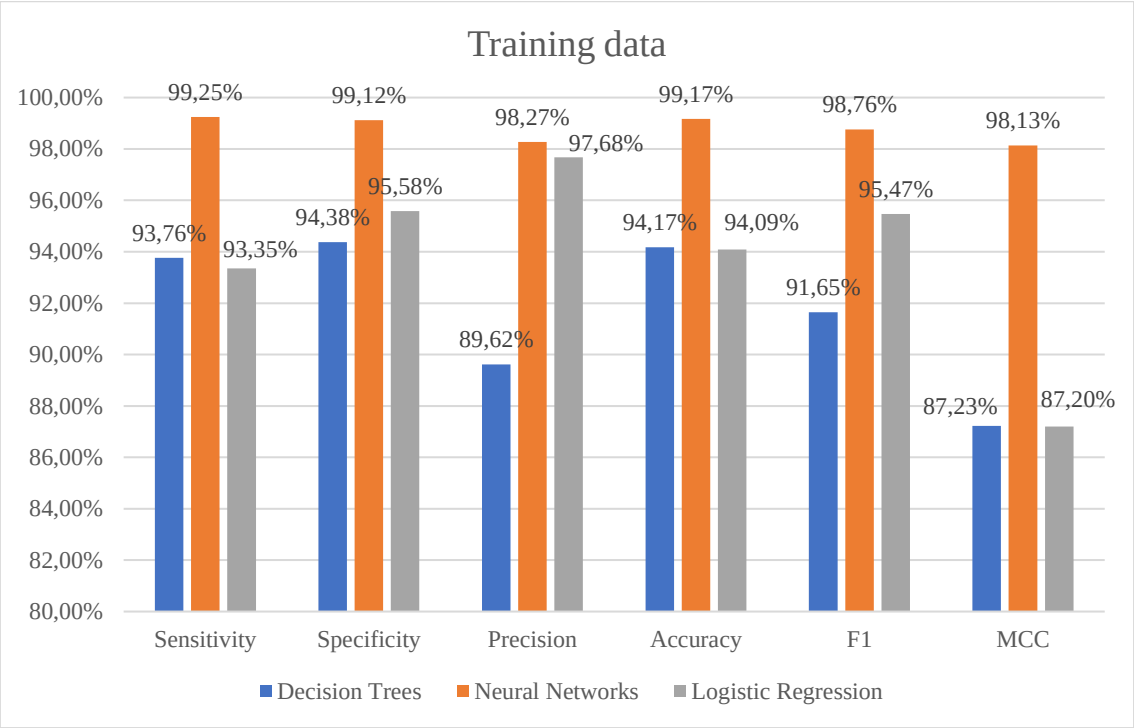
Model	Sensitivity	Specificity	Precision	Accuracy	F1	MCC
Decision Trees	93,76%	94,38%	89,62%	94,17%	91,65%	87,23%
Neural Networks	99,25%	99,12%	98,27%	99,17%	98,76%	98,13%
Logistic Regression	93,35%	95,58%	97,68%	94,09%	95,47%	87,20%

Validation

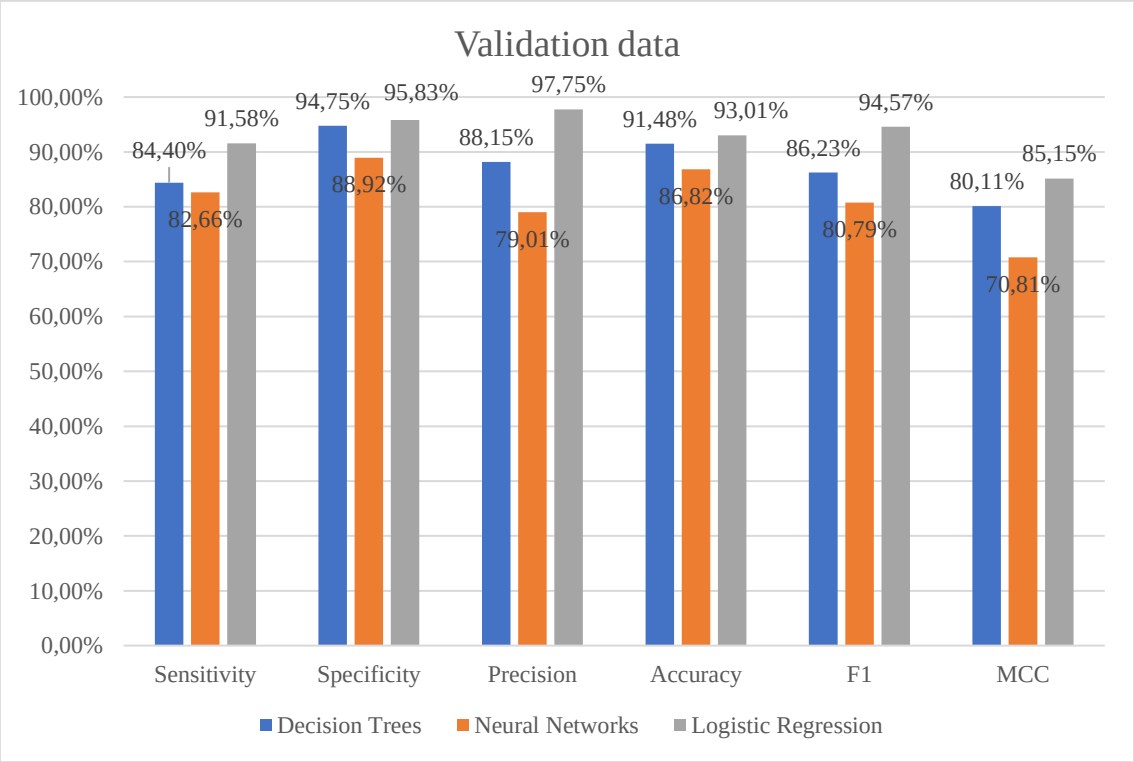
Model	Sensitivity	Specificity	Precision	Accuracy	F1	MCC
Decision Trees	84,40%	94,75%	88,15%	91,48%	86,23%	80,11%
Neural Networks	82,66%	88,92%	79,01%	86,82%	80,79%	70,81%
Logistic Regression	91,58%	95,83%	97,75%	93,01%	94,57%	85,15%

Όπως προκύπτει από τη σύγκριση των αποτελεσμάτων για τα δεδομένα εκπαίδευσης, το μοντέλο Neural Network έχει τα καλύτερα αποτελέσματα για όλες τις μετρικές αξιολόγησης, ενώ τα μοντέλα Logistic Regression και Decision Tree έχουν παρόμοια αποτελέσματα με σχεδόν ίδιο accuracy. Συγκρίνοντας ωστόσο τα αποτελέσματα για τα δεδομένα επαλήθευσης, παρατηρείται ότι το μοντέλο Logistic Regression έχει καλύτερα αποτελέσματα σε όλες τις μετρικές αξιολόγησης με δεύτερο το μοντέλο Decision Tree και τρίτο το μοντέλο Neural Networks. Αυτό σημαίνει ότι το μοντέλο Neural Network κάνει υπερπροσαρμογή στα δεδομένα εκπαίδευσης και δε μπορεί να γενικεύσει το μοτίβο σε νέα δεδομένα.

Για την καλύτερη οπτικοποίηση της σύγκρισης των αποτελεσμάτων, ακολούθως παρατίθεται ο πίνακας 22 σε δύο διαγράμματα (Διάγραμμα 26 και Διάγραμμα 27).



Διάγραμμα 26: Δεδομένα εκπαίδευσης.



Διάγραμμα 27: Δεδομένα επαλήθευσης.

Συμπερασματικά, το μοντέλο Logistic Regression έχει καλύτερη προβλεπτική ικανότητα σε σχέση με τα μοντέλα Decision Tree και Neural Network. Αν συγκριθεί το accuracy των δεδομένων επαλήθευσης των μοντέλων ως ο κύριος δείκτης που θέλει η επιχείρηση

να χρησιμοποιήσει για την επιλογή του καλύτερου μοντέλου πρόβλεψης, τότε παρατηρείται ότι το Logistic Regression και το Decision Tree είναι πολύ κοντά. Η πολύ καλή απόδοση του μοντέλου Logistic Regression δεν είναι σε απόλυτη συμφωνία με τη βιβλιογραφία που αναλύθηκε στην ενότητα 2.6 ενώ η καλή απόδοση του μοντέλου Decision Tree που παρατηρείται σε πολλές μελέτες, αποδεικνύεται και σε αυτήν. Οι περισσότερες μελέτες απέδειξαν ότι τα μοντέλα Decision Trees και Random Forests, που αποτελούν ουσιαστικά προέκταση των Decision Trees, δίνουν τα καλύτερα αποτελέσματα στη μελέτη του φαινομένου της απώλειας πελατών. Το μοντέλο Neural Network σε όσες μελέτες αναλύθηκαν δεν ήταν ποτέ το καλύτερο μοντέλο που περιγράφει καλύτερα την απώλεια πελατών όπως παρατηρείται και στην παρούσα μελέτη.

7.2. Μεταβλητές με τη μεγαλύτερη βαρύτητα

Όσον αφορά στις μεταβλητές που επηρεάζουν περισσότερο το αν ένας πελάτης θα αποχωρήσει από την επιχείρηση ή όχι, αυτό που παρατηρήθηκε από την ανάλυση είναι ότι η πιο σημαντική μεταβλητή είναι το date of last transaction σε όλα τα μοντέλα. Η μεταβλητή αυτή είναι ουσιαστικά η μεταβλητή Recency που αναλύθηκε και στην ανάλυση RFM και είναι εξ ορισμού σημαντική στην πρόβλεψη του churn καθώς ο χαρακτηρισμός ενός πελάτη ως churned ή όχι, εξαρτάται από τον αριθμό ημερών τελευταίας συναλλαγής που έχει ορίσει η μελέτη. Όσο μεγαλύτερος αριθμός ημερών από την τελευταία συναλλαγή, τόσο πιο πιθανό είναι ένας πελάτης να αποχωρήσει.

Η μεταβλητή Frequency συναντάται σε όλα τα μοντέλα και θεωρείται επίσης σημαντική. Όσο μεγαλύτερος ο αριθμός του Frequency, δηλαδή της συχνότητας παραγγελιών, τόσο μειώνεται η πιθανότητα αποχώρησης του πελάτη. Αυτό μπορεί να ερμηνευτεί ως δείκτης αφοσίωσης του πελάτη καθώς μεγάλος αριθμός παραγγελιών, σημαίνει τακτικός και επαναλαμβανόμενος πελάτης που έχει χτίσει μακροχρόνια σχέση με την επιχείρηση.

Τέλος, σε δύο από τα τρία μοντέλα, σημαντικές μεταβλητές είναι οι Total Value και Product types per customer. Η μεταβλητή Total Value είναι η συνολική αξία των παραγγελιών και μπορεί να ερμηνευτεί επίσης ως αφοσίωση πελάτη καθώς μεγάλος τζίρος σημαίνει πολλές παραγγελίες ή ακριβές (premium) παραγγελίες που και στις δύο περιπτώσεις σημαίνει ότι ο πελάτης εμπιστεύεται την επιχείρηση και άρα η πιθανότητα churn μειώνεται. Όσον αφορά στη μεταβλητή Product types per customer, φαίνεται

επίσης να υπάρχει συσχέτιση του churn με τον αριθμό των διαφορετικών προϊόντων που προμηθεύεται ο πελάτης.

Πίνακας 23: Σημαντικότερες μεταβλητές με φθίνουσα βαρύτητα ανά μοντέλο πρόβλεψης.

Decision Tree	**Neural Network**	**Logistic Regression**
date of last transaction	date of last transaction	date of last transaction
Total Value	SalesPerson	Frequency
Product types per customer	Frequency	Country code
Frequency	total number of missing parts	Product types per customer
Manufacturer code	Total Value	

7.3. Μελλοντική Έρευνα – Προτάσεις Βελτίωσης

Τα μοντέλα που αναλύθηκαν και βάσει των καλών αποτελεσμάτων τους μπορούν να αποδειχθούν ιδιαιτέρως χρήσιμα για την επιχείρηση. Αν η επιχείρηση μπορεί να προβλέπει με αρκετά καλή ακρίβεια τους πελάτες που πιθανώς θα αποχωρήσουν το προσεχές διάστημα, μπορεί να αξιολογεί τον κάθε πελάτη και να εφαρμόζει αντίστοιχα κάποια ενέργεια για τη διατήρηση τους.

Μελλοντικά, θα είχε ενδιαφέρον να συγκριθούν περισσότερα μοντέλα καθώς οι περιορισμοί στη δοκιμαστική έκδοση του προγράμματος απέκλεισε αλγόριθμους που η βιβλιογραφία έχει δείξει ότι μπορούν να δώσουν καλά αποτελέσματα όπως SVM και Random Forest. Επίσης θα μπορούσαν να προστεθούν δεδομένα δοκιμής (test) ώστε να υπάρξει μια πιο ολοκληρωμένη αξιολόγηση της προβλεπτικής απόδοσης των μοντέλων.

Μια πρόταση βελτίωσης μπορεί να αποτελέσει ο εμπλουτισμός των ανεξάρτητων μεταβλητών με στοιχεία όπως ημερομηνία τελευταίας επικοινωνίας, αριθμός παραπόνων και χρόνος επίλυσης τους, αριθμός προσφορών Portal, αριθμός παραγγελιών ανταλλακτικών μέσω e-shop, αποτελέσματα ετήσιας έρευνας ικανοποίησης πελατών, δείκτης NPS (Net Promoter Score) και άλλα. Κάποιες από τις παραπάνω μεταβλητές μπορεί να δώσουν ενδιαφέροντα αποτελέσματα ειδικά ό,τι έχει να κάνει με την ικανοποίηση πελάτη που αποτελεί μια σημαντική διάσταση της μελέτης απώλειας πελατών.

8. Βιβλιογραφία

Ahh, J.-H., Han, S.-P., & Lee, Y.-S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy, 30*, σσ. 552-568.

Aleksandrova, Y. (2018, June). Application of machine learning for churn prediction based on transactional data (RFM Analysis). *18th International Multidisciplinary Scientific GeoConference SGEM 2018, 18(2.1)*, 125-132.

Alwis, P., Kumara, B., & Hapuarachchi, H. (2018, August 25-26). Customer Churn Analysis and Prediction in Telecommunication for Decision Making. *2018 International Conference On Business Innovation (ICOBI)*, 40-45. Colombo, Sri Lanka.

Bagul, N., Surana, P., Berad, P., & Khachane, C. (2021, March). Retail Customer Churn Analysis using RFM Model and K-Means Clustering. *International Journal of Engineering Research & Technology (IJERT), 10*(3), σσ. 349-354. Ανάκτηση October 22, 2023, από https://www.researchgate.net/publication/366248713_Retail_Customer_Churn_Analysis_using_RFM_Model_and_K-Means_Clustering

Berkay. (2021, June 6). *Medium*. Ανάκτηση July 27, 2023, από https://iambideniz.medium.com/customer-segmentation-with-rfm-analysis-ed1c17aa57a6

Birant, D. (2011). Data Mining Using RFM Analysis. Στο K. Funatsu, *Knowledge-Oriented Applications in Data Mining* (σσ. 91-108). InTech.

Caigny, A., Coussement, K., Verbeke, W., Idbenjra, K., & Phan, M. (2021). Uplift modeling and its implications for B2B customer churn prediction: A segmentation-based modeling approach. *Industrial Marketing Management, 99*, σσ. 28-39.

Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., . . . Zdeborová, L. (2019, December 6). Machine learning and the physical sciences. *Reviews of Modern Physics, 91*(4), σσ. 045002_1-045002_39.

Chen, K., Hu, Y.-H., & Hsieh, Y.-C. (2014, October 2). Predicting customer churn from valuable B2B customers in the logistics industry: a case study. *Information Systems and e-Business Management, 13*(6), σσ. 475-494.

Christy, A., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University – Computer and Information Sciences, 33*, σσ. 1251-1257.

Collins-Thompson, K. (χ.χ.). *Applied Machine Learning - Unsupervised machine learning.* Ανάκτηση December 24, 2023, από Coursera: https://www.coursera.org/learn/python-machine-learning?specialization=data-science-python

Gattermann-Itschert, T., & Thonemann, U. W. (2022, October 7). Proactive customer retention management in a non-contractual B2B setting based on churn prediction with random forests. *Industrial Marketing Management, 107*, σσ. 134-147.

Gordini, N., & Veglio, V. (2017). Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry. *Industrial Marketing Management, 62*, σσ. 100-107.

Gupta, S. (2020, February 28). *Towards Data Science*. Ανάκτηση October 18, 2023, από https://towardsdatascience.com/pros-and-cons-of-various-classification-mlalgorithms-3b5bfb3c87d6

Hadden, J., Tiwari, A., Roy, R., & Ruta, D. (2007, October). Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research, 34*(10), σσ. 2902-2917.

Hills, W., Daniel, W., Lu, M., Schaer, O., & Adams, S. (2020). Modeling Client Churn for Small Business-to-Business Firms. *2020 Systems and Information Engineering Design Symposium (SIEDS)*, 1-7. Charlottesville, VA, USA.

Hughes, A. (2007). Churn reduction in the telecom industry. *Direct Marketing News*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning with Applications in R* (Second εκδ.). Springer.

Kohavi, R., & Parekh, R. (2004). Visualizing RFM Segmentation. *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM)* (σσ. 391-399). Florida: Society for Industrial and Applied Mathematics.

Kule, D., Brentari, E., & Alberici, A. (2022, November 17-18). Data-Driven Prediction-Making on customer churn in a circular economy through RFM and clustering algorithms. *Circular Economy: Opportunities and Challenges*, σσ. 132-139.

Lamb, C. W., Hair, J. F., & McDaniel, C. (2011). *Marketing* (11 εκδ.). United States of America : South-Western, Cengage Learning.

loginom. (2021, April 14). Ανάκτηση December 14, 2023, από Customer segmentation by loyalty or RFM analysis: https://loginom.com/blog/rfm

Mitchell, T. M. (1997). *Machine Learning.* McGraw-Hill.

Müller, A. C., & Guido, S. (2017). *Introduction to Machine Learning with Python.* United States of America: O'REILLY.

Mutuvi, S. (2019, April 16). *medium*. Ανάκτηση August 17, 2023, από Heartbeat: https://heartbeat.comet.ml/introduction-to-machine-learning-model-evaluation-fa859e1b2d7f

Nighania, K. (2018, December 30). *towardsdatascience*. Ανάκτηση August 18, 2023, από https://towardsdatascience.com/various-ways-to-evaluate-a-machine-learning-models-performance-230449055f15

Oghojafor, B., Mesike, G., Bakarea, R., Omoera, C., & Adeleke, I. (2012, March). Discriminant Analysis of Factors Affecting Telecoms Customer Churn. *International Journal of Business Administration, 3*(2).

Ravish, R. (χ.χ.). *enjoy algorithms.* (G. Shubham, Επιμ.) Ανάκτηση August 2023, από https://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning

Roshan, B. (2020). *Kaggle*. Ανάκτηση Οκτώβριος 15, 2023, από https://www.kaggle.com/code/benroshan/divide-and-rule-customer-segmentation-via-k-means

Sabbeh, S. F. (2018). Machine-Learning Techniques for Customer Retention: A Comparative Study. *International Journal of Advanced Computer Science and Applications (IJACSA), 9*(2).

Sarker, I. H. (2021, March 22). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science, 2*(160). doi:https://doi.org/10.1007/s42979-021-00592-x

Schwartz, H. M. (2014). *Mutli-agent machine learning - A reinforcement aprroach.* Wiley.

Sheikh, A., Ghanbarpour, T., & Gholamiangonabadi, D. (2019, April 3). A Preliminary Study of Fintech Industry: A Two-Stage Clustering Analysis for Customer Segmentation in the B2B Setting. *Journal of Business-to-Business Marketing, 26*(2), σσ. 197-207.

Silpa, S., & Chandran, A. S. (2020, March). Literature Survey On Customer Churn Prediction. *International Journal of Research and Analytical Reviews (IJRAR), 7*(1), σσ. 347-350.

Van Haver, J. (2017). Benchmarking analytical techniques for churn modelling in a B2B context. *Master's Dissertation*. Faculteit Economie en Bedrijfskunde.

Wei, J.-T., Lin, S.-Y., & Wu, H.-H. (2010, December 9). A review of the application of RFM model. *African Journal of Business Management, 4*(19), σσ. 4199-4206.

Yang. (2019, September 9). *medium.* Ανάκτηση December 24, 2023, από https://towardsdatascience.com/introduction-to-na%C3%AFve-bayes-classifier-fa59e3e24aaf

Μητρόπουλος, Μ. (2022, Οκτώβριος). Μεθοδολογία εκτίμησης πιθανότητας απώλειας πελατών. *Διπλωματική Εργασία*. Αθήνα: ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ - ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ.

Παξιμάδης, Δ. (χ.χ.). *ΚΕΜΕΛ: ΚΕΝΤΡΟ ΕΘΕΛΟΝΤΩΝ ΜΑΝΑΤΖΕΡ ΕΛΛΑΔΟΣ*. Ανάκτηση Ιούλιος 26, 2023, από kemel: https://www.kemel.gr/library/i-diacheirisi-tis-schesis-me-ton-pelati-customer-relationship-management-crm

Πρίφτης, Α. Γ. (2021, Οκτώβριος). Εξόρυξη δεδομένων στη διερεύνηση και κατηγοριοποίηση της αποχώρησης / αφοσίωσης πελατών. *Μεταπτυχιακή Διατριβή*. (Α. Γεωργίου, Επιμ.) Θεσσαλονίκη, Ελλάδα: ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ ΣΤΗΝ ΑΝΑΛΥΤΙΚΗ ΤΩΝ ΕΠΙΧΕΙΡΗΣΕΩΝ ΚΑΙ ΤΗΝ ΕΠΙΣΤΗΜΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ - ΤΜΗΜΑ ΟΡΓΑΝΩΣΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ.