Πρόγραμμα Μεταπτυχιακών Σπουδών στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Τμήμα Οργάνωσης και Διοίκησης Επιχειρήσεων

Διπλωματική Εργασία

ΕΦΑΡΜΟΓΗ RFM ΑΝΑΛΥΣΗΣ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΤΕΧΝΙΚΕΣ ΕΞΟΡΥΞΗΣ ΔΕΔΟΜΕΝΩΝ ΓΙΑ ΤΗΝ ΔΗΜΙΟΥΡΓΙΑ ΣΤΡΑΤΗΓΙΚΩΝ ΔΙΑΧΕΙΡΙΣΗΣ ΠΕΛΑΤΕΙΑΚΩΝ ΣΧΕΣΕΩΝ ΚΑΙ ΤΗΝ ΕΚΤΙΜΗΣΗ ΤΗΣ ΠΙΣΤΟΤΗΤΑΣ ΤΩΝ ΠΕΛΑΤΩΝ

ΤΗΣ

ΚΟΡΔΕΛΛΑ ΑΦΡΟΔΙΤΗΣ ΤΟΥ ΝΙΚΟΛΑΟΥ

Επιβλέπων Καθηγητής: Χατζηθωμάς Λεωνίδας

Υποβλήθηκε ως απαιτούμενο για την απόκτηση του μεταπτυχιακού διπλώματος στην Αναλυτική των Επιχειρήσεων και Επιστήμη των Δεδομένων

Θεσσαλονίκη, Φεβρουάριος 2024

Ευχαριστίες

Πρώτα απ' όλα, θέλω να ευχαριστήσω θερμά τον σύμβουλο της διπλωματικής εργασίας μου, Επίκουρο Καθηγητή Λεωνίδα Χατζηθωμά, για την ανεκτίμητη βοήθεια, την καθοδήγηση και την παρακίνηση που μου παρείχε καθ' όλη τη διάρκεια του μεταπτυχιακού μου προγράμματος. Οι εκτεταμένες γνώσεις, η εμπειρία του, και κυρίως η αμέριστη υποστήριξή του, ήταν οι κινητήριες δυνάμεις που όχι μόνο εμπλούτισαν το ακαδημαϊκό μου ταξίδι, αλλά και με παρακινούσαν καθ’ όλη την διάρκεια της συγγραφής μου. Επίσης, θα ήθελα να εκφράσω την ειλικρινή εκτίμησή μου στον Υποψήφιο Μεταδιδάκτορα της σχολής, τον κ. Παλταγιάν Γεώργιο, ο οποίος συνέβαλε στη διαδικασία εκπόνησης της διπλωματικής μου, παρέχοντας κρίσιμα άρθρα, πολύτιμες συμβουλές, προσέφερε διορατικά σχόλια και μοιράστηκε καινοτόμες ιδέες που συνέβαλαν σημαντικά στην έρευνά μου.

Τέλος, οφείλω ιδιαίτερη ευγνωμοσύνη στο οικογενειακό και φιλικό μου περιβάλλον για την αμέριστη συναισθηματική και ηθική υποστήριξή της κατά τη διάρκεια της συγγραφής αυτής της διπλωματικής εργασίας.

Abstract

The digital transformation of commerce has necessitated a deeper understanding of customer behavior, leading to enhanced customer relationship management (CRM) strategies. This research investigates the integration of Recency, Frequency, Monetary value (RFM) analysis and sentiment analysis using machine learning and natural language processing (NLP) techniques to predict Customer Lifetime Value (CLV) and understand customer loyalty in e-commerce. By analyzing transactional data through RFM and evaluating customer sentiment from online reviews, the study provides a comprehensive view of customer behavior. The research applies various machine learning models to segment customers based on RFM criteria and employs sentiment analysis to extract qualitative insights from customer feedback. The findings demonstrate that combining quantitative RFM analysis with qualitative sentiment insights significantly enhances the prediction of CLV and the formulation of targeted marketing strategies. This dual approach not only aids in identifying the most valuable customers but also in understanding the underlying factors driving customer satisfaction and loyalty. The research underscores the importance of leveraging both transactional and emotional customer data to inform more nuanced CRM and marketing tactics in the realm of e-commerce.

Keywords: RFM Analysis, Customer Lifetime Value, Sentiment Analysis, Machine Learning, Natural Language Processing, E-commerce, Customer Relationship Management, Customer Behavior, Qualitative Insights, Predictive Analytics.

Περίληψη

Η παρούσα μελέτη διερευνά την εφαρμογή της ανάλυσης RFM (Recency, Frequency, Monetary) χρησιμοποιώντας τεχνικές εξόρυξης δεδομένων για την ανάπτυξη στρατηγικών διαχείρισης πελατειακών σχέσεων και την εκτίμηση της πιστότητας των πελατών στο πλαίσιο του ηλεκτρονικού εμπορίου. Με τη ραγδαία εξέλιξη του ηλεκτρονικού εμπορίου, οι επιχειρήσεις αντιμετωπίζουν την πρόκληση όχι μόνο της προσέλκυσης αλλά και της διατήρησης μιας κερδοφόρας πελατειακής βάσης. Η παρούσα έρευνα εμβαθύνει στην τμηματοποίηση των πελατών, αξιοποιώντας την ανάλυση RFM για την αξιολόγηση των προτύπων συμπεριφοράς των πελατών στις αγορές και ενσωματώνει αλγορίθμους μηχανικής μάθησης για την πρόβλεψη της αξίας ζωής των πελατών (CLV) με μεγαλύτερη ακρίβεια.

Η μελέτη επεκτείνεται περαιτέρω στην ανάλυση συναισθήματος, χρησιμοποιώντας την επεξεργασία φυσικής γλώσσας (NLP) για τη μέτρηση της ικανοποίησης και της αφοσίωσης των πελατών με βάση τις αξιολογήσεις των συναλλαγών τους. Αναλύοντας τα σχόλια κειμένου με τη χρήση προηγμένων τεχνικών NLP, η έρευνα στοχεύει στην παροχή μιας πιο διαφοροποιημένης κατανόησης των συναισθημάτων των πελατών, η οποία, όταν συνδυάζεται με τις προβλέψεις RFM και CLV, μπορεί να προσφέρει ολοκληρωμένες γνώσεις σχετικά με τη συμπεριφορά των πελατών.

Τα βασικά ευρήματα υπογραμμίζουν την αποτελεσματικότητα του συνδυασμού της ανάλυσης RFM με μοντέλα μηχανικής μάθησης για την πρόβλεψη του CLV, αναδεικνύοντας τις δυνατότητες των επιχειρήσεων να προσαρμόσουν αποτελεσματικότερα τις στρατηγικές μάρκετινγκ και διατήρησης πελατών τους. Η συνιστώσα ανάλυσης συναισθήματος υπογραμμίζει την αξία της κατανόησης των ανατροφοδοτήσεων των πελατών σε βάθος, παρέχοντας μια διπλή προσέγγιση για την αφοσίωση των πελατών: μέσω της προγνωστικής ανάλυσης και των ποιοτικών γνώσεων.

Η έρευνα αναγνωρίζει τους περιορισμούς, συμπεριλαμβανομένων των προκλήσεων στην τμηματοποίηση των πελατών αποκλειστικά βάσει δεδομένων συναλλαγών και της πολυπλοκότητας της ακριβούς καταγραφής του συναισθήματος των πελατών. Προτείνεται η μελλοντική εργασία για τη διερεύνηση πιο εξελιγμένων

αλγορίθμων μηχανικής μάθησης και τεχνικών ανάλυσης συναισθήματος, με στόχο μια ολιστική προσέγγιση της διαχείρισης των πελατειακών σχέσεων στο ηλεκτρονικό εμπόριο. Η παρούσα έρευνα συμβάλλει στο υπάρχον σώμα γνώσεων, καταδεικνύοντας την πρακτική εφαρμογή της ανάλυσης RFM, της πρόβλεψης CLV και της ανάλυσης συναισθήματος για την ενίσχυση των στρατηγικών διαχείρισης πελατειακών σχέσεων στον τομέα του ψηφιακού εμπορίου, δημιουργώντας ένα θεμέλιο πάνω στο οποίο μπορεί να βασιστεί μια μελλοντική έρευνα με δυνατότητα να επηρεάσει σημαντικά τον τρόπο με τον οποίο οι επιχειρήσεις κατανοούν και αλληλεπιδρούν με τους πελάτες τους σε μια ολοένα και πιο ψηφιακή αγορά.

Περιεχόμενα

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κατάλογος Συντομογραφιών

AI Artificial Intelligence

AUC Area under the ROC Curve

BN Bayesian Network

B2B Business to Business

B2C Business to Customer

CLV Customer Lifetime Value

CNN Convolutional Neural Network

COVID Coronavirus disease

CRM Customer Relationship Management

CSI Customer Satisfaction Index

DF DataFrame

EC Ecommerce

ECT Expectation Confirmation Theory

EM Elbow Method

IT Information Technology

LGBM Light Gradient-Boosting Machine

ME Maximum Entropy

MEC Maximum Entropy Classifier

NB Naïve Bayes Classifier

NLP Natural Language Processing

NLTK Natural Language Toolkit

NN Neural Networks

PLS Partial Least Squares

PSQ Perceived Service Quality

RFM Recency Frequency Monetary

RNN Recurrent Neural Network

SI Silhouette Index

SMB Small & Midsize Business

SVM Support Vector Machines

WSS Within-Clusters Sum of Squares

WWW World Wide Web

Κεφάλαιο 1: Εισαγωγή

1.1 ΕΙΣΑΓΩΓΗ

Τις τελευταίες δεκαετίες το ηλεκτρονικό εμπόριο έχει εξελιχθεί ραγδαία. Όλα ξεκίνησαν στις αρχές του 1990 με την άφιξη του Παγκόσμιου Ιστού, η οποία έθεσε τα θεμέλια για το ηλεκτρονικό εμπόριο (Ashton, 2020). Στα τέλη του 1990 αρχές του 2000, ξεκίνησαν την λειτουργία τους πρωτοποριακές εταιρείες του ηλεκτρονικού εμπορίου όπως η Amazon και το eBay, οι οποίες έφεραν στο προσκήνιο την νέα συνθήκη πώλησης και αγοράς προϊόντων στο διαδίκτυο σε μεγάλη κλίμακα. Ταυτόχρονα το 2000 παρατηρήθηκε και η εμφάνιση διαφόρων ηλεκτρονικών επιχειρήσεων, από εταιρείες που αναγνώρισαν την προοπτική εξέλιξης στην παγκόσμια αγορά μέσω της χρήσης διαδικτύου (Deng et al., 2021). Με την εξέλιξη της τεχνολογίας, η υιοθέτηση έξυπνων τηλεφώνων (smartphones) και η εξέλιξη ασφαλών πυλών πληρωμής έκαναν τις ηλεκτρονικές αγορές ακόμη πιο προσιτές στους καταναλωτές. Τα τελευταία έτη έχει παρατηρηθεί ραγδαία ανάπτυξη των ηλεκτρονικών αγορών μέσω κινητών τηλεφώνων, ενώ η εισαγωγή της Τεχνητής Νοημοσύνης (AI) και της ανάλυσης δεδομένων έχουν μετατρέψει το ηλεκτρονικό εμπόριο σε μια εξατομικευμένη εμπειρία αγορών για τον κάθε καταναλωτή. Επιπλέον, η πανδημία του Κορονοϊού (COVID-19) κατά την διάρκεια του 2020 επιτάχυνε την στροφή προς τις ηλεκτρονικές αγορές, καθιστώντας το ηλεκτρονικό εμπόριο αναπόσπαστο μέρος του σύγχρονου εμπορίου (Loginova, 2022). Σήμερα το ηλεκτρονικό εμπόριο συνεχίζει να εξελίσσεται με καινοτομίες στην επαυξημένη πραγματικότητα, την εικονική πραγματικότητα και την τεχνολογία blockchain, υπόσχοντας ένα μέλλον όπου η εμπειρία αγορών στο διαδίκτυο γίνεται ακόμη πιο καθηλωτική και ασφαλής.

Τρεις είναι οι κατηγορίες ηλεκτρονικού εμπορίου που έχουν κυριαρχήσει ανά τις δεκαετίες:

α) Το μοντέλο Business-to-Business (B2B) περιλαμβάνει μια επιχείρηση, η οποία πουλάει προϊόντα ή υπηρεσίες σε μια άλλη επιχείρηση. Βασικό παράδειγμα αυτού του μοντέλου είναι οι εταιρείες λογισμικού που εστιάζουν στην παροχή υπηρεσιών, αλλά,

επίσης αντίστοιχες εταιρείες που μπορούν να παρέχουν ηλεκτρονικές συσκευές ή έπιπλα γραφείου.

β) Το μοντέλο Business-to-Customer (B2C) που είναι η πιο συνήθης μορφή εμπορίου, ακολουθεί το κλασσικό μοντέλο λιανικής πώλησης όπου η επιχείρηση πουλάει τα προϊόντα της στον καταναλωτή. Η διαδικασία της αγοράς συνήθως είναι πιο σύντομη, οι αποφάσεις είναι βασισμένες στο συναίσθημα του καταναλωτή, ενώ ταυτόχρονα και τα ποσά αγοράς είναι σημαντικά μικρότερα από ότι στο προηγούμενο επιχειρησιακό μοντέλο.

γ) Τέλος, στις κατηγορίες ηλεκτρονικού εμπορίου συγκαταλέγεται και η μορφή Consumer-to-Consumer, στην οποία μεμονωμένα άτομα πωλούν απευθείας τα προϊόντα τους ο ένας στον άλλο, συχνά μέσω της χρήσης κάποιας ιστοσελίδα ενός τρίτου ατόμου. Τέτοιες ιστοσελίδες, συνήθως, χρεώνουν μια προμήθεια για τη φιλοξενία αυτών των συναλλαγών μεταξύ των καταναλωτών (Cano et al., 2023).

Σε αυτό το πλαίσιο της ραγδαίας ανάπτυξης του ηλεκτρονικού εμπορίου κάθε επιχείρηση επιθυμεί την κατάκτηση όλο και μεγαλύτερου μεριδίου της αγοράς. Για πολλά χρόνια οι επιχειρήσεις εστίαζαν περισσότερο στα προϊόντα παρά στους τελικούς χρήστες (Mosaddegh et al., 2021). Στην σημερινή εποχή, όμως, οι επιχειρήσεις έχουν αντιληφθεί την σημαντική αξία των καταναλωτών και την διατήρηση του αγοραστικού κοινού τους, για αυτό το λόγο επενδύουν πολύ περισσότερο σε στρατηγικές μάρκετινγκ και μελέτες με τελικό σκοπό την μεγιστοποίηση των κερδών (Mosaddegh et al., 2021).

Η έρευνα της συμπεριφοράς των καταναλωτών είναι ένα διευρυμένο πεδίο που συγκεντρώνει πολλές μεθόδους μελέτης. Πολλές μελέτες έχουν ερευνήσει τις διαφορετικές μεθόδους κατηγοριοποίησης πελατών. Η κατηγοριοποίηση των πελατών είναι ένα στρατηγικό εργαλείο των επιχειρήσεων που χρησιμοποιείται για την ομαδοποίηση των πελατών βάσει των κοινών τους αναγκών, με σκοπό να αναγνωρίσουν τελικά τους πιο κερδοφόρους και να διαχειριστούν τους πόρους της επιχείρησης αποτελεσματικά για την προσέλκυση και διατήρηση αυτών (Mosaddegh et al., 2021). Τα δεδομένα είναι το «καύσιμο» της ηλεκτρονικής οικονομίας. Μέσω της τεχνολογίας και των κινητών τηλεφώνων πλέον παράγεται τεράστιος όγκος δεδομένων για το πως οι καταναλωτές αισθάνονται, συμπεριφέρονται, αλληλοεπιδρούν με τα προϊόντα, τις υπηρεσίες αλλά και τις στρατηγικές μάρκετινγκ που οι επιχειρήσεις προωθούν στο κοινό (Wedel & Kannan, 2016).

Η ανάλυση δεδομένων και η ανάπτυξη της τεχνητής νοημοσύνης στον κλάδο του μάρκετινγκ επιχειρεί την εφεύρεση ευρημάτων από τα δεδομένα για την ανάληψη αποφάσεων. Η εξόρυξη δεδομένων μπορεί να βοηθήσει τον σκοπό αυτό και να οδηγήσει στην δημιουργία μοντέλων πρόβλεψης της συμπεριφοράς των καταναλωτών από μεγάλες βάσεις δεδομένων (Fan et al., 2015). Η κατηγοριοποίηση πελατών είναι μια μέθοδος εξόρυξης δεδομένων που θα μελετήσουμε και αναλύσουμε περεταίρω στην συνέχεια. Μία μέθοδος, η οποία μπορεί να βοηθήσει στην πρόβλεψη του κόστους απόκτησης του πελάτη (Customer Lifetime Value) (Mosaddegh et al., 2021). Το CLV στο μάρκετινγκ ορίζεται ως η αξία διάρκειας ζωής του πελάτη. Η αξία αυτή αναφέρεται στην συνολική αξία του καταναλωτή κατά τη διάρκεια της σχέσης του με την επιχείρηση (Stahl et al., n.d.). Θεωρείται από τις πιο σημαντικές μετρήσεις στον επιχειρησιακό κόσμο του εμπορίου. Η παρατήρηση αυτού του μέτρου βοηθά μακροπρόθεσμα τις επιχειρήσεις να κατανείμουν τους πόρους πιο αποτελεσματικά.

Στην συγκεκριμένη μελέτη θα χρησιμοποιηθεί η RFM ανάλυση (Recency, Frequency, Monetary), μια διάσημη τεχνική κατηγοριοποίησης πελατών για την αξιολόγηση τους με βάση τα μοτίβα συμπεριφοράς στις αγορές τους (Christy et al., 2021). Μια μέθοδο αξιολόγησης που μετρά πότε (recency), πόσο συχνά (frequency) και με τί έσοδα (monetary) πραγματοποιήθηκαν οι συναλλαγές από τους πελάτες της επιχείρησης. Στο τέλος, αυτές οι μετρήσεις των τριών μεταβλητών ενώνονται σε ένα RFM σκορ (Ma & Guo, 2010), το οποίο χρησιμοποιείται για να προβλέψει μελλοντικά μοτίβα συμπεριφοράς. Εφόσον έχουν ολοκληρωθεί οι μετρήσεις των τριών αυτών μεταβλητών, ο K-means αλγόριθμος εφαρμόζεται στις παραπάνω μεταβλητές για να ομαδοποιήσει τους πελάτες (Christy et al., 2021). Η συμπεριφορά της κάθε ομάδας πελατών αναλύεται για να βρεθεί τελικώς, η πιο κερδοφόρα για την επιχείρηση. Σίγουρα η επέκταση και η προσέλκυση νέων πελατών με στρατηγικές μάρκετινγκ είναι πολύ ελκυστική, ταυτόχρονα, όμως η διατήρηση του ήδη υπάρχων πελατολογίου είναι ακόμη πιο σημαντική αλλά και απαιτητική (Kastouni & Lahcen, 2020).

Στο ταχέως εξελισσόμενο τοπίο του ηλεκτρονικού εμπορίου, όπου η κατανόηση της συμπεριφοράς των πελατών έχει καταστεί υψίστης σημασίας για τις επιχειρήσεις που στοχεύουν στην ενίσχυση της αφοσίωσης και την αύξηση των πωλήσεων. Οι παραδοσιακές αναλυτικές μέθοδοι, όπως η ανάλυση RFM (συχνότητα, συχνότητα, νομισματική αξία), έχουν προσφέρει πολύτιμες πληροφορίες για την τμηματοποίηση των πελατών και τα αγοραστικά πρότυπα. Ωστόσο, η έλευση των μεγάλων δεδομένων και των προηγμένων τεχνολογιών ανάλυσης προσφέρει την ευκαιρία να εμβαθύνουμε

στις αποχρώσεις των αλληλεπιδράσεων των πελατών. Η παρούσα μελέτη αναγνωρίζει τη σημασία της ανάλυσης RFM και το συνδυασμό της με την εκτίμηση της CLV ως ένα θεμελιώδη εργαλεία για την αξιολόγηση της αξίας των πελατών και εισάγει την ανάλυση συναισθήματος ως μια συμπληρωματική προσέγγιση. Εξετάζοντας όχι μόνο τις συναλλαγές αλλά και τις συναισθηματικές διαστάσεις των σχέσεων με τους πελάτες, η παρούσα έρευνα στοχεύει στην παροχή μιας ολιστικής άποψης της αφοσίωσης και της ικανοποίησης των πελατών.

1.2 ΣΚΟΠΟΣ

Στην περίπτωση του ηλεκτρονικού καταστήματος που θα μελετηθεί παρακάτω, θα χρησιμοποιηθούν όλες οι προαναφερόμενες τεχνικές ανάλυσης δεδομένων με σκοπό την μελέτη της συμπεριφοράς του καταναλωτή και την εξαγωγή συμπερασμάτων. Συμπεράσματα γύρω από τις αγορές του και την κερδοφορία αυτών για την επιχείρηση. Ο πρωταρχικός σκοπός της παρούσας μελέτης είναι να διερευνήσει την εφαρμογή της ανάλυσης RFM στο πλαίσιο του ηλεκτρονικού εμπορίου για την τμηματοποίηση των πελατών και την πρόβλεψη της μελλοντικής συμπεριφοράς τους. Αναγνωρίζοντας τους περιορισμούς των αμιγώς ποσοτικών μεθόδων στην καταγραφή του πλήρους φάσματος της αφοσίωσης των πελατών, η παρούσα έρευνα διευρύνει το πεδίο εφαρμογής της ώστε να συμπεριλάβει την ανάλυση συναισθήματος. Με την ενσωμάτωση τεχνικών επεξεργασίας φυσικής γλώσσας (NLP) για την ανάλυση των κριτικών και των αναφορών των πελατών, η μελέτη επιδιώκει να αποκαλύψει τις ποιοτικές πτυχές της ικανοποίησης των πελατών που επηρεάζουν την αφοσίωση και τις αγοραστικές αποφάσεις. Αυτή η διπλή προσέγγιση αποσκοπεί στην ενίσχυση της προβλεπτικής δύναμης των μοντέλων αξίας διάρκειας ζωής των πελατών και στην προσφορά αξιοποιήσιμων πληροφοριών για στοχευμένες στρατηγικές μάρκετινγκ και εξατομικευμένη δέσμευση πελατών.

Κεφάλαιο 2: Επισκόπηση Αρθρογραφίας

2.1 ΗΛΕΚΤΡΟΝΙΚΟ ΕΜΠΟΡΙΟ (e-commerce)

Το ηλεκτρονικό εμπόριο σαν όρος (EC) έκανε την εμφάνισή του κατά τη δεκαετία του 1970 (Wigand, 2003). Το EC περιλαμβάνει όλες τις οικονομικές δραστηριότητες που διεξάγονται μέσω ηλεκτρονικών συνδέσεων, συμπεριλαμβανομένου ενός ευρέος φάσματος βιομηχανιών όπως η χρηματοδότηση, ο τουρισμός, η μεσιτεία, η ασφάλιση, η εφοδιαστική και η διαχείριση πελατειακών σχέσεων. Χαρακτηρίζεται από μηχανισμούς συντονισμού της αγοράς και περιλαμβάνει αλλαγές στα οργανωτικά όρια, στις αλυσίδες αξίας και στην ενσωμάτωση των προμηθευτών και των πελατών σε αυτές τις αλυσίδες. Οι τεχνολογίες πληροφοριών και επικοινωνιών είναι ζωτικής σημασίας στο ηλεκτρονικό εμπόριο, επιτρέποντας την εξατομίκευση και τη μαζική προσαρμογή (Gierlich-Joas et al., 2019). Η ανάπτυξη των υπολογιστών, των τηλεπικοινωνιών, του Διαδικτύου και του Παγκόσμιου Ιστού έχει προωθήσει σημαντικά το ηλεκτρονικό εμπόριο, καθιστώντας το προσιτό στο ευρύ κοινό (Ashton, 2020). Η εξέλιξη αυτή εισήγαγε νέους παίκτες στην αγορά, συμπεριλαμβανομένων των ιδιωτών, και οδήγησε σε φαινόμενα όπως η αποδιαμεσολάβηση (disintermediation) και η επαναδιαμεσολάβηση (reintermediation), συχνά με τη συμμετοχή διαφορετικών παικτών.

Η αποδιαμεσολάβηση στο πλαίσιο των ηλεκτρονικών αγορών αναφέρεται στη μείωση ή την εξάλειψη των μεσαζόντων μεταξύ παραγωγών και καταναλωτών, οδηγώντας σε άμεσες συναλλαγές. Αυτό οφείλεται στην ικανότητα της τεχνολογίας των πληροφοριών να μειώνει το κόστος των συναλλαγών, επιτρέποντας την άμεση επαφή μεταξύ πωλητών και αγοραστών (Giaglis et al., 2002). Η επαναδιαμεσολάβηση, από την άλλη πλευρά, περιγράφει την εμφάνιση ή την επανεμφάνιση των διαμεσολαβητών σε νέους ρόλους. Οι παραδοσιακοί μεσάζοντες μπορεί να προσαρμοστούν και να βρουν ευκαιρίες να αξιοποιήσουν την τεχνογνωσία και τις οικονομίες κλίμακας που διαθέτουν στην ηλεκτρονική αγορά (Adelaar, 2000), ή μπορεί να εμφανιστούν μεσάζοντες άλλης μορφής, οι οποίοι παρέχουν υποστήριξη υποδομών για λειτουργίες της αγοράς που αναδιαρθρώνονται ριζικά στον κόσμο του ηλεκτρονικού εμπορίου (Giaglis et al., 2002).

Το Διαδίκτυο, και συγκεκριμένα ο Παγκόσμιος Ιστός, έχει διαδραματίσει καθοριστικό ρόλο στην επέκταση της εμβέλειας και της σημασίας του EC. Προσφέρει ένα δυναμικό μέσο που ξεπερνά τα παραδοσιακά μέσα όπως ο τύπος, το ραδιόφωνο ή η τηλεόραση (Ashton, 2020). Η καθολική συνδεσιμότητα και η πρόσβαση που παρέχει το Διαδίκτυο έχουν γίνει απαραίτητες για τις επιχειρήσεις και τους ιδιώτες, επηρεάζοντας τόσο την προσφορά όσο και τη ζήτηση του εμπορίου (Adelaar, 2000). Κατά συνέπεια, το ηλεκτρονικό εμπόριο αποτελεί βασικό στοιχείο των στρατηγικών σχεδίων πολλών επιχειρήσεων, με το Διαδίκτυο να λειτουργεί ως "παγκόσμιος τηλεφωνικός θάλαμος" για τις επιχειρηματικές δραστηριότητες. Αν και το EC που αφορά τους καταναλωτές λαμβάνει συχνά μεγαλύτερη προσοχή από τα μέσα ενημέρωσης, το ηλεκτρονικό εμπόριο μεταξύ επιχειρήσεων είναι σημαντικά μεγαλύτερο (Feike & Rösch, 2024a). Οι προβλέψεις για το συνολικό μέγεθος της αγοράς EC είναι τεράστιες, αντανακλώντας την εκτεταμένη επιρροή και εμβέλειά της στον σύγχρονο επιχειρηματικό κόσμο.

Οι κορυφαίες διαδικτυακές αγορές στον κόσμο, συμπεριλαμβανομένων των Alibaba, Amazon και το eBay, πούλησαν 2,7 τρισεκατομμύρια δολάρια το 2020, που ισοδυναμεί με το 62% των παγκόσμιων διαδικτυακών πωλήσεων εκείνο το έτος (Jap et al., n.d.). Το ηλεκτρονικό εμπόριο έχει καταστεί ζωτικής σημασίας για το παγκόσμιο λιανικό εμπόριο λόγω του διαδικτύου. Δύο χρόνια μετά, από το 2022, οι πωλήσεις λιανικού ηλεκτρονικού εμπορίου ξεπέρασαν τα 5,7 τρισεκατομμύρια δολάρια παγκοσμίως (Koen van Gelder, 2023). Οι διαδικτυακές αγορές, με επικεφαλής την Amazon όσον αφορά την επισκεψιμότητα, κυριαρχούν στις παγκόσμιες διαδικτυακές αγορές. Ωστόσο, η Amazon βρίσκεται πίσω από την Taobao και την Tmall (Alibaba Group) σε ακαθάριστη αξία εμπορευμάτων. Μια σημαντική τάση είναι η άνοδος του εμπορίου μέσω κινητών τηλεφώνων (m-commerce), με πάνω από το 70% των επισκέψεων σε ιστότοπους λιανικής πώλησης και την πλειονότητα των παραγγελιών να πραγματοποιούνται μέσω smartphones το 2023 (Koen van Gelder, 2023). Στον τομέα B2C, η μόδα και τα ηλεκτρονικά είδη ευρείας κατανάλωσης κατέχουν ηγετική θέση στις διαδικτυακές λιανικές πωλήσεις.

Αν και οι πλατφόρμες B2C και B2B βασίζονται σε μεγάλο βαθμό στις επιδράσεις του δικτύου, οι επιδράσεις του δικτύου μπορεί να διαδραματίζουν διαφορετικό ρόλο ανάλογα με το αν η μία πλευρά της αγοράς είναι καταναλωτές ή αν και οι δύο πλευρές της αγοράς είναι επιχειρηματικοί χρήστες (Feike & Rösch, 2024b). Ένας σημαντικός παράγοντας για τις πλατφόρμες B2B είναι, για παράδειγμα, ότι το μεγαλύτερο μέρος

της χρήσης γίνεται στο πλαίσιο μιας προκαθορισμένης διαδικασίας ή ως εκπρόσωπος ενός συγκεκριμένου ρόλου εντός ενός οργανισμού αντί για ιδιωτικά συμφέροντα. Επιπλέον, η υψηλή πολυπλοκότητα των αγορών B2B λόγω των εξειδικευμένων προϊόντων, της ετερογενούς και κατακερματισμένης δομής της αγοράς καθιστά πιο δύσκολη τη μεταφορά ικανοτήτων από τη μία αγορά στην άλλη (Schreieck et al., 2019). Αντίθετα, τα δυνατά σημεία των πλατφορμών B2C είναι, μεταξύ άλλων, η υλοποίηση οικονομιών εμβέλειας στην πλευρά της προσφοράς μεταξύ των αγορών (Armstrong & Wright, 2007).

Η χρήση δεδομένων θεωρείται αμοιβαία επωφελής, στο πλαίσιο των πλατφορμών B2B και B2C αλλά με διαφορετικές εφαρμογές. Οι πλατφόρμες B2C είναι γνωστές για την προσέγγισή τους με γνώμονα τα δεδομένα, αξιοποιώντας τη συλλογή και ανάλυση δεδομένων για τη δημιουργία γνώσεων και την εισαγωγή νέων προϊόντων και υπηρεσιών στην αγορά (Condorelli & Padilla, 2020). Ομοίως, στις πλατφόρμες B2B, η πρόσβαση σε υψηλής ποιότητας και συνεπή δεδομένα αποτελούν σημαντικό πλεονέκτημα για τους συμμετέχοντες (Feike & Rösch, 2024a). Η κοινή χρήση δεδομένων προάγει τη μεγαλύτερη εξάρτηση και αφοσίωση μεταξύ των συμμετεχόντων, ενισχύοντας τις σχέσεις αγοραστή-προμηθευτή. Η δυνατότητα ανταλλαγής δεδομένων μεταξύ των οργανισμών αποτελεί βασική κινητήρια δύναμη για τη δημιουργία πλατφορμών, τη βελτίωση των βάσεων δεδομένων τους και την αξιοποίησή τους προσβάσιμα στην πλατφόρμα (Gierlich-Joas et al., 2019).

2.2 ΤΜΗΜΑΤΟΠΟΙΗΣΗ ΠΕΛΑΤΩΝ (customer segmentation)

Στο εξελισσόμενο τοπίο του ηλεκτρονικού εμπορίου, η εστίαση στον πελάτη, η ποιότητα των προϊόντων και η άριστη εξυπηρέτηση έχουν γίνει ο πυρήνας των πωλήσεων. Τα προϊόντα υψηλής ποιότητας, οι υπηρεσίες υψηλού επιπέδου, η ανάλυση και η πρόβλεψη της συμπεριφοράς των πελατών είναι ζωτικής σημασίας για αποτελεσματικές στρατηγικές μάρκετινγκ. Η υιοθέτηση μιας πελατοκεντρικής προσέγγισης είναι απαραίτητη, καθώς βοηθά στον εντοπισμό των ομάδων-στόχων καταναλωτών, στην κατανόηση των αναγκών τους και στην παροχή αξίας (He & Li, 2017). Η τμηματοποίηση των πελατών περιλαμβάνει τη διαίρεση των πελατών σε διακριτές ομάδες με βάση συγκεκριμένα κριτήρια (Vagn et al., 2001). Αυτό είναι σημαντικό επειδή οι ανάγκες και οι αξίες των πελατών ποικίλλουν - ορισμένοι, αν και λιγότεροι σε αριθμό, μπορεί να συμβάλλουν σημαντικά στα κέρδη. Η κατανόηση

αυτών των διαφορών επιτρέπει στις επιχειρήσεις να προσαρμόζουν τις στρατηγικές τους, να προσφέρουν εξατομικευμένα προϊόντα και υπηρεσίες και τελικά να μεγιστοποιούν τη χρήση των πόρων των πελατών για βέλτιστη κερδοφορία.

Το κλειδί της επιτυχίας στο λιανικό εμπόριο είναι η δημιουργία ισχυρών σχέσεων με τους πελάτες, οι οποίες οδηγούν στην αφοσίωση στη μάρκα, στις επαναλαμβανόμενες επισκέψεις και στις μετατροπές πωλήσεων (Morales et al., 2017). Οι πρόσφατες οικονομικές και κοινωνικές αλλαγές έχουν πιέσει τους λιανοπωλητές να χαράξουν αποτελεσματικότερη στρατηγική, να κατανοήσουν σε βάθος τους καταναλωτές, τους ανταγωνιστές τους και να εστιάσουν στη συμπεριφορά και την αφοσίωση των πελατών τους (M. C. Chen et al., 2005). Η τμηματοποίηση της αγοράς, ο διαχωρισμός των δυνητικών πελατών σε παρόμοιες ομάδες με βάση κοινά χαρακτηριστικά, είναι απαραίτητη. Η προσέγγιση αυτή βοηθά τους λιανοπωλητές να αποφεύγουν τα μη κερδοφόρα προϊόντα και να διαχειρίζονται καλύτερα τους πόρους στοχεύοντας σε πολλά υποσχόμενα τμήματα της αγοράς (Yoseph et al., 2020).

Πολλές μελέτες έχουν επικεντρωθεί στη συμπεριφορά αγοράς των πελατών και στη δια βίου αξία σε διάφορα προϊόντα, χρησιμοποιώντας την τμηματοποίηση της αγοράς με βάση δημογραφικές μεταβλητές και χαρακτηριστικά. Ωστόσο, οι περισσότερες από αυτές τις μελέτες εξετάζουν μόνο μια συγκεντρωτική άποψη των ιστορικών δεδομένων των καταναλωτών για να συμπεράνουν ομοιότητες μεταξύ τους. Αυτή η προσέγγιση συχνά παραβλέπει σημαντικές λεπτομέρειες σχετικά με τους μεμονωμένους καταναλωτές, καθώς γενικεύει με βάση τα χαρακτηριστικά της ομάδας αντί να εστιάζει στην ατομική αγοραστική συμπεριφορά (Yoseph et al., 2020). Αυτό αναδεικνύει ένα σημαντικό κενό στην κατανόηση των αποχρώσεων της ατομικής συμπεριφοράς των καταναλωτών στο πλαίσιο της έρευνας για την τμηματοποίηση της αγοράς.

2.2.1 Μέθοδος RFM Ανάλυσης

Το μοντέλο RFM είναι μια ευρέως χρησιμοποιούμενη στατιστική προσέγγιση στο λιανικό μάρκετινγκ για την τμηματοποίηση της συμπεριφοράς των πελατών και την αξιολόγηση της αξίας ζωής των καταναλωτών. Η ανάλυση RFM έκανε για πρώτη φορά την εμφάνισή της το 1995 από τους ερευνητές Bult και Wansbeek (Blattberg et al., 2008), και φάνηκε πολλά υποσχόμενη κατά την εφαρμογή της σε δεδομένα μάρκετινγκ.

Η μέθοδος RFM κατηγοριοποιεί τους πελάτες με βάση το ιστορικό των αγορών τους, λαμβάνοντας υπόψη πόσο πρόσφατες, συχνές και σημαντικές είναι οι αγορές τους (Yoseph et al., 2020). Το "R" συμβολίζει το «Recency», που ισούται με το χρόνο από την τελευταία αγορά. Πιο συγκεκριμένα, αναφέρεται στο χρονικό διάστημα μεταξύ του χρόνου που συμβαίνει η τελευταία καταναλωτική συμπεριφορά και του παρόντος. Πολλοί έμποροι άμεσων πωλήσεων πιστεύουν ότι οι πιο πρόσφατοι αγοραστές είναι πιο πιθανό να αγοράσουν ξανά από ό,τι οι λιγότερο πρόσφατοι αγοραστές (Birant, n.d.). Το "F" σημαίνει «Frequency», δηλαδή πόσο συχνά γίνονται αγορές. Η συχνότητα είναι ο αριθμός των συναλλαγών που έχει πραγματοποιήσει ένας πελάτης μέσα σε μια ορισμένη περίοδο. Το μέτρο αυτό χρησιμοποιείται με βάση την υπόθεση ότι οι πελάτες με περισσότερες αγορές είναι πιο πιθανό να αγοράσουν προϊόντα από τους πελάτες με λιγότερες αγορές (Birant, n.d.). Τέλος, το "M" σημαίνει «Monetary value» δηλαδή χρηματική αξία. Η νομισματική αυτή μετρική αναφέρεται στο αθροιστικό σύνολο των χρημάτων που δαπανήθηκαν από ένα συγκεκριμένο πελάτη.

Οι πελάτες στη συνέχεια βαθμολογούνται σε ομάδες από το καλύτερο προς το χειρότερο, συνήθως σε αύξουσα κλίμακα. Η συνολική βαθμολογία RFM, η οποία συχνά αναπροσαρμόζεται σε ένα εύρος 0-1, είναι ένας σταθμισμένος μέσος όρος αυτών των τριών παραγόντων, βοηθώντας στον εντοπισμό πελατών που είναι πιο πιθανό να ανταποκριθούν σε προωθητικές ενέργειες και προσπάθειες εξατομίκευσης (Shreya Tripathi et al., 2018).

Βαθμολογία RFM = (βαθμολογία χρονικής εγγύτητας × βάρος χρονικής εγγύτητας) + (βαθμολογία συχνότητας × βάρος συχνότητας + (νομισματική βαθμολογία x νομισματικό βαρος)

Αναβαθμισμένη βαθμολογία RFM = (βαθμολογία RFM - ελάχιστη βαθμολογία RFM) / (Μέγιστη βαθμολογία RFM - ελάχιστη βαθμολογία RFM)

Το μοντέλο RFM είναι το πιο συχνά υιοθετούμενο μοντέλο τμηματοποίησης που περιλαμβάνει τρία μέτρα (χρονική εγγύτητα, συχνότητα και χρηματικό ποσό), τα οποία συνδυάζονται σε έναν τριψήφιο κωδικό RFM. Μεταξύ των τριών μέτρων RFM, η χρονική εγγύτητα θεωρείται συχνά ως το πιο σημαντικό μέτρο (Wei et al., 2010). Ωστόσο, σύμφωνα με προηγούμενα ευρήματα, οι τιμές RFM τείνουν να είναι συγκεκριμένες για την επιχείρηση και βασίζονται στη φύση των προϊόντων (Lumsden et al., 2008). Για παράδειγμα, μελέτες (Fader et al., 2005) διαπίστωσαν ότι για χαμηλότερη ανακύκλωση, οι πελάτες με υψηλότερη συχνότητα τείνουν να έχουν

χαμηλότερη μελλοντική πιθανότητα αγοραστικής ικανότητας από εκείνους με χαμηλότερα ποσοστά προαγοράς. Πρόσφατα, ερευνητές (Lumsden et al., 2008) ενίσχυσαν την θεωρία αυτή καθώς είχαν παρόμοια ευρήματα, ότι υπάρχουν σημαντικές διαφορές μεταξύ των ομάδων κατά την επαναληπτικότητα και τη συχνότητα.

Η χρονική εγγύτητα ή αλλιώς περιοδικότητα, ορίζεται συνήθως από τον αριθμό των περιόδων από την τελευταία αγορά, ο οποίος μετρά το διάστημα μεταξύ του πιο πρόσφατου χρόνου συναλλαγής και του χρόνου ανάλυσης (ημέρες ή μήνες). Συνεπώς, όσο μικρότερος είναι ο αριθμός των ημερών, τόσο υψηλότερη είναι η βαθμολογία της περιοδικότητας. Ένας πελάτης που έχει υψηλή βαθμολογία επαναφοράς συνεπάγεται ότι είναι πιο πιθανό να προβεί σε επαναλαμβανόμενη αγορά (Lin, 2010).

Ο ορισμός της συχνότητας συχνά απλοποιείται για να εξετάσει δύο καταστάσεις, συμπεριλαμβανομένων των απλών και επαναλαμβανόμενων αγορών. Ένας πελάτης που έχει υψηλή βαθμολογία συχνότητας υποδηλώνει ότι έχει μεγάλη ζήτηση για το προϊόν και είναι περισσότερο πιθανό να αγοράζει τα προϊόντα της επιχείρησης επανειλημμένα. Για τη χρηματική μετρική «Monetary value», οι πελάτες κωδικοποιούνται με βάση το συνολικό ποσό των χρημάτων που δαπανήθηκαν κατά τη διάρκεια μιας συγκεκριμένης χρονικής περιόδου. Ωστόσο, μια άλλη έρευνα στο παρελθόν, (Claudio Marcus, 1998) πρότεινε ότι είναι καλύτερο να χρησιμοποιείται ο μέσος όρος αγοράς, παρά το συνολικό συσσωρευμένο ποσό αγοράς, ώστε να μειωθεί η συν-γραμμικότητα της συχνότητας και της χρηματικής μετρικής.

Στη μια διαφορετική μελέτη (Miglautsch, 2000), η βαθμολόγηση RFM περιλαμβάνει την κατάταξη των πελατών από τους καλύτερους στους χειρότερους, εξασφαλίζοντας ίσο αριθμό σε κάθε τμήμα. Ενώ η μέθοδος αυτή κατανέμει ομοιόμορφα τους πελάτες, αντιμετωπίζει προκλήσεις, ιδίως όσον αφορά τη μέτρηση της συχνότητας. Τείνει να διαχωρίζει ανακριβώς τους πελάτες με παρόμοιες συμπεριφορές σε χαμηλότερα τμήματα και, αντίθετα, ομαδοποιεί πελάτες με αξιοσημείωτα διαφορετικές αγοραστικές συνήθειες, όπως παρατήρησαν άλλοι ερευνητές αργότερα (Alam et al., 2009).

Ο John Wirth, ιδρυτής της Woodworker's Supply of New Mexico, ανέπτυξε μια εναλλακτική μέθοδο βαθμολόγησης, που ονομάζεται μέθοδος βαθμολόγησης πεντάδων συμπεριφοράς. Αυτή η προσέγγιση κατατάσσει τους πελάτες με βάση τη συμπεριφορά τους, οδηγώντας σε διαφορετικό αριθμό πελατών σε κάθε πεντάδα. Αποδίδει βαθμολογίες με βάση το χρονικό διάστημα από την τελευταία αγορά, με διαστήματα που κυμαίνονται από 0 έως πάνω από 25 μήνες (McCarty & Hastak, 2007).

Παρά την αντιμετώπιση ορισμένων ζητημάτων της παραδοσιακής μεθόδου πεντάδας πελατών, εξακολουθεί να μοιράζεται παρόμοιες προκλήσεις όσον αφορά την ακριβή μέτρηση της συχνότητας. Ο Miglautsch πρότεινε μια υβριδική προσέγγιση, που συνδυάζει τη βαθμολόγηση με βάση το πεντάγραμμο συμπεριφοράς με τη βαθμολόγηση με βάση τον μέσο όρο, για την καλύτερη αντιμετώπιση αυτών των προκλήσεων (Miglautsch, 2000). Σε αυτή τη μέθοδο, οι μεμονωμένοι αγοραστές λαμβάνουν χαμηλή βαθμολογία και οι μέσες τιμές συχνότητας χρησιμοποιούνται για την απόδοση βαθμολογίας στους υπόλοιπους. Το νομισματικό στοιχείο εξακολουθεί να διαιρεί τους πελάτες σε πέντε πεμπτημόρια με ίσα ποσά πωλήσεων.

Πέρα από την αξιολόγηση των πελατών με βάση τις μεμονωμένες τιμές RFM, ορισμένες έρευνες προτείνουν μια μέθοδο δημιουργίας τμημάτων πελατών συγκρίνοντας τις μέσες τιμές RFM μιας ομάδας με τους συνολικούς μέσους όρους (Lin, 2010). Εάν ο μέσος όρος RFM μιας συστάδας είναι μικρότερος ή μεγαλύτερος από τον συνολικό μέσο όρο, αποδίδεται μια δυαδική ταξινόμηση. Η συνδυασμένη τιμή RFM υπολογίζεται πολλαπλασιάζοντας τις κανονικοποιημένες τιμές RFM κάθε πελάτη με τα βάρη των μεταβλητών RFM, όπως διερευνήθηκε σε σύγχρονες μελέτες (D.-R. Liu & Shih, 2004), (D.-R. Liu & Shih, 2005), (Sohrabi & Amir, 2007).

Στο μοντέλο RFM, η δημιουργία μιας ενιαίας τιμής μπορεί να γίνει με δύο βασικούς τρόπους. Η μέθοδος του Libey περιλαμβάνει το άθροισμα της χρονικής εγγύτητας, της συχνότητας και των χρηματικών αξιών, συμπεριλαμβανομένης της μέσης παραγγελίας και της συχνότητας ανά έτος (Lin, 2010). Μια άλλη συνήθης μέθοδος είναι η απευθείας πρόσθεση των βαθμολογιών RFM. Ο Hughes (Arthur Middleton Hughes, 1994) υποστηρίζει την ίση στάθμιση κάθε μέτρου RFM, ενώ ο Miglautsch (Miglautsch, 2000) προτείνει τη δυνατότητα απόδοσης διαφορετικών βαρών. Ο Miglautsch παρουσιάζει, επίσης, συγκεκριμένους τύπους για τον υπολογισμό μιας σύνθετης βαθμολογίας με διαφορετικά βάρη για κάθε στοιχείο RFM. Αντίθετα, άλλοι μελετητές (D.-R. Liu & Shih, 2004) συνιστούν ότι η συνολική βαρύτητα όλων των μέτρων RFM πρέπει να ισούται με 1. Ο Stone (Stone & Jacobs, 1995) προσαρμόζει τα βάρη με βάση τα χαρακτηριστικά του προϊόντος και του κλάδου.

Το βάρος της συχνότητας στο μοντέλο RFM καθορίζεται πολλαπλασιάζοντας τη συχνότητα αγοράς επί 4. Για τη χρηματική αξία, το ποσό αγοράς πολλαπλασιάζεται επί 10%, με μέγιστη τιμή το 9. Οι Liu και Shih χρησιμοποίησαν μια αναλυτική ιεραρχική διαδικασία για να αποδώσουν σχετικά βάρη στις μεταβλητές RFM, αποφεύγοντας την αυθαίρετη στάθμιση (D.-R. Liu & Shih, 2005). Αντίθετα, οι McCarty και Hastak

χρησιμοποίησαν την προηγούμενη εμπειρία και την κρίση του έμπορου για να σταθμίσουν τις μεταβλητές RFM, γνωστή ως RFM με βάση την κρίση (McCarty & Hastak, 2007). Μια εναλλακτική μέθοδος RFM που βασίζεται στην εμπειρία περιλαμβάνει μια διαδικασία δύο σταδίων με δοκιμαστικές αποστολές και ανάλυση του ποσοστού ανταπόκρισης για τον προσδιορισμό της σημασίας των μέτρων RFM (Arthur Middleton Hughes, 1994), η συχνότητα μετράται για μια περίοδο επτά εβδομάδων, με κάθε περίπτωση να διαρκεί τρεις ώρες και να ξεπερνά ένα καθορισμένο όριο δικτύου, και η χρηματική αξία υπολογίζεται με βάση την αναλογία της μηνιαίας μίσθωσης δικτύου προς την κίνηση. Οι Lumsden κ.ά. προσέγγισαν διαφορετικά το θέμα αυτό στη μελέτη τους για ένα ιδιωτικό ταξιδιωτικό κλαμπ, χρησιμοποιώντας το πιο πρόσφατο έτος αγοράς διακοπών για τη χρονική εγγύτητα, την αναλογία των διακοπών προς τη διάρκεια της ιδιότητας μέλους για τη συχνότητα και τη μέση δαπάνη διακοπών για τη χρηματική αξία (Lumsden et al., 2008). Ο Chan ανέλυσε τη συμπεριφορά σε διαδικτυακές δημοπρασίες, ορίζοντας την περιοδικότητα ως τη διάρκεια της προσφοράς, τη συχνότητα ως το σύνολο των προσφορών που έγιναν και τη νομισματική αξία ως το τελικό ποσό της προσφοράς (M. C. Chen et al., 2005). Η εφαρμογή του μοντέλου RFM ποικίλλει ανάλογα με την εστίαση της μελέτης, με διαφορετικές παραμέτρους που χρησιμοποιούνται για τη μέτρηση των συνιστωσών RFM σε διάφορα πλαίσια.

2.2.2 Αξία διάρκειας ζωής πελάτη (CLV)

Η διαχείριση πελατειακών σχέσεων, γνωστή με τον όρο CRM θεωρείται ως ένας συνδυασμός μεθοδολογιών και διαδικασιών σε έναν οργανισμό, που εστιάζει στην απόκτηση και διατήρηση πελατών, ενισχύοντας την αφοσίωσή τους και ικανοποιώντας τις ανάγκες τους. Ενσαρκώνει μια επιχειρηματική φιλοσοφία που επικεντρώνεται στην προσέλκυση και διατήρηση πελατών, στην ενίσχυση της αφοσίωσης και της αξίας τους και στην εφαρμογή στρατηγικών που εστιάζουν στον πελάτη (Nenonen & Storbacka, 2016). Οι εταιρείες που υιοθετούν το CRM βελτιώνουν την αφοσίωση και την κερδοφορία των πελατών τους (Saarijarvi et al., 2013). Το CRM ενσωματώνει διάφορες οργανωτικές λειτουργίες, όπως οι πωλήσεις, το μάρκετινγκ και η τεχνική υποστήριξη, χρησιμοποιώντας συχνά την τεχνολογία της πληροφορίας για την αποτελεσματική διαχείριση των σχέσεων με τους πελάτες (Shim et al., 2012). Σύμφωνα με την Αμερικανική Ένωση Διοίκησης, η διατήρηση ενός υφιστάμενου πελάτη είναι σημαντικά λιγότερο δαπανηρή από την απόκτηση ενός νέου, υπογραμμίζοντας τη

σημασία των μακροχρόνιων σχέσεων με τους πελάτες για τη διατήρηση του κέρδους και της ικανοποίησης (Kotler, 1974). Οι επιτυχημένες επιχειρήσεις δίνουν προτεραιότητα στη διατήρηση των πελατών και στην ανάπτυξη σχέσεων, σε αντίθεση με τις προϊοντοκεντρικές επιχειρήσεις που εστιάζουν περισσότερο στο χαρτοφυλάκιο των προϊόντων τους. Η έρευνα δείχνει ότι η πελατοκεντρική προσέγγιση οδηγεί σε σημαντικά κέρδη, αμφισβητώντας το παραδοσιακό προϊοντοκεντρικό μοντέλο (Kumar, 2007).

Στον τομέα του μάρκετινγκ, για την επιτυχή εφαρμογή του CRM, είναι απαραίτητο οι εταιρείες να υιοθετήσουν στρατηγικές με επίκεντρο τον πελάτη, όπως η αξιολόγηση της αξίας του πελάτη, μια έννοια που έχει υπογραμμιστεί από τις πρώτες μελέτες γύρω από το μάρκετινγκ (Kotler, 1974). Αυτό περιλαμβάνει την αναγνώριση της μακροπρόθεσμης αξίας που συνεισφέρει ένας πελάτης σε μια εταιρεία. Ο προσδιορισμός αυτής της αξίας απαιτεί από τις επιχειρήσεις να επιλέξουν τις κατάλληλες μετρήσεις για την αξιολόγηση των πελατών (Asnawi & Setyaningsih, 2020). Η αξία διάρκειας ζωής του πελάτη (Customer Lifetime Value - CLV) είναι μια κρίσιμη έννοια στο CRM, η οποία αναγνωρίζεται για την αποτελεσματική αξιολόγηση της αξίας ενός πελάτη για μια επιχείρηση (Kumar, 2007).

Η Αξία Διάρκειας Ζωής Πελάτη (CLV), όπως ορίστηκε αρχικά (Kotler, 1974), είναι μια κρίσιμη μετρική στο μάρκετινγκ, που αντιπροσωπεύει την παρούσα αξία των μελλοντικών ροών κέρδους που αναμένονται από έναν πελάτη για ένα καθορισμένο χρονικό διάστημα. Η έννοια της CLV έχει προσελκύσει όλο και περισσότερο την προσοχή τα τελευταία χρόνια, με πολυάριθμες μελέτες που διερευνούν τις εφαρμογές της σε διάφορα πλαίσια του μάρκετινγκ (Reinartz & Kumar, 2000). Οι ερευνητές έχουν αναπτύξει μοντέλα για τον υπολογισμό της CLV, ορισμένα από τα οποία εστιάζουν σε ιστορικά δεδομένα πελατών και άλλα στην πρόβλεψη της μελλοντικής συμπεριφοράς των πελατών (Grover & Vriens, 2006). Η τελευταία προσέγγιση προτιμάται συχνά λόγω της έλλειψης ολοκληρωμένων ιστορικών δεδομένων. Ο ακριβής υπολογισμός της CLV είναι ζωτικής σημασίας για τις επιχειρήσεις ώστε να τμηματοποιούν αποτελεσματικά και να ιεραρχούν τους πελάτες, επιτρέποντας την ανάπτυξη στοχευμένων στρατηγικών μάρκετινγκ για κάθε τμήμα (Nenonen & Storbacka, 2016). Μια σύγχρονη μελέτη υπογραμμίζει τη σημασία του υπολογισμού της CLV για ομοιογενείς ομάδες πελατών αντί για μεμονωμένα άτομα, επιτρέποντας στις επιχειρήσεις να κατατάξουν αυτές τις ομάδες με βάση τη συνεισφορά τους στην κερδοφορία (Kumar, 2007). Αυτός ο υπολογισμός της CLV με βάση τις ομάδες βοηθά

τις επιχειρήσεις να διαφοροποιήσουν τις στρατηγικές μάρκετινγκ και σχέσεων με τους πελάτες τους, αντιμετωπίζοντας κάθε ομάδα ανάλογα με την αξία της. Επιπλέον, η κατανόηση της CLV βοηθά τις επιχειρήσεις να προσδιορίσουν το βέλτιστο επίπεδο επένδυσης στη διατήρηση πελατών, ώστε να εξασφαλίσουν θετική απόδοση της επένδυσης (Kumar, 2007). Αυτή η στρατηγική προσέγγιση στην αξιολόγηση και την επένδυση της αξίας των πελατών διασφαλίζει ότι οι πόροι κατανέμονται αποτελεσματικά για τη μεγιστοποίηση της κερδοφορίας των πελατών και της επιτυχίας της επιχείρησης.

Η αυξανόμενη σημασία του CLV στο μάρκετινγκ είναι εμφανής τόσο στην ακαδημαϊκή έρευνα όσο και στις επιχειρηματικές πρακτικές. Επιχειρήσεις όπως η Harrah's και η IBM χρησιμοποιούν το CLV για τη διαχείριση και την αξιολόγηση των επιχειρήσεων. Αυτή η στροφή προς την CLV οφείλεται στην ανάγκη για υπεύθυνο μάρκετινγκ πέρα από τις παραδοσιακές μετρήσεις όπως η αναγνωσιμότητα της μάρκας. Μια μελέτη σημείωσε τον πιθανό αρνητικό αντίκτυπο ορισμένων ενεργειών μάρκετινγκ στη μακροπρόθεσμη κερδοφορία (Gupta et al., 2006a). Τα χρηματοοικονομικά μέτρα, αν και χρήσιμα, αποτυγχάνουν να διαφοροποιήσουν επαρκώς μεταξύ κερδοφόρων και λιγότερο κερδοφόρων πελατών, οδηγώντας σε πιο διαφοροποιημένες στρατηγικές (Robert C. Blattberg et al., 2002), (Gupta et al., 2006a), (Rust et al., 2004). Η CLV χρησιμεύει ως λεπτομερής μετρική για την κερδοφορία των πελατών και καθοδηγεί την κατανομή των πόρων (Kumar, 2007), λειτουργώντας επίσης ως δείκτης της συνολικής αξίας μιας επιχείρησης (Gupta & Lehmann, 2003). Οι εξελίξεις στην τεχνολογία των πληροφοριών έχουν απλοποιήσει σημαντικά τη συλλογή εκτεταμένων δεδομένων από τις συναλλαγές των πελατών για τις επιχειρήσεις, επιτρέποντας την εστίαση στις πραγματικές προτιμήσεις αντί των προθέσεων. Με την πρόσβαση σε πλήρεις βάσεις δεδομένων πελατών, η δειγματοληψία έχει καταστεί περιττή και οι βελτιωμένες τεχνικές μοντελοποίησης δεδομένων επιτρέπουν στους εμπόρους να αντλούν πολύτιμες πληροφορίες (Gupta et al., 2006b).

Μια βασική πτυχή του CRM, λοιπόν, η CLV - επικεντρώνεται στην ενίσχυση των αλληλεπιδράσεων με τους πελάτες για τη βελτίωση της απόκτησης, της διατήρησης, της αφοσίωσης και της κερδοφορίας (Swift & Ronald S., 2000). Το CRM στοχεύει στη μεγιστοποίηση της αξίας ζωής ενός πελάτη για έναν οργανισμό (Peppers et al., 1999). Υπάρχουν διάφορες ταξινομήσεις μοντέλων CLV (Gupta et al., 2006a), συμπεριλαμβανομένων των μοντέλων RFM, των μοντέλων πιθανοτήτων, των οικονομετρικών μοντέλων, των μοντέλων επιμονής, των μοντέλων πληροφορικής και

των μοντέλων διάχυσης/ανάπτυξης (Gupta et al., 2006a). Το μοντέλο RFM, το οποίο σημειώνεται ιδιαίτερα για την απλότητα και την αποτελεσματικότητά του στο CRM, αξιολογεί την Ανακύκλωση, τη Συχνότητα και τη Χρηματική Αξία (Cheng & Chen, 2009).

Η CLV, εν γένει περιγράφεται συχνά ως η παρούσα αξία των μελλοντικών κερδών που θα αποφέρει ένας πελάτης κατά τη διάρκεια της σχέσης του με μια εταιρεία. Η έννοια αυτή είναι συγγενής με τη μέθοδο προεξόφλησης ταμειακών ροών στα χρηματοοικονομικά, αλλά με δύο σημαντικές διαφοροποιήσεις. Πρώτον, η CLV επικεντρώνεται σε μεμονωμένους πελάτες ή τμήματα, επιτρέποντας στις επιχειρήσεις να εντοπίζουν πιο κερδοφόρους πελάτες αντί να βασίζονται στη μέση κερδοφορία. Δεύτερον, σε αντίθεση με τις παραδοσιακές χρηματοοικονομικές προσεγγίσεις, η CLV συνυπολογίζει την πιθανότητα αποστασίας των πελατών από τους ανταγωνιστές σε μελλοντικές εκτιμήσεις (Gupta & Lehmann, 2003), (Reinartz & Kumar, 2000).

$$\mathrm{CLV} = \sum_{t=0}^{T} \frac{(p_t - c_t) r_t}{(1+i)^t} - \mathrm{AC}$$

p_t = τιμή πληρωμής από πελάτη κατά το χρονικό διάστημα t

c_t = κόστος εξυπηρέτησης πελάτη κατά το χρονικ΄πο διάστημα t

i = προεξοφλητικό επιτόκιο ή κόστος κεφαλαίου για την επιχείρηση

r_t = πιθανότητα επαναληπτικής αγοράς από τον πελάτη ή "ζωντανός" κατά το χρονική στιγμή t

AC = κόστος απόκτησης

T = χρονική στιγμή για την εκτίμηση της CLV

Ενώ η βασική έννοια της CLV είναι απλή, υπάρχουν διαφορές στη μοντελοποίηση και την εκτίμησή της μεταξύ των ερευνητών. Ορισμένοι έχουν χρησιμοποιήσει μια σταθερή διάρκεια ζωής των πελατών ή έναν προκαθορισμένο χρονικό ορίζοντα (Reinartz & Kumar, 2000), (Thomas, 2001), ενώ άλλοι έχουν εφαρμόσει έναν άπειρο χρονικό ορίζοντα (Fader, Hardie και Lee 2005- Gupta, Lehmann και Stuart 2004). Οι Gupta και Lehmann διαπίστωσαν ότι η χρήση της αναμενόμενης διάρκειας ζωής του πελάτη οδηγεί συχνά σε υπερεκτίμηση της CLV. Επιπλέον, οι ίδιοι ερευνητές απέδειξαν ότι εάν τα περιθώρια κέρδους και τα ποσοστά διατήρησης παραμένουν σταθερά με την πάροδο του χρόνου με άπειρο ορίζοντα, η CLV μπορεί να απλοποιηθεί (Gupta & Lehmann, 2003), (Gupta et al., 2004).

$$\mathrm{CLV} = \sum_{t=0}^{\infty} \frac{(p - c)r^t}{(1 + i)^t} = m \frac{r}{(1 + i - r)}$$

Στον παραπάνω τύπο το CLV απλά γίνεται περιθώριο (m) επί a πολλαπλάσιο περιθωρίου (r/1 + i - r). Όταν το ποσοστό διατήρησης είναι 90% και το προεξοφλητικό επιτόκιο είναι 12%, ο πολλαπλασιαστής περιθωρίου είναι περίπου 4. Οι Gupta και Lehmann έδειξαν ότι όταν τα περιθώρια αυξάνονται με σταθερό ρυθμό "g", ο πολλαπλασιαστής περιθωρίου γίνεται r/[1 + i - r(1 + g)], (Gupta et al., 2004). Τα περισσότερα μοντέλα για την εκτίμηση του CLV τείνουν να παραβλέπουν τον ανταγωνισμό λόγω μη διαθέσιμων ανταγωνιστικών δεδομένων. Η συχνότητα των επικαιροποιήσεων του CLV εξαρτάται από τη δυναμική της αγοράς, ιδίως σε ευμετάβλητες αγορές που επηρεάζονται από ανταγωνιστικούς παράγοντες. Στην έρευνα, τα μοντέλα συχνά επικεντρώνονται χωριστά στην απόκτηση πελατών, τη διατήρηση και το περιθώριο κέρδους ή συνδυάζουν στοιχεία αυτών. Μελετητές έχουν μοντελοποιήσει την απόκτηση και τη διατήρηση πελατών μαζί (Thomas, 2001), (Reinartz & Kumar, 2000). Αντίθετα, άλλοι ερευνητές ενσωμάτωσαν την επανάληψη και τη συχνότητα σε ένα μοντέλο, ενώ αντιμετώπισαν τη χρηματική αξία ξεχωριστά (Fader et al., 2004). Ωστόσο, οι μεθοδολογίες για τη μοντελοποίηση αυτών των στοιχείων ή της ίδιας της CLV διαφέρουν μεταξύ των ερευνητών.

2.2.3 Αλγόριθμος K-means

Η ομαδοποίηση K-Means είναι μια ευρέως αναγνωρισμένη μη ιεραρχική τεχνική ομαδοποίησης δεδομένων (Macqueen, 1967), (Güçdemir & Selim, 2015), η οποία χωρίζει ένα σύνολο δεδομένων σε έναν προκαθορισμένο αριθμό συστάδων, με πρωταρχικό στόχο την εξασφάλιση υψηλής ομοιότητας εντός των συστάδων (επίτευξη ενός ελάχιστου αθροίσματος τετραγώνων εντός κάθε συστάδας), διατηρώντας παράλληλα χαμηλή ομοιότητα μεταξύ διαφορετικών συστάδων (μεγιστοποίηση του αθροίσματος τετραγώνων μεταξύ συστάδων) (P. Li et al., 2022). Γνωστή για την απλή αλγοριθμική της προσέγγιση και τον αποτελεσματικό υπολογισμό της στον προσδιορισμό των κεντροειδών των συστάδων, η μέθοδος K-Means ξεχωρίζει ως μια ιδιαίτερα ευνοημένη μέθοδος στο πεδίο των τεχνικών συσταδοποίησης.

Η βασική διαδικασία της ομαδοποίησης K-Means περιστρέφεται γύρω από τη χρήση της ευκλείδειας απόστασης για την επαναληπτική αξιολόγηση και ενίσχυση της

ομοιογένειας εντός μιας συστάδας (P. Li et al., 2022). Αυτή η επαναληπτική διαδικασία μπορεί να περιγράφει σε διάφορα βασικά βήματα:

I. Επιλογή του αριθμού των συστάδων k,
II. Τυχαία επιλογή k κεντροειδών, που αντιπροσωπεύουν τα αρχικά κέντρα των συστάδων,
III. Ανάθεση κάθε σημείου δεδομένων στην πλησιέστερη συστάδα, μετά από υπολογισμούς ευκλείδειας απόστασης,
IV. Επανυπολογισμός του κεντροειδούς κάθε συστάδας, που καθορίζεται από τη μέση θέση όλων των σημείων που έχουν ανατεθεί στη συγκεκριμένη συστάδα,
V. Επαναπροσδιορισμός των σημείων των δεδομένων σε συστάδες, χρησιμοποιώντας τα ενημερωμένα πλέον κεντροειδή, και επανάληψη αυτού του βήματος έως ότου τα κεντροειδή σταθεροποιηθούν και δεν υπάρχουν περαιτέρω αλλαγές,
VI. Ολοκλήρωση της διαδικασίας ομαδοποίησης όταν τα κεντροειδή παραμένουν αμετάβλητα.

Συνεπώς, ο αλγόριθμος K-Means απαιτεί κάποιες συγκεκριμένες παραμέτρους και έναν προκαθορισμένο αριθμό συστάδων ως είσοδο. Λειτουργεί με την κατάτμηση του συνόλου δεδομένων σε έναν καθορισμένο αριθμό συστάδων, διασφαλίζοντας ότι η ομοιότητα εντός κάθε συστάδας μεγιστοποιείται. Ο αλγόριθμος λειτουργεί επαναληπτικά, υπολογίζοντας εκ νέου τις θέσεις των κεντροειδών πριν από κάθε επανάληψη. Κατά τη διάρκεια αυτών των επαναλήψεων, τα σημεία δεδομένων ανακατανέμονται σε διάφορες συστάδες με βάση τα νεοϋπολογισθέντα κεντροειδή. Αυτή η επανατοποθέτηση και ο επαναϋπολογισμός συνεχίζονται μέχρι να μην είναι δυνατή η περαιτέρω μείωση του αθροίσματος των αποστάσεων (Gustriansyah et al., 2019a).

Το K-Means προσπαθεί να ελαχιστοποιήσει το τετραγωνικό σφάλμα, το οποίο λειτουργεί ως μέτρο της απόστασης μεταξύ των σημείων δεδομένων και των αντίστοιχων κέντρων συστάδων τους. Ενώ ο αλγόριθμος φτάνει πάντα σε ένα σημείο τερματισμού, η βέλτιστη διαμόρφωση των συστάδων δεν είναι εγγυημένη. Η αποτελεσματικότητα του αλγορίθμου επηρεάζεται επίσης από την αρχική τυχαία επιλογή των κέντρων συστάδων, γεγονός που καθιστά αναγκαία την εκτέλεση πολλαπλών εκτελέσεων για τον μετριασμό αυτής της τυχαιότητας. Παρά την επαναληπτική του φύση, ο αλγόριθμος K-Means επιδεικνύει ισχυρή απόδοση με μεγάλα σύνολα δεδομένων. Για την αξιολόγηση της αποτελεσματικότητας της

συσταδοποίησης χρησιμοποιούνται διάφοροι δείκτες εγκυρότητας συστάδων, όπως οι Dunn, Davies Bouldin και Silhouette, καθώς και η μετρική Sum of Squares within Cluster (SSWC) (Güçdemir & Selim, 2015). Ειδικότερα, οι χαμηλότερες τιμές SSWC συσχετίζονται με καλύτερα αποτελέσματα ομαδοποίησης (Güçdemir & Selim, 2015).

Στο παρελθόν μελέτες αναφέρονται στον αλγόριθμο Fuzzy C-Means, αυτός ο αλγόριθμος αποτελεί μια προηγμένη μεθοδολογία ομαδοποίησης που επιτρέπει σε ένα σημείο δεδομένων να είναι μέλος πολλών ομάδων ταυτόχρονα (Memon & Lee, 2017). Η μέθοδος αυτή έρχεται σε αντίθεση με τις παραδοσιακές προσεγγίσεις συσταδοποίησης, όπως η K-Means, όπου ένα σημείο δεδομένων κατατάσσεται αυστηρά σε μία μόνο συστάδα. Στον ασαφή C-Means, αντί να αποδίδει απόλυτη ένταξη ενός σημείου δεδομένων σε μια συγκεκριμένη συστάδα, ο αλγόριθμος υπολογίζει την πιθανότητα ή τον βαθμό στον οποίο ένα σημείο δεδομένων ανήκει σε κάθε συστάδα. Αυτό το χαρακτηριστικό του Fuzzy C-Means προσφέρει ένα αξιοσημείωτο πλεονέκτημα, ιδίως όταν πρόκειται για μεγάλα σύνολα δεδομένων που παρουσιάζουν ομοιότητες μεταξύ τους (Memon & Lee, 2017). Σε τέτοια σενάρια, ο Fuzzy C-Means τείνει να αποδίδει πιο διαφοροποιημένα και κατατοπιστικά αποτελέσματα σε σύγκριση με τον αλγόριθμο K-Means, όπου η συμμετοχή ενός σημείου δεδομένων περιορίζεται σε μία μόνο συστάδα. Η εφαρμογή του Fuzzy C-Means είναι επωφελής και σε μελέτες τμηματοποίησης πελατών. Για παράδειγμα, ένας πελάτης θα μπορούσε να συσχετιστεί με πολλαπλές συστάδες, επιτρέποντας έτσι μια πιο προσαρμοσμένη προσέγγιση στις στρατηγικές διατήρησης πελατών. Οι επιχειρήσεις μπορούν να επωφεληθούν από αυτό προσφέροντας ποικίλα κίνητρα ή υπηρεσίες προσαρμοσμένες σε κάθε τμήμα πελάτη (Christy et al., 2021).

Ο παραδοσιακός αλγόριθμος K-Means, αν και δημοφιλής για την ομαδοποίηση δεδομένων, δεν είναι απαλλαγμένος από τις αδυναμίες του (C. C. Liu et al., 2014). Ένα βασικό μειονέκτημα της μεθόδου K-Means είναι η εξάρτησή της από τυχαία επιλεγμένα αρχικά κεντροειδή. Αυτή η τυχαιότητα μπορεί να οδηγήσει σε πολύ στενή ομαδοποίηση των κεντροειδών, με αποτέλεσμα λιγότερο σημαντικές ομαδοποιήσεις (Christy et al., 2021). Η επιλογή των αρχικών κεντροειδών είναι ζωτικής σημασίας, καθώς επηρεάζει διάφορες πτυχές της διαδικασίας ομαδοποίησης, συμπεριλαμβανομένης της μείωσης του αριθμού των απαιτούμενων επαναλήψεων, της επίτευξης συνολικά βέλτιστων λύσεων και της συμπαγούς μορφής των συστάδων. Η αποτελεσματικότητα της μεθόδου K-Means μπορεί να υπονομευθεί σημαντικά από αυτά τα τυχαία επιλεγμένα αρχικά κεντροειδή.

Για την αντιμετώπιση αυτών των προκλήσεων, μια μελέτη παρουσίασε μια νέα προσέγγιση για την επιλογή των αρχικών κεντροειδών στον αλγόριθμο K-Means (Christy et al., 2021). Η μέθοδος αυτή περιλαμβάνει την ταξινόμηση των τριών μεταβλητών - Recency (R), Frequency (F) και Monetary (M) - σε αύξουσα σειρά και την τοποθέτησή τους σε τρία ξεχωριστά διανύσματα, τα οποία συμβολίζονται ως R', F' και M'. Στη συνέχεια, προσδιορίζεται η διάμεση τιμή κάθε διανύσματος και ορίζεται ως το αρχικό κεντροειδές για τον αλγόριθμο K-Means (Christy et al., 2021). Αυτή η διαδικασία επαναλαμβάνεται επαναληπτικά, υπολογίζοντας εκ νέου τις διάμεσες τιμές από τα R', F' και M' συνολικά k φορές, που αντιστοιχούν στον αριθμό των τμημάτων (k). Αυτή η μέθοδος επιλογής αρχικών κεντροειδών με βάση τη μέση κατανομή τους έχει αποδειχθεί ότι μειώνει τόσο τον αριθμό των επαναλήψεων όσο και τον υπολογιστικό χρόνο που απαιτεί ο παραδοσιακός αλγόριθμος K-Means. Οι συστάδες που προκύπτουν από αυτή την τροποποιημένη προσέγγιση, η οποία μπορεί να ονομαστεί Repetitive Median K-Means (RM K-Means), παρατηρείται ότι είναι πιο ουσιαστικές και κατάλληλες σε σύγκριση με εκείνες που παράγονται με τυχαία επιλογή κεντροειδών. Παρά τις βελτιώσεις αυτές, η υπολογιστική πολυπλοκότητα του RM K-Means παραμένει η ίδια με εκείνη του παραδοσιακού K-Means, η οποία είναι O(n + k + i). Ωστόσο, το πλεονέκτημα του RM K-Means έγκειται στην ικανότητά του να μειώνει τον αριθμό των επαναλήψεων που συνήθως απαιτούνται στην τυπική διαδικασία K-Means, χάρη στη μέθοδο που βασίζεται στη διάμεσο για τον υπολογισμό των αρχικών κεντροειδών. Αυτή η βελτίωση καθιστά τον RM K-Means μια πιο αποδοτική και αποτελεσματική εναλλακτική λύση για εργασίες ομαδοποίησης δεδομένων (Christy et al., 2021).

Κατά τον συνδυασμό του K-Means με την ανάλυση RFM, ένα κρίσιμο βήμα είναι η κανονικοποίηση των μεταβλητών - επαναληπτικότητα, συχνότητα και χρηματική αξία - χρησιμοποιώντας την τεχνική κανονικοποίησης min-max (Alamsyah et al., 2022). Αυτή η κανονικοποίηση είναι ζωτικής σημασίας, καθώς αντιμετωπίζει το ζήτημα των στρεβλών τιμών των δεδομένων, οι οποίες θα μπορούσαν δυνητικά να εμποδίσουν τη διαδικασία ομαδοποίησης (Hsin Hung Wu et al., 2009). Στη συνέχεια, ο αλγόριθμος εφαρμόζεται στα κανονικοποιημένα δεδομένα. Ο αριθμός των συστάδων περιορίζεται σε δέκα κατ' ανώτατο όριο. Μια βασική εφαρμογή αυτής της διαδικασίας είναι ο προσδιορισμός των τμημάτων πελατών που παράγουν τα περισσότερα έσοδα για μια εταιρεία, αναλύοντας τη χρηματική αξία που σχετίζεται με κάθε συστάδα. Η υπολογιστική πολυπλοκότητα του αλγορίθμου K-Means εκφράζεται ως O(n + k + i),

όπου το "n" αντιπροσωπεύει τον αριθμό των περιπτώσεων δεδομένων, το "k" είναι ο αριθμός των συστάδων και το "i" υποδηλώνει τον αριθμό των επαναλήψεων που απαιτούνται για τη σύγκλιση του αλγορίθμου. Αυτός ο συμβολισμός πολυπλοκότητας παρέχει μια κατανόηση του τρόπου με τον οποίο ο αλγόριθμος κλιμακώνεται με το μέγεθος του συνόλου δεδομένων, τον αριθμό των συστάδων και την επαναληπτική διαδικασία (Christy et al., 2021).

Όπως αναφέρθηκε παραπάνω, μια πρόκληση – ίσως η πιο καθοριστική για την επιτυχία της εφαρμογής της μεθόδου K-Means είναι ο προσδιορισμός του καταλληλότερου αριθμού συστάδων, k. Μελέτες, όπως αυτές που διεξήχθησαν (Subbalakshmi et al., 2015), έχουν δείξει ότι η ακρίβεια της συσταδοποίησης K-Means μπορεί να βελτιωθεί σημαντικά με την προσεκτική επιλογή των αρχικών κεντροειδών και του αριθμού των συστάδων (Hsin Hung Wu et al., 2009). Υπάρχουν διάφορες μεθοδολογίες για την εκτίμηση του βέλτιστου αριθμού συστάδων k, όπως η μέθοδος Elbow, ο δείκτης Silhouette, ο δείκτης Calinski-Harabasz, ο δείκτης Davies-Bouldin, ο δείκτης Ratkowski, ο δείκτης Hubert, ο δείκτης Ball-Hall και ο δείκτης Krzanowski-Lai.

Η μέθοδος Elbow (EM) είναι μια τεχνική για τον προσδιορισμό του ιδανικού αριθμού συστάδων σε ένα σύνολο δεδομένων (Liu et al., 2014). Λειτουργεί εξετάζοντας το πώς η προσθήκη κάθε συστάδας επηρεάζει τη συνολική διακύμανση, η οποία αναπαρίσταται μέσω μιας καμπύλης (Gustriansyah et al., 2019). Μια σημαντική πτώση της διακύμανσης, που σχηματίζει έναν "αγκώνα" ή μια οξεία γωνία σε αυτή την καμπύλη, υποδεικνύει τον καταλληλότερο αριθμό συστάδων. Αυτό το βέλτιστο σημείο είναι εκεί όπου βρίσκεται το άθροισμα τετραγώνων εντός των συστάδων (Within-Clusters Sum of Squares - WSS), το οποίο μετρά τη συνολική διακύμανση εντός των συστάδων και παρουσιάζει τη μεγαλύτερη μείωση ως συνάρτηση του αριθμού των συστάδων. Καθώς προστίθενται περισσότερες συστάδες, το WSS γενικά μειώνεται, αλλά ο στόχος είναι να βρεθεί το σημείο της αύξησης του αριθμού των συστάδων, όπου δεν οδηγεί πλέον σε σημαντική μείωση της διακύμανσης. Επίσης πολύ σημαντικός δείκτης είναι ο δείκτης σιλουέτας (SI), μια μετρική για την αξιολόγηση της ποιότητας ενός συγκεκριμένου σχηματισμού συστάδων (Luna-Romera et al., 2016). Η μέθοδος αυτή εισήχθη από τον Rousseeuw και υπολογίζει τη μέγιστη τιμή του δείκτη για τον προσδιορισμό της προσαρμογής των συστάδων (Rousseeuw, 1987). Ο δείκτης σιλουέτας αξιολογεί τόσο τη συνοχή εντός των συστάδων όσο και τον διαχωρισμό

μεταξύ τους, παρέχοντας ένα μέτρο του πόσο κατάλληλα έχουν ομαδοποιηθεί τα σημεία δεδομένων σε συστάδες.

Όπως είναι φανερό μία από τις κύριες προκλήσεις του αλγορίθμου K-means σε κάθε εφαρμογή έγκειται στον προσδιορισμό του βέλτιστου αριθμού συστάδων, K. Ενώ ο K-Means υπερέχει σε υπολογιστική ταχύτητα σε σύγκριση με τις ιεραρχικές μεθόδους συσταδοποίησης (Badase et al., 2015), η εξάρτησή του από τα αρχικά κέντρα συστάδων μπορεί να οδηγήσει σε ανακριβή αποτελέσματα εάν τα κέντρα αυτά δεν επιλεγούν κατάλληλα (Xiong et al., 2016). Η χρονική πολυπλοκότητα της μεθόδου K-Means είναι O(nkl), όπου n είναι ο αριθμός των σημείων δεδομένων, k είναι ο αριθμός των συστάδων και l είναι ο αριθμός των επαναλήψεων που απαιτούνται για τη σύγκλιση (O. A. Abbas, 2008).

Συμπερασματικά, ο αλγόριθμος K-Means ξεχωρίζει ως ένας εξαιρετικά αποδοτικός και δημοφιλής αλγόριθμος ομαδοποίησης, ο οποίος χρησιμοποιείται εκτενώς σε διάφορους τομείς, ιδίως στην τμηματοποίηση πελατών και στην ανάλυση λιανικής πώλησης (Gustriansyah et al., 2019b). Η απλότητά του, σε συνδυασμό με την ικανότητά του να χωρίζει αποτελεσματικά μεγάλα σύνολα δεδομένων σε διακριτές συστάδες, τον καθιστά πολύτιμο εργαλείο στην ανάλυση δεδομένων (P. Li et al., 2022). Ο επαναληπτικός χαρακτήρας του αλγορίθμου και η εξάρτηση από τη μέθοδο του αγκώνα για τον προσδιορισμό του βέλτιστου αριθμού συστάδων υπογραμμίζουν περαιτέρω την προσαρμοστική αλλά μεθοδική προσέγγισή του. Παρά την ευαισθησία του στις αρχικές συνθήκες και την έλλειψη εγγυημένης βέλτιστης διαμόρφωσης, ο αλγόριθμος K-Means υπερέχει σε υπολογιστική ταχύτητα σε σύγκριση με τις ιεραρχικές μεθόδους και η ευρεία εφαρμογή του σε τομείς όπως το λιανικό εμπόριο και η ανάλυση πελατών υπογραμμίζει την πρακτική του χρησιμότητα. Οι επιδόσεις του αλγορίθμου, ιδίως όταν πρόκειται για μεγάλα, καλά κλιμακωμένα σύνολα δεδομένων, παραμένουν ισχυρές, καθιστώντας τον μια τεχνική ακρογωνιαίο λίθο στη σφαίρα των αλγορίθμων ομαδοποίησης (Y. Li et al., 2021).

2.3 ΠΙΣΤΟΤΗΤΑ ΠΕΛΑΤΩΝ (customer loyalty)

Η αφοσίωση χαρακτηρίζεται ως μια ισχυρή τάση να συνεχίζει κανείς να αγοράζει ή να χρησιμοποιεί ένα προτιμώμενο προϊόν ή υπηρεσία, ακόμη και όταν οι περιστάσεις θα μπορούσαν να προτείνουν τη μετάβαση σε μια άλλη μάρκα (Oliver, 1999). Αυτή η έννοια της αφοσίωσης των πελατών περιλαμβάνει διάφορες πτυχές,

συμπεριλαμβανομένων των γνωστικών, συναισθηματικών και συμπεριφορικών διαστάσεων (Dapena-Baron et al., 2020). Η συμπεριφορική διάσταση αφορά τη συνεπή αγορά αγαθών ή υπηρεσιών από την ίδια μάρκα ή εταιρεία. Αντίθετα, η συμπεριφορική αφοσίωση είναι εμφανής όταν οι πελάτες αντιστέκονται στην παρόρμηση να στραφούν σε άλλη μάρκα, με αποτέλεσμα τη συνεχή κατανάλωση ενός συγκεκριμένου προϊόντος ή μάρκας από την ίδια εταιρεία (Oliver, 1999). Η συμπεριφορική αφοσίωση συνδέεται επίσης με την προθυμία να πληρώσουν υψηλή τιμή (Zeithaml et al., 1996) και την επιθυμία να συστήσουν τη μάρκα σε άλλους.

Η αφοσίωση των πελατών μπορεί να αποφέρει πολλά πλεονεκτήματα για τις επιχειρήσεις, όπως η δημιουργία μιας σταθερής πελατειακής βάσης, η οποία μειώνει το κόστος συναλλαγών και απόκτησης, ενώ παράλληλα μετριάζει τη μεταβλητότητα των κερδών (Srivastava et al., 1999). Αποφέρει επίσης πρόσθετα οφέλη, όπως μειωμένες δαπάνες μάρκετινγκ, βελτιωμένη διατήρηση των πελατών, μεγαλύτερο μερίδιο αγοράς και προθυμία των πελατών να πληρώσουν περισσότερα (Rauyruen et al., 2009). Κατά συνέπεια, η αφοσίωση των πελατών είναι ζωτικής σημασίας για την επίτευξη μακροπρόθεσμης ανταγωνιστικής υπεροχής έναντι των ανταγωνιστών και χρησιμεύει ως πρωταρχικός στόχος στον τομέα του μάρκετινγκ. Επιπλέον, οι εταιρείες που υποστηρίζουν κοινωνικούς σκοπούς τείνουν να απολαμβάνουν μεγαλύτερη εύνοια μεταξύ των καταναλωτών (Sharma & Jain, 2019). Το βιώσιμο μάρκετινγκ είναι μια πολύτιμη στρατηγική για την προώθηση της αφοσίωσης των πελατών, καθώς περιστρέφεται γύρω από τη δημιουργία και τη διατήρηση διαρκών πελατειακών σχέσεων (Almeida & Coelho, 2019).

Τα εμπειρικά στοιχεία των μελετών υπογραμμίζουν τη θετική επίδραση της ευαισθητοποίησης των πελατών σχετικά με τις πρακτικές περιβαλλοντικής βιωσιμότητας στις προθέσεις συμπεριφοράς τους. Υποστηρίζεται, επίσης, ότι οι πρακτικές βιώσιμου μάρκετινγκ επηρεάζουν θετικά τη συμπεριφορική αφοσίωση, επηρεάζοντας στη συνέχεια τη συμπεριφορική αφοσίωση των πελατών (Khandai et al., 2023). Καταδεικνύεται η θετική σχέση μεταξύ της περιβαλλοντικής διάστασης, της αειφορίας και της καταναλωτικής αφοσίωσης (Han et al., 2019). Ομοίως, επικυρώνεται ότι η κοινωνική διάσταση της βιωσιμότητας επηρεάζει θετικά την καταναλωτική αφοσίωση, με τη διαμεσολάβηση της φήμης (Aramburu & Pescador, 2019). Πολυάριθμες μελέτες στον τομέα των καταναλωτικών ηλεκτρονικών ειδών παρέχουν περαιτέρω επιβεβαίωση της θετικής σχέσης μεταξύ βιώσιμων πρακτικών και καταναλωτικής πίστης (Edeh et al., 2021).

Είναι πολλές οι έρευνες που έχουν ασχοληθεί με τους παράγοντες που επηρεάζουν την ικανοποίηση και την αφοσίωση των πελατών σε διάφορα συστήματα ηλεκτρονικού εμπορίου, όπως ηλεκτρονικά καταστήματα (Wu et al., 2018), (F. Liu et al., 2020), ειδικές πλατφόρμες ηλεκτρονικού εμπορίου (Gajewska et al., 2019) και συστήματα ηλεκτρονικών αγορών (Childers et al., 2001), (Rita et al., 2019). Η μετάβαση στις online υπηρεσίες προήλθε κυρίως από χρηστικούς στόχους που αποσκοπούσαν στην ικανοποίηση συγκεκριμένων αναγκών των πελατών και στην επίτευξη των επιθυμητών αποτελεσμάτων αποδοτικά (W.-K. Liu et al., 2017). Μια ευρέως χρησιμοποιούμενη μέθοδος για τη διερεύνηση του τρόπου με τον οποίο οι πελάτες καθορίζουν και αξιολογούν την εκπλήρωση αυτών των χρηστικών στόχων με βάση τις προϋπάρχουσες προσδοκίες τους και τις εμπειρίες τους μετά την κατανάλωση, είναι η θεωρία επιβεβαίωσης προσδοκιών (Expectation Confirmation Theory - ECT) (Bhattacherjee, 2001). Η θεωρία αυτή υπογραμμίζει το ρόλο των προσδοκιών στη διαμόρφωση των κρίσεων ικανοποίησης των πελατών και της συνέχισης της χρήσης. Κατά συνέπεια, η υποκειμενική αξιολόγηση της ποιότητας του συστήματος από τους πελάτες έχει καθοριστικό ρόλο στη διαμόρφωση της συμπεριφοράς τους μετά την κατανάλωση. Όταν ένα σύστημα ηλεκτρονικών υπηρεσιών ξεπερνά τις αρχικές προσδοκίες του πελάτη, δημιουργεί θετική επιβεβαίωση, ενισχύοντας στη συνέχεια την ικανοποίηση μετά την υιοθέτηση (Hossain & Quaddus, 2012). Αυτή η υποκειμενική αξιολόγηση της αντιλαμβανόμενης χρησιμότητας είναι στενά συνυφασμένη με τους ωφελιμιστικούς στόχους των πελατών.

Προηγούμενες μελέτες έχουν χρησιμοποιήσει την έννοια της αντιλαμβανόμενης χρησιμότητας στο πλαίσιο της ECT για να διερευνήσουν τη συμπεριφορά των πελατών και των χρηστών μετά την υιοθέτηση συστημάτων ηλεκτρονικού εμπορίου (Zhou et al., 2019). Δεδομένης της υψηλής ανταγωνιστικότητας στον κλάδο, οι εταιρείες αυτές πρέπει να παρέχουν εξαιρετικές υπηρεσίες για να προσελκύσουν και να διατηρήσουν πελάτες, καλλιεργώντας μακροχρόνιες σχέσεις. Η ενσωμάτωση της αντιλαμβανόμενης χρησιμότητας σε αυτά τα συστήματα έχει εξελιχθεί ώστε να περιλαμβάνει ένα ευρύτερο φάσμα χαρακτηριστικών ποιότητας υπηρεσιών (Askariazad & Babakhani, 2015). Η ποιότητα των ηλεκτρονικών υπηρεσιών αφορά το βαθμό στον οποίο το ηλεκτρονικό εμπόριο διευκολύνει τις αποδοτικές και αποτελεσματικές ηλεκτρονικές συναλλαγές, που περιλαμβάνουν αγορές, πληρωμές και παράδοση (Parasuraman et al., 2005). Σε μια μελέτη που εξέτασε τον δείκτη ικανοποίησης των ηλεκτρονικών πελατών (e-CSI), εισήγαγαν πέντε χαρακτηριστικά ποιότητας ηλεκτρονικών υπηρεσιών:

διαθεσιμότητα και περιεχόμενο πληροφοριών, ευκολία χρήσης, ιδιωτικότητα/ασφάλεια, σχεδιασμός γραφικών και εκπλήρωση/αξιοπιστία (Hsu, 2008). Η μελέτη αυτή ανακάλυψε, επίσης, ότι η προτεινόμενη ποιότητα ηλεκτρονικών υπηρεσιών επηρέασε θετικά την αφοσίωση μέσω της διαμεσολάβησης της ικανοποίησης των πελατών. Σε μια άλλη μελέτη έγινε πρόταση ενός μοντέλου ποιότητας ηλεκτρονικών υπηρεσιών τρίτης τάξης που περιλαμβάνει τέσσερις διαστάσεις: σχεδιασμός ιστοτόπου, εκπλήρωση, εξυπηρέτηση πελατών και ασφάλεια/ιδιωτικότητα (Blut, 2016). Οι Sheu & Chang, 2022 οριοθέτησαν την ποιότητα υπηρεσιών σε τέσσερις διαστάσεις: εκπλήρωση, αποτελεσματικότητα, διαθεσιμότητα του συστήματος και ιδιωτικότητα. Τα ευρήματά τους έδειξαν ότι όλες οι διαστάσεις της ποιότητας υπηρεσιών είχαν θετική συσχέτιση με την ικανοποίηση των πελατών, με την εκπλήρωση, την αποτελεσματικότητα και την ιδιωτικότητα να επηρεάζουν σημαντικά την αφοσίωση. Μια άλλη μελέτη, η οποία εξέταζε τον ρόλο της ποιότητας ηλεκτρονικών υπηρεσιών στην αφοσίωση των διαδικτυακών πελατών, εισήγαγε δύο διαστάσεις: την ασφάλεια των πληροφοριών και την απόδοση του δικτυακού τόπου αγορών (Shafiee & Bazargan, 2018). Η έρευνα έδειξε θετική συσχέτιση μεταξύ όλων των διαστάσεων και της ποιότητας ηλεκτρονικών υπηρεσιών, με την ποιότητα ηλεκτρονικών υπηρεσιών να συνδέεται θετικά με την ηλεκτρονική αφοσίωση.

Η θεωρία της επιβεβαίωσης των προσδοκιών (ECT) υποστηρίζει ότι η επιβεβαίωση των προσδοκιών αξιολογεί την ευθυγράμμιση μεταξύ των προσδοκιών των χρηστών των συστημάτων πληροφορικής (IT) και της πραγματικής χρήσης τους, γεγονός που συνεπάγεται την υλοποίηση των αναμενόμενων οφελών από τη χρήση του συστήματος IT. Η θετική επιβεβαίωση συμβαίνει όταν οι πραγματικές εμπειρίες ευθυγραμμίζονται με τις αρχικές προσδοκίες των χρηστών ή τις ξεπερνούν (Oghuma et al., 2016). Στη συνέχεια, η επίτευξη ή η υπέρβαση της υλοποίησης της αντιλαμβανόμενης ποιότητας υπηρεσιών μέσω της πραγματικής χρήσης συμβάλλει και ενισχύει την ικανοποίηση των χρηστών (Hsiao, 2018). Επιπλέον, η θετική επιβεβαίωση όχι μόνο προάγει την ικανοποίηση επικυρώνοντας τις αρχικές προσδοκίες των χρηστών, αλλά δημιουργεί, επίσης, αίσθημα εκπλήρωσης και γενικά, θετικά συναισθήματα. Προηγούμενες έρευνες σε συστήματα ηλεκτρονικού εμπορίου έχουν επιβεβαιώσει την ευνοϊκή επίδραση της επιβεβαίωσης των προσδοκιών στην ικανοποίηση των χρηστών (Tam et al., 2020). Η χρήση των συστημάτων ηλεκτρονικού εμπορίου, συμπεριλαμβανομένων των ηλεκτρονικών αγορών, γίνεται συνήθης για

πολλούς χρήστες, έτσι, οι προσδοκίες τους επεκτείνονται πέρα από τους χρηστικούς στόχους, προσφέροντας και ηδονικές εμπειρίες (Marinkovic & Kalinic, 2017). Μια βασική πτυχή των ηδονικών εμπειριών στις δραστηριότητες αγορών είναι η αντίληψη των χρηστών για την απόλαυση. Προηγούμενες μελέτες έχουν καταδείξει μια θετική σχέση μεταξύ της επιβεβαίωσης των προσδοκιών των χρηστών και της αντιλαμβανόμενης απόλαυσης (Oghuma et al., 2016). Η επίτευξη της αντιλαμβανόμενης απόλαυσης στις δραστηριότητες αγορών οδηγεί σε θετικές συναισθηματικές συσχετίσεις, αυξημένη δέσμευση και αυξημένη ικανοποίηση των πελατών (Marinkovic & Kalinic, 2017), (Lee et al., 2019), συμβάλλοντας τελικά στη μακροπρόθεσμη αφοσίωση.

Σε ένα έντονα ανταγωνιστικό τοπίο ηλεκτρονικού εμπορίου, οι πελάτες έχουν το πάνω χέρι λόγω της πληθώρας των επιλογών που έχουν στη διάθεσή τους. Το επίπεδο ικανοποίησης από την αρχική αγορά παίζει καθοριστικό ρόλο στον καθορισμό των γεγονότων μετά την αγορά, όπως οι αλλαγές συμπεριφοράς, η δέσμευση και η αφοσίωση. Καθώς τα επίπεδα ικανοποίησης αυξάνονται, οι πελάτες είναι πιο πιθανό να επιδείξουν αφοσίωση, να επαναγοράσουν, να συστήσουν προϊόντα ή υπηρεσίες και να τα υποστηρίξουν ακούσια σε άλλους (Bowen & Mccain, 2015), (Boonlertvanich, 2019). Ωστόσο, στον τομέα του ηλεκτρονικού εμπορίου, η ικανοποίηση από μόνη της δεν εγγυάται επαναλαμβανόμενες συναλλαγές ή συνεχή χρήση. Ο ψηφιακός χαρακτήρας των συστημάτων ηλεκτρονικού εμπορίου δίνει τη δυνατότητα στους πελάτες να συγκρίνουν τιμές, προσφορές προϊόντων και εμπειρίες σε διάφορες πλατφόρμες με τα χέρια τους (Puspitasari et al., 2023). Αυτή η διαφάνεια επιτρέπει στους πελάτες να λαμβάνουν τεκμηριωμένες αποφάσεις και να μεταπηδούν σε πλατφόρμες που προσφέρουν καλύτερες προσφορές ή ανώτερες εμπειρίες. Μια κοινή στρατηγική για τη διατήρηση των πελατών είναι η προσφορά εξατομικευμένων υπηρεσιών σε διάφορα σημεία επαφής, ενισχύοντας την επίτευξη τόσο χρηστικών όσο και ηδονικών στόχων, καλλιεργώντας την αφοσίωση στην μάρκα και τελικά καλλιεργώντας μακροχρόνιες σχέσεις (Kasiri et al., 2017), (Pappas et al., 2017).

Η ικανοποίηση και η αφοσίωση των πελατών αναγνωρίζονται ευρέως ως κρίσιμοι δείκτες μακροπρόθεσμης επιχειρηματικής επιτυχίας. Πολλές επιχειρήσεις θεωρούν την αφοσίωση των πελατών κρίσιμο περιουσιακό στοιχείο για τη δημιουργία κερδών (Guimaraes & Paranjape, 2014). Οι μελετητές υποστηρίζουν ότι η υψηλή ικανοποίηση των πελατών οδηγεί σε υψηλή αφοσίωση των πελατών (Kim et al., 2004). Είναι ευρέως αποδεκτό ότι η ικανοποίηση των πελατών χρησιμεύει ως ζωτικός

πρόδρομος για την αφοσίωση των πελατών (Morshedlou & Meybodi, 2014). Μέχρι και τα τελευταία χρόνια, δεν υπάρχει καθιερωμένο ακαδημαϊκό πλαίσιο για τον προσδιορισμό των παραγόντων που επηρεάζουν την αφοσίωση των πελατών (Kandampully & Suhartanto, 2000). Παρ' όλα αυτά, υπάρχει συναίνεση ότι τόσο η ποιότητα των υπηρεσιών όσο και η ικανοποίηση των πελατών αποτελούν προϋποθέσεις για την προώθηση της αφοσίωσης των πελατών (Cronin & Taylor, 1992). Οι εμπειρικές παρατηρήσεις μελετών έδειξαν ότι οι πιο ικανοποιημένοι πελάτες είναι πιο πιθανό να επιδείξουν αφοσίωση (Helgesen, 2006), (Rahim, 2016), (van Lierop & El-Geneidy, 2016).

2.4 ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ (sentiment analysis)

Οι υπάρχουσες έρευνες έχουν δείξει ότι τα τυπικά χαρακτηριστικά των διαδικτυακών κριτικών, όπως ο αριθμός και η διάρκεια των κριτικών, επηρεάζουν σημαντικά την αγοραστική συμπεριφορά των καταναλωτών ή τις πωλήσεις προϊόντων (Delre & Luffarelli, 2023). Στον τομέα της φιλοξενίας (Mauri & Minazzi, 2013), ερευνητές εντόπισαν θετική συσχέτιση μεταξύ των προθέσεων αγοράς των καταναλωτών και του τόνου των κριτικών μέσω μιας πειραματικής μελέτης. Σε άλλη μελέτη εξέτασαν πώς ο όγκος των διαδικτυακών κριτικών των πελατών μεσολαβεί στην επίδραση των κριτικών των εμπειρογνωμόνων στις αποφάσεις των καταναλωτών (Luo et al., 2022). Σε μια έρευνα του 2016 ανακάλυψαν ότι τόσο ο όγκος όσο και η μεταβλητότητα των κριτικών επηρεάζουν τις αποφάσεις αγοράς σε ένα διαδικτυακό περιβάλλον λιανικής πώλησης (Minnema et al., 2016). Παρομοίως, σε άλλη έρευνα ανέλυσαν δεδομένα από το Τμήμα Εσόδων της Αϊόβα και διαπίστωσαν θετική σχέση μεταξύ του όγκου των κριτικών και των κερδών των εστιατορίων (L. Wang et al., 2021). Επιπλέον, έχει γίνει χρήση ενός μοντέλου ταυτόχρονων εξισώσεων για να δείξουν τις φθίνουσες προληπτικές επιδράσεις του αθροιστικού όγκου eWOM στις πωλήσεις εισιτηρίων ταινιών κατά τη διάρκεια του κύκλου ζωής της ταινίας (Delre & Luffarelli, 2023).

Άλλες μελέτες έχουν εξετάσει τις πτυχές του περιεχομένου των διαδικτυακών αξιολογήσεων πελατών, όπως η πυκνότητά τους, το επίπεδο εμπειρογνωμοσύνης και η σαφήνειά τους, καθώς και την επιρροή τους στις αποφάσεις αγοράς των καταναλωτών ή στις πωλήσεις προϊόντων (Cai et al., 2023). Σε μια μελέτη κατέληξαν στο συμπέρασμα ότι η αντιλαμβανόμενη χρησιμότητα, ο επαγγελματισμός, η επικαιρότητα

και η πληρότητα των διαδικτυακών κριτικών πελατών επηρεάζουν θετικά τις κρατήσεις ξενοδοχείων, βάσει μελέτης ερωτηματολογίου (Zhao et al., 2015). Οι Kaushik κ.ά. κατέδειξαν την επίδραση των συνοπτικών στατιστικών στοιχείων των κριτικών, του όγκου, του περιεχομένου και της αλληλουχίας των κριτικών στις πωλήσεις προϊόντων, χρησιμοποιώντας μια ολοκληρωμένη προσέγγιση (Kaushik et al., 2018). Σε άλλη ανάλυση χρησιμοποίησαν την μέθοδο PLS-SEM για να καταδείξουν την ευεργετική επίδραση της αντιλαμβανόμενης χρησιμότητας των διαδικτυακών κριτικών πελατών στην ικανοποίηση των πελατών, χρησιμοποιώντας δεδομένα του TripAdvisor (Filieri et al., 2021). Ενώ, τέλος, σε μια πρόσφατη μελέτη διαπίστωσαν ότι η αναγνωσιμότητα των κριτικών επηρεάζει άμεσα τις πωλήσεις προϊόντων, με την πληροφορία των κριτικών να παίζει μετριαστικό ρόλο (Cai et al., 2023).

Μέχρι τώρα η έρευνα έχει επικεντρωθεί στη σχέση μεταξύ του συναισθηματικού τόνου των διαδικτυακών κριτικών πελατών και της αγοραστικής συμπεριφοράς ή των πωλήσεων προϊόντων (Guo et al., 2020). Ο Ludwig χρησιμοποίησε εξόρυξη κειμένου για να παρακολουθήσουν τις αλλαγές στο συναισθηματικό περιεχόμενο των κριτικών της Amazon.com, διαπιστώνοντας ότι η μεγαλύτερη αύξηση του θετικού συναισθηματικού περιεχομένου των κριτικών είχε μικρότερο αντίκτυπο στα ποσοστά μετατροπής για τους διαδικτυακούς ιστότοπους λιανικής πώλησης (Ludwig et al., 2013). Σε άλλη έρευνα πρότειναν ότι η απόλαυση που προέρχεται από τις διαδικτυακές κριτικές πελατών επηρεάζει σημαντικά την εμπιστοσύνη των πελατών και την προθυμία τους να προβούν σε ψηφιακές αγορές, βάσει συζητήσεων ομάδων εστίασης και ερευνών (Elwalda et al., 2016). Οι Guo κ.ά. διεξήγαγαν και εργαστηριακά πειράματα για να διερευνήσουν την επίδραση των συναισθημάτων των πελατών στις διαδικτυακές κριτικές, αποκαλύπτοντας ότι οι θετικές κριτικές αύξησαν την πιθανότητα αγορών περισσότερο από τις αρνητικές (Guo et al., 2020).

Η υπάρχουσα βιβλιογραφία έχει επικεντρωθεί στον αντίκτυπο μεμονωμένων, ευρέων χαρακτηριστικών των διαδικτυακών κριτικών πελατών σε πλαίσια όπως οι πωλήσεις διαδικτυακών κρατήσεων ή η αγορά λογισμικού. Στον τομέα των ανακατασκευασμένων προϊόντων, η έρευνα έχει εξετάσει κυρίως τις επιδράσεις των αντιλήψεων για την ποιότητα (Abbey et al., 2015), των στρατηγικών τιμολόγησης (Frota Neto et al., 2016), των παραγόντων της αγοράς (Van Nguyen et al., 2020) και των πολιτικών φορολογίας άνθρακα (Luo et al., 2022) στη συμπεριφορά των καταναλωτών. Ωστόσο, υπάρχει κενό στην έρευνα που ποσοτικοποιεί τα λεπτομερή και πολύπλευρα χαρακτηριστικά των διαδικτυακών κριτικών των πελατών που

ενδέχεται να επηρεάζουν τις πωλήσεις των ανακατασκευασμένων προϊόντων. Ενώ οι περισσότερες μελέτες έχουν χρησιμοποιήσει ερωτηματολόγια ή έχουν συλλέξει δεδομένα από ανεξάρτητες βάσεις δεδομένων για την αξιολόγηση του περιεχομένου και των συναισθηματικών ιδιοτήτων των διαδικτυακών κριτικών πελατών, λιγότερες έχουν εφαρμόσει την επεξεργασία φυσικής γλώσσας (NLP) για την αυτόματη δημιουργία μεταβλητών για την ανάλυση αυτών των διαστάσεων και των υποκατηγοριών τους. Επιπλέον, αν και προηγούμενες έρευνες έχουν διερευνήσει τις τυπικές, περιεχομένου και συναισθηματικές ιδιότητες των διαδικτυακών κριτικών πελατών, οι ολοκληρωμένες μελέτες που ενσωματώνουν όλες αυτές τις πτυχές με πιο λεπτομερή τρόπο είναι περιορισμένες.

Η ανάλυση συναισθήματος, γνωστή και ως εξόρυξη γνώμης, διαδραματίζει κρίσιμο ρόλο στην αποκρυπτογράφηση του υποκειμενικού περιεχομένου ενός κειμένου (Pang & Lee, 2008). Οι προσεγγίσεις που βασίζονται σε λεξικά περιλαμβάνουν τη χρήση λεξικών συναισθημάτων με προκαθορισμένα συναισθήματα για τις λέξεις/φράσεις (B. Liu, 2012), ενώ οι τεχνικές μηχανικής μάθησης βασίζονται στην εκπαίδευση μοντέλων που χρησιμοποιούν επισημασμένα δεδομένα για την κατηγοριοποίηση συναισθημάτων (Turney, 2002). Οι μέθοδοι βαθιάς μάθησης (deep learning methods), όπως τα συνεπαγωγικά νευρωνικά δίκτυα (CNN) και τα επαναλαμβανόμενα νευρωνικά δίκτυα (RNN), έχουν δείξει ενθαρρυντικά αποτελέσματα σε αυτόν τον τομέα (Y. Wang et al., 2021). Τα συναισθήματα κατηγοριοποιούνται σε δύο τύπους σύμφωνα με τον (B. Liu, 2012):

- Τακτικές γνώμες (π.χ., "Η ποιότητα του ήχου του τηλεφώνου είναι εξαιρετική") ή έμμεσες γνώμες για ένα αντικείμενο με βάση τον αντίκτυπό του σε άλλα πράγματα (π.χ., "Αφού έκανα το εμβόλιο, άρχισα να έχω έντονους πονοκεφάλους")
- Συγκριτικές γνώμες, αυτές εκφράζουν μια σύγκριση μεταξύ δύο ή περισσότερων αντικειμένων, αντικατοπτρίζοντας την προτίμηση του κατόχου της γνώμης με βάση κοινά χαρακτηριστικά (π.χ. "Προτιμώ τον καφέ από το τσάι")

Η ανάλυση συναισθήματος λειτουργεί σε διαφορετικά επίπεδα:

- Ανάλυση σε επίπεδο πρότασης (θετική, αρνητική ή ουδέτερη) των προτάσεων, κάνοντας διάκριση μεταξύ πραγματικών (αντικειμενικών) και βασισμένων στη γνώμη (υποκειμενικών) προτάσεων (B. Liu, 2012)

- Ανάλυση σε επίπεδο εγγράφου, η οποία αξιολογεί αν ένα ολόκληρο έγγραφο μεταφέρει θετικό ή αρνητικό συναίσθημα. Αυτό το επίπεδο είναι κατάλληλο για έγγραφα που συζητούν ένα μόνο αντικείμενο, καθώς δεν μπορεί να χειριστεί πολλαπλά αντικείμενα που αξιολογούνται ή συγκρίνονται (B. Liu, 2012)
- Ανάλυση σε επίπεδο πτυχής (Aspect-Level Analysis), αυτό το επίπεδο επικεντρώνεται στον στόχο της γνώμης, κατανοώντας ότι μια γνώμη περιλαμβάνει ένα συναίσθημα και έναν στόχο. Η αναγνώριση του στόχου μιας γνώμης είναι ζωτικής σημασίας στην ανάλυση συναισθήματος (Hailu & Tilahun, 2014)

Παρόλο που η βιβλιογραφία για την ανάλυση συναισθήματος καλύπτει διάφορες μεθόδους και προσεγγίσεις, υπάρχει σημαντικό ερευνητικό κενό στην αξιολόγηση της ευρείας εφαρμογής των μοντέλων και της δυνατότητας μεταφοράς τους σε διάφορους τομείς. Η αξιολόγηση της ευελιξίας των μοντέλων συναισθήματος σε διάφορα γλωσσικά στυλ, πλαίσια και πηγές είναι ζωτικής σημασίας για πρακτικές, ευρείες εφαρμογές στην ανάλυση συναισθήματος (Zhai et al., 2024). Αυτός ο ανεξερεύνητος τομέας είναι κομβικός για τη μελλοντική έρευνα. Η ανάλυση συναισθήματος, ένας κλάδος της τεχνητής νοημοσύνης, διερευνά τις ανθρώπινες στάσεις, τα συναισθήματα, τις απόψεις, τις προοπτικές, τις εκτιμήσεις, τις αξιολογήσεις και τα συναισθήματα σχετικά με διάφορα θέματα, όπως προϊόντα, υπηρεσίες, οργανισμούς, άτομα, ζητήματα, γεγονότα και τις πτυχές τους (B. Liu, 2012). Τρία βασικά στοιχεία ορίζουν ένα συναίσθημα: ο κάτοχος του συναισθήματος (ο φορέας της γνώμης), το αντικείμενο του συναισθήματος (το υποκείμενο της γνώμης) και το ίδιο το συναίσθημα (η γνώμη ή το συναίσθημα που εκφράζεται). Τα συναισθήματα μπορεί να είναι θετικά, αρνητικά ή ουδέτερα και προσδιορίζονται είτε μέσω μη γραμμικής ταξινόμησης είτε μέσω μιας αριθμητικής βαθμολογίας που υποδεικνύει την ένταση του συναισθήματος (B. Liu, 2012).

Σε πολλές μελέτες διερευνήθηκαν οι παράγοντες που συμβάλλουν στην αφοσίωση του πελάτη ως αγοραστή. Μια έρευνα στην αγορά της Ινδονησίας (Puspitasari et al., 2023), μελέτησε το μοντέλο αφοσίωσης πελατών στην ηλεκτρονική αγορά, ενσωματώνοντας τις διαστάσεις του PSQ, τις δομές του μοντέλου προσδοκίας-επιβεβαίωσης, το μοντέλο χρήστη αποδοχή των ηδονικών συστημάτων πληροφοριών και την εξατομίκευση. Χρησιμοποιώντας δύο διαφορετικές προσεγγίσεις (Μερικά

ελαχίστων τετραγώνων Hierarchical Component Model και αλγόριθμοι ταξινόμησης μηχανικής μάθησης) στον έλεγχο υποθέσεων, ερεύνησαν την ανάπτυξη ενός ισχυρού μοντέλου, που παρέχει περαιτέρω στοιχεία για την υποστήριξη των αποτελεσμάτων των υποθέσεων. Δεδομένου ότι οι ηλεκτρονικές αγορές έχουν γίνει μια τακτική δραστηριότητα για τους περισσότερους πελάτες, οι στόχοι των πελατών για τη χρήση των ηλεκτρονικών αγορών έχουν επεκταθεί ώστε να συμπεριλάβουν στόχους με περισσότερο ηδονικό προσανατολισμό. Η εκπλήρωση των ηδονικών αξιών μέσω της αντιλαμβανόμενης απόλαυσης δημιουργεί μεγαλύτερο αντίκτυπο στην ικανοποίηση των πελατών (Puspitasari et al., 2023).

Έχουν διερευνηθεί πολλές τεχνικές ταξινόμησης συναισθημάτων. Τρεις βασικές που συναντώνται συχνά είναι machine learning, lexicon-based, and hybrid methods. Οι προσεγγίσεις μηχανικής μάθησης (ML) αξιοποιούν αλγορίθμους και γλωσσικά χαρακτηριστικά, ενώ οι μέθοδοι που βασίζονται σε «λεξικό» (lexicon-based) χρησιμοποιούν μια προσυγκεντρωμένη συλλογή φράσεων συναισθήματος, οι οποίες κατηγοριοποιούνται περαιτέρω σε προσεγγίσεις που βασίζονται σε λεξικό και σε σώματα κειμένων (Medhat et al., 2014). Η υβριδική μέθοδος συνδυάζει και τα δύο, με τα λεξικά συναισθήματος να είναι ζωτικής σημασίας για την ενίσχυση της απόδοσης (Medhat et al., 2014). Η ταξινόμηση συναισθήματος με μηχανική μάθηση διακρίνεται σε μεθόδους με επίβλεψη (supervised), όπου τα έγγραφα με ετικέτες είναι απαραίτητα, και σε μεθόδους χωρίς επίβλεψη (unsupervised methods) για τις περιπτώσεις που τα έγγραφα αυτά είναι σπάνια(Sisodiya & Mannepalli, 2021). Οι βασικοί ταξινομητές επιβλεπόμενης μάθησης περιλαμβάνουν, τον ταξινομητή Naïve Bayes (NB) (Medhat et al., 2014) (Sisodiya & Mannepalli, 2021), το Bayesian Network (BN), ο οποίος απομακρύνεται από την υπόθεση του NB προς ένα μοντέλο που αναπαριστά τις μεταβλητές και τις εξαρτήσεις τους μέσω ενός κατευθυνόμενου ακυκλικού γράφου, τον ταξινομητής μέγιστης εντροπίας (ME), που κωδικοποιεί σύνολα χαρακτηριστικών με ετικέτες σε διανύσματα για τον υπολογισμό των βαρών των χαρακτηριστικών, παράγοντας την πιο πιθανή «ετικέτα» (Medhat et al., 2014). Επιπλέον, οι γραμμικοί ταξινομητές, όπως οι Μηχανές Διανυσμάτων Υποστήριξης (SVM) και τα Νευρωνικά Δίκτυα (NN), στοχεύουν στην εύρεση των καλύτερων γραμμικών διαχωριστών για τη διαφοροποίηση των κλάσεων. Τα SVM λειτουργούν εντοπίζοντας το ευρύτερο περιθώριο μεταξύ των κλάσεων για βέλτιστο διαχωρισμό, ενώ τα NN λειτουργούν μέσω νευρώνων που επεξεργάζονται συχνότητες εισόδου για να προβλέψουν τις ετικέτες κλάσεων (Abate & Rashid, 2024).

Κεφάλαιο 3: Ερευνητική Προσέγγιση

3.1 ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ

3.1.1 Εισαγωγή στη Μεθοδολογία

Η επιλογή της μεθοδολογίας συμβαδίζει με το αντικείμενο μελέτης και την φύση των δεδομένων για την δημιουργία μιας δομημένης και κατανοητής ανάλυσης. Αυτό το τμήμα θα αναλύσει λεπτομερώς τα βήματα που ακολουθήθηκαν για την εξόρυξη των δεδομένων (data mining), αξιοποιώντας την RFM ανάλυση με τελικό σκοπό να εμβαθύνουμε στην μέτρηση της πιστότητας των πελατών (customer loyalty), ενώ ταυτόχρονα θα ενισχύσουμε τα συμπεράσματα αναφορικά με τη σχέση επιχείρησης και καταναλωτή, αναλύοντας και το συναίσθημα των πελατών (sentiment analysis) μέσα από τις κριτικές τους μετά την ολοκλήρωση της παραγγελίας τους. Η πηγή των δεδομένων που χρησιμοποιήθηκαν σε αυτή την μελέτη είναι μια από τις μεγαλύτερες πλατφόρμες ηλεκτρονικών αγορών στην Βραζιλία. Αυτό το ηλεκτρονικό κατάστημα είναι το νούμερο ένα οικοσύστημα εμπορικών εφαρμογών για μικρομεσαίες επιχειρήσεις (SMB) που ειδικεύεται στους τομείς των logistics και του κεφαλαίου στη Λατινική Αμερική.

Τα δεδομένα υποβλήθηκαν σε προ επεξεργασία για την απομάκρυνση τυχόν ακραίων τιμών, και στη συνέχεια συνοψίστηκαν τα κύρια χαρακτηριστικά του συνόλου δεδομένων (διαδικασία EDA), προκειμένου να αποκτηθεί μια βαθιά κατανόηση των δεδομένων, να εντοπιστούν μοτίβα, να εντοπιστούν ανωμαλίες και να δημιουργηθούν υποθέσεις. Ο αλγόριθμος K-means χρησιμοποιήθηκε για την ομαδοποίηση του συνόλου δεδομένων πελατών με βάση την ανάλυση RFM για την κατανόηση και την ομαδοποίηση των πελατών με βάση τη συμπεριφορά τους. Ταυτόχρονα, μελετήθηκε και η αξία διάρκειας ζωής πελάτης, μια μετρική εξίσου σημαντική, κατά τον υπολογισμό της οποίας χρησιμοποιήθηκε ξανά ο αλγόριθμος K-means. Τέλος, με την χρήση της python βιβλιοθήκης «nltk» οι κριτικές των πελατών επεξεργάστηκαν με βάση κάποιων σημαντικών NLP μεθόδων για την ασφαλή εξαγωγή συμπερασμάτων μετά την ανάλυση των δεδομένων κειμένου.

3.1.2 Το δείγμα

Το σύνολο δεδομένων παρέχει μια ολιστική εικόνα του ηλεκτρονικού εμπορίου, περιλαμβάνοντας λεπτομέρειες της παραγγελίας, όπως η κατάσταση, η τιμολόγηση, οι μέθοδοι πληρωμής και οι επιδόσεις αποστολής, εκτός από τα δημογραφικά στοιχεία των πελατών, τις ιδιαιτερότητες των προϊόντων και τα σχόλια των καταναλωτών.

Εικόνα 3.1 Σχεσιακή βάση δεδομένων ηλεκτρονικού καταστήματος

Το χρονικό διάστημα που καλύπτεται από το σύνολο δεδομένων είναι από το 2016 έως το 2018 σε διάφορες αγορές της Βραζιλίας. Για να ενισχυθεί η χρησιμότητά του, περιλαμβάνεται ένα στοιχείο γεωγραφικού εντοπισμού που αντιστοιχίζει τους ταχυδρομικούς κώδικες της Βραζιλίας σε γεωγραφικές συντεταγμένες. Το ηλεκτρονικό αυτό κατάστημα αποτελεί τον κορυφαίο όμιλο στον τομέα των αγορών της Βραζιλίας, διευκολύνοντας την απρόσκοπτη ενσωμάτωση των μικρών επιχειρήσεων σε εθνικό επίπεδο σε μεγάλα κανάλια πωλήσεων στο πλαίσιο μιας μοναδικής συμφωνίας. Οι έμποροι επωφελούνται από την καταχώριση των προσφορών τους στο ηλεκτρονικό κατάστημα και την αξιοποίηση του δικτύου υλικοτεχνικής υποστήριξης αυτού για άμεσες αποστολές στους πελάτες.

Ο πίνακας περιγραφικών στατιστικών στοιχείων (Πίνακας 1) παρέχει μια ολοκληρωμένη εικόνα των διαφόρων βασικών χαρακτηριστικών του συνόλου δεδομένων. Παρατηρώντας τις κεντρικές τάσεις, είναι αξιοσημείωτο ότι αρκετά χαρακτηριστικά παρουσιάζουν σημαντική μεταβλητότητα. Για παράδειγμα, ο μέσος όρος της μεταβλητής «customer_zip_code_prefix» των περίπου 35 δισεκατομμυρίων υποδηλώνει σημαντική ποικιλομορφία στις τοποθεσίες των πελατών, η οποία υποστηρίζεται από ένα ευρύ φάσμα τιμών ταχυδρομικού κώδικα περίπου 99 δισεκατομμυρίων. Ωστόσο, η τυπική απόκλιση περίπου 30 δισεκατομμυρίων από τον μέσο όρο υποδηλώνει διασπορά σε αυτή την κατανομή, ενισχύοντας τη διακύμανση στις τοποθεσίες των πελατών. Αντίθετα, άλλα χαρακτηριστικά, όπως το «review_score», υποδεικνύοντας ενδεχομένως μια τάση ανόδου λόγω του περιορισμένου εύρους βαθμολογίας.

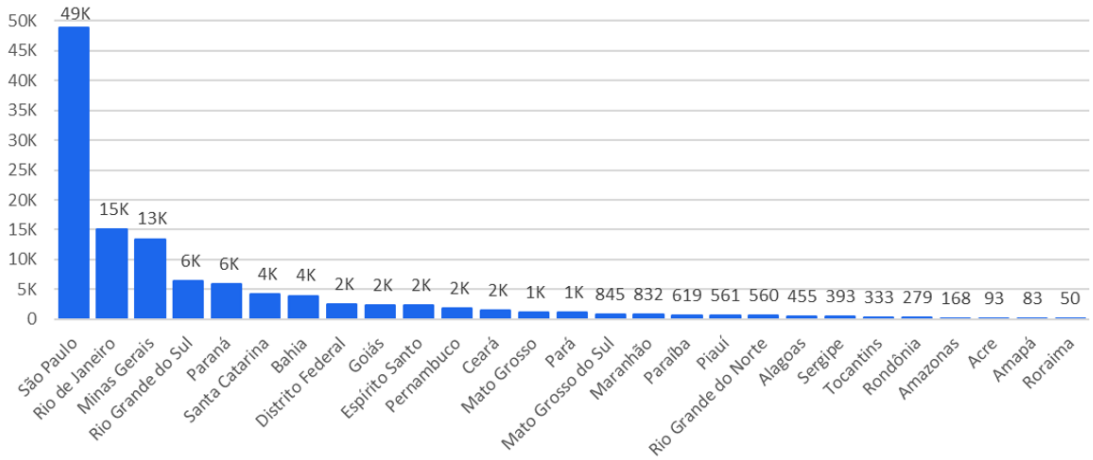
	count	*mean*	*std*	*min*	*0,25*	*0,5*	*0,75*	*max*
zip_code	115.609.000.000	35.061.537.597	29.841.671.732	1.003.000.000	11.310.000.000	24.241.000.000	58.745.000.000	99.980.000.000
order_item_id	115.609.000.000	1.194.535	0.685926	1.000.000	1.000.000	1.000.000	1.000.000	21.000.000
price	115.609.000.000	120.619.850	182.653.476	0.850000	39.900.000	74.900.000	134.900.000	6.735.000.000
freight_value	115.609.000.000	20.056.880	15.836.184	0.000000	13.080.000	16.320.000	21.210.000	409.680.000
product_name_lenght	115.609.000.000	48.766.541	10.034.187	5.000.000	42.000.000	52.000.000	57.000.000	76.000.000
product_description_lenght	115.609.000.000	785.808.198	652.418.619	4.000.000	346.000.000	600.000.000	983.000.000	3.992.000.000
product_photos_qty	115.609.000.000	2.205.373	1.717.771	1.000.000	1.000.000	1.000.000	3.000.000	20.000.000
product_weight_g	115.608.000.000	2.113.907.697	3.781.754.895	0.000000	300.000.000	700.000.000	1.800.000.000	40.425.000.000
product_length_cm	115.608.000.000	30.307.903	16.211.108	7.000.000	18.000.000	25.000.000	38.000.000	105.000.000
product_height_cm	115.608.000.000	16.638.477	13.473.570	2.000.000	8.000.000	13.000.000	20.000.000	105.000.000
product_width_cm	115.608.000.000	23.113.167	11.755.083	6.000.000	15.000.000	20.000.000	30.000.000	118.000.000
payment_sequential	115.609.000.000	1.093.747	0.729849	1.000.000	1.000.000	1.000.000	1.000.000	29.000.000
payment_installments	115.609.000.000	2.946.233	2.781.087	0.000000	1.000.000	2.000.000	4.000.000	24.000.000
payment_value	115.609.000.000	172.387.379	265.873.969	0.000000	60.870.000	108.050.000	189.480.000	13.664.080.000
seller_zip_code_prefix	115.609.000.000	24.515.713.958	27.636.640.968	1.001.000.000	6.429.000.000	13.660.000.000	28.605.000.000	99.730.000.000
review_score	115.609.000.000	4.034.409	1.385.584	1.000.000	4.000.000	5.000.000	5.000.000	5.000.000

Πίνακας 1 Περιγραφικά στατιστικά στοιχεία δεδομένων

Για την ανάλυση χρησιμοποιήθηκαν διάφορα σύνολα δεδομένων, πινάκων από την βάση δεδομένων της επιχείρησης, όπως απεικονίζεται στην Εικόνα 3.1, ωστόσο παρά το μέγεθος των συνολικών εγγραφών όπως απεικονίζονται στον παραπάνω πίνακα (Πίνακας 1), μετά τον καθαρισμό και την εύρεση των μοναδικών αναγνωριστικών το δείγμα μας μεταμορφώνεται, όπως απεικονίζεται παρακάτω (Πίνακας 2). Το σύνολο δεδομένων «Παραγγελίες» περιλαμβάνει 99.441 εγγραφές, ωστόσο ορισμένες στήλες που σχετίζονται με την έγκριση και την παράδοση έχουν ελλιπείς τιμές, αυτές θα εξαιρεθούν από την ανάλυση. Το δείγμα περιέχει 99.441 διακριτά αναγνωριστικά παραγγελιών (order ids) και ισάριθμα μοναδικά αναγνωριστικά πελατών (customer ids). Το σύνολο δεδομένων «Customers», σε συμφωνία με το σύνολο δεδομένων «Orders». Στο σύνολο δεδομένων «Items», υπάρχουν 112.650 καταχωρήσεις που αντιστοιχούν σε 98.666 διαφορετικές παραγγελίες. Υπάρχουν διπλά αναγνωριστικά παραγγελίας, λόγω του γεγονότος ότι η ίδια παραγγελία μπορεί να πληρωθεί με πολλαπλούς τρόπους πληρωμής.

DataFrame	Nunique	Count
customers_df	customer_id	99441
geo_df	geolocation_zip_code_prefix	19015
orderitem_df	order_id	98666
orderpay_df	order_id	99440
orderreviews_df	review_id	98410
orders_df	order_id	99441
products_df	product_id	32951
sellers_df	seller_id	3095
categname_df	product_category_name	71

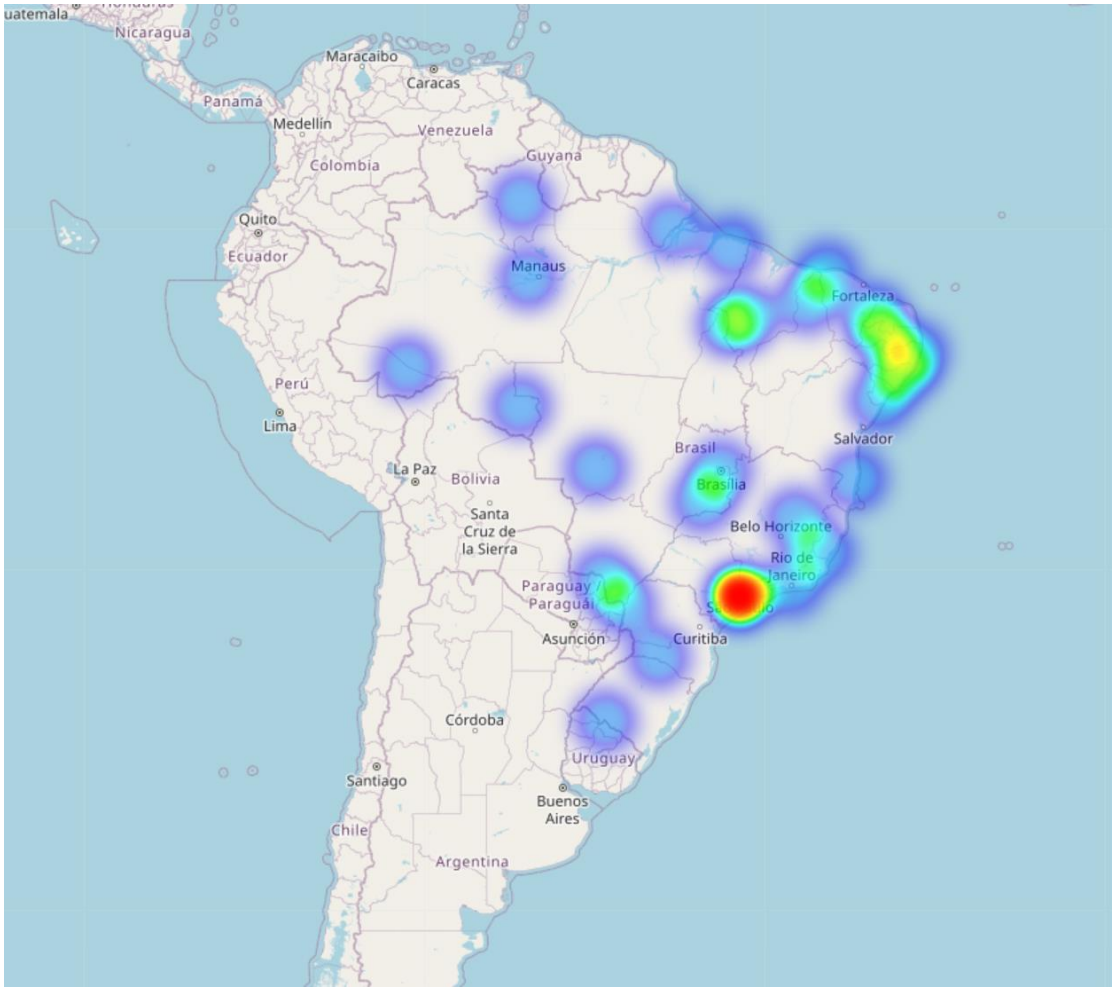
Πίνακας 2 Μοναδικές τιμές αναγνωριστικών πινάκων



Εικόνα 3.2 Κατανομή πελατών ανά περιοχή

Η συλλογή δεδομένων για τη συγκεκριμένη μελέτη ήταν δύσκολη, λόγω της ευαισθησίας των προσωπικών δεδομένων που αφορούν τους πελάτες. Τα δεδομένα

προέρχονται από ένα από τα πιο δημοφιλή ηλεκτρονικά καταστήματα στη Βραζιλία και δημοσιεύτηκαν στο Kaggle - τη μεγαλύτερη κοινότητα επιστήμης δεδομένων. Επιπλέον, διατεθεί ένα σύνολο δεδομένων γεωγραφικού εντοπισμού, το οποίο συνδέει τους ταχυδρομικούς κώδικες της Βραζιλίας με τις αντίστοιχες συντεταγμένες γεωγραφικού πλάτους και μήκους. Σημειώστε ότι το εν λόγω σύνολο δεδομένων περιλαμβάνει αυθεντικά εμπορικά δεδομένα που έχουν ανωνυμοποιηθεί. Τυχόν αναφορές σε εταιρείες και συνεργάτες εντός των αξιολογήσεων πελατών έχουν αντικατασταθεί με ονόματα από τους ευγενείς οίκους του Game of Thrones για την διατήρηση της ανωνυμίας και την αποφυγή συσχέτισης με άλλα επώνυμα εμπορικά σήματα.



Εικόνα 3.3 Πελάτες Geolocation Heatmap

3.1.3 Διαδικασία επεξεργασίας δεδομένων

Με σκοπό την μελέτη και πρόβλεψη της αξία διάρκειας ζωής των πελατών του καταστήματος, λίγο πριν την εφαρμογή τεχνικών μηχανικής μάθησης, έγινε μια ειδική επεξεργασία των δεδομένων. Πιο συγκεκριμένα, πραγματοποιήθηκε ειδική μετατροπή

στην ημερομηνία στο σύνολο δεδομένων, για να οριοθετηθούν οι περίοδοι για ανάλυση (3 μήνες και 6 μήνες). Αυτή η μετατροπή είναι θεμελιώδης για τη δόμηση της ανάλυσης, καθώς επιτρέπει τον ακριβή και συνεπή χειρισμό των δεδομένων ημερομηνίας και ώρας. Στο σενάριό μας, ορίζονται συγκεκριμένες ημερομηνίες που σηματοδοτούν την έναρξη και το τέλος δύο διαφορετικών περιόδων - η μία καλύπτει τρεις μήνες και η άλλη έξι μήνες. Οι ημερομηνίες αυτές είναι:

Ημερομηνία έναρξης: 1 Οκτωβρίου 2017

Ημερομηνία λήξης για την περίοδο 3 μηνών: 1 Ιανουαρίου 2018

Ημερομηνία λήξης για περίοδο 6 μηνών: Απριλίου 2018

Με τη μετατροπή αυτών των ημερομηνιών, το σενάριο αποτυπώνει με ακρίβεια την ακριβή διάρκεια κάθε περιόδου. Αυτή η ακρίβεια είναι ζωτικής σημασίας για τη μετέπειτα ανάλυση, διασφαλίζοντας ότι όλοι οι υπολογισμοί και οι συγκρίσεις που σχετίζονται με το χρόνο είναι ακριβείς και ουσιαστικές.

Πέρα από τα παραπάνω για το τελικό κομμάτι της μελέτης των δεδομένων που παρείχε το ηλεκτρονικό κατάστημα, την ανάλυση συναισθήματος των αξιολογήσεων από τους πελάτες, χρειάστηκε μια πιο λεπτομερής προεπεξεργασία κειμένου(text processing) σχετικά με συχνές επαναλαμβανόμενες εκφράσεις(regular expressions). Στην ανάλυση εφαρμόζονται διάφορες επεξεργασίες για τον καθαρισμό και τη τυποποίηση των δεδομένων κειμένου. Αυτό περιλαμβάνει το χειρισμό διακεκομμένων γραμμών, υπερσυνδέσμων, ημερομηνιών, νομισματικών τιμών, αριθμητικών τιμών, εκφράσεων άρνησης και ειδικών χαρακτήρων. Επίσης, απομακρύνθηκαν διάφορα μέρη του λόγου όπως άρθρα, σύνδεσμοι, προθέσεις και άλλες άκλιτες λέξεις που συνήθως αφαιρούνται κατά την επεξεργασία και ανάλυση κειμένων γνωστά ως «stop words», με σκοπό τη μείωση του θορύβου στα δεδομένα κειμένου. Σε αυτή την μελέτη περίπτωσης χρησιμοποιήθηκε η διάσημη στην NLP ανάλυση βιβλιοθήκη της python που ονομάζεται «NLTK» και ενίσχυσε την επεξεργασία κειμένου που ήταν ιδιαίτερα απαιτητική μιας και η γλώσσα των κειμένων ήταν τα πορτογαλικά. Τέλος, χρησιμοποιήθηκε και η μέθοδος «Stemming», σύμφωνα με την οποία οι λέξεις μειώνονται στη ριζική τους μορφή- ένα ακόμη βήμα που βοηθά στη γενίκευση του μοντέλου και στη μείωση της πολυπλοκότητας του συνόλου δεδομένων.

3.2 ΔΙΕΡΕΥΝΗΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ (EDA)

Με την ενοποίηση των διαφόρων συνόλων δεδομένων χρησιμοποιώντας κοινά κλειδιά (όπως αναγνωριστικά παραγγελίας, αναγνωριστικά πελάτη, αναγνωριστικά προϊόντος κ.λπ.), στόχος ήταν η δημιουργία ενός ολοκληρωμένου συνόλου δεδομένων που θα ενσωματώνει πολλαπλές διαστάσεις του συστήματος ηλεκτρονικού εμπορίου. Αυτή η μεθοδολογία συνένωσης συνόλων δεδομένων βοηθά στον εμπλουτισμό του συνόλου δεδομένων με την ενσωμάτωση σχετικών πληροφοριών σε διάφορους τομείς. Το ολοκληρωμένο σύνολο δεδομένων θα χρησιμεύσει ως βάση για την εκτέλεση αναλύσεων, τη δημιουργία μοντέλων και την παραγωγή πληροφοριών που καλύπτουν ολόκληρο τον κύκλο αγοράς των πελατών - επιτρέποντας μια πιο ολιστική κατανόηση της συμπεριφοράς των πελατών, των επιδόσεων των προϊόντων και των λειτουργικών πτυχών στο πλαίσιο της πλατφόρμας ηλεκτρονικού εμπορίου.

3.3 ΜΗΧΑΝΙΚΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ (FEATURE ENGINEERING)

Στη σφαίρα της ανάλυσης πελατών, η δημιουργία μεταβλητών για την εύρεση χαρακτηριστικών χρησιμεύει ως βασικό βήμα για τη μετατροπή των ακατέργαστων δεδομένων σε πληροφορίες που αποτυπώνουν βασικές πτυχές της συμπεριφοράς των πελατών. Ένα θεμελιώδες μέτρο σε αυτόν τον τομέα είναι η μεταβλητή «Recency», που αντιπροσωπεύει τη χρονική εγγύτητα της τελευταίας αγοράς ενός πελάτη. Στην παρούσα μελέτη, η μεταβλητή «Recency» υπολογίζεται σχολαστικά μέσω μιας σειράς βημάτων. Πρώτον, το σύνολο δεδομένων ομαδοποιείται με βάση τα μοναδικά αναγνωριστικά πελατών και εξάγεται η μέγιστη ημερομηνία αγοράς παραγγελίας για κάθε πελάτη. Η μεταβλητή που υπολογίζεται, ονομάζεται «LastPurchaseDate» και κατά αυτό τον τρόπο, υπολογίζεται η βαθμολογία «Recency» με τον προσδιορισμό του αριθμού των ημερών μεταξύ της πιο πρόσφατης ημερομηνίας αγοράς στο σύνολο δεδομένων και της τελευταίας ημερομηνίας αγοράς κάθε πελάτη. Αυτή η χρονική διάσταση επιτρέπει μια λεπτομερή κατανόηση της συχνότητας των αλληλεπιδράσεων των πελατών με την πλατφόρμα. Για να βελτιωθεί η ερμηνευσιμότητα της μεταβλητής «Recency», χρησιμοποιείται μια τεχνική ομαδοποίησης, συγκεκριμένα η ομαδοποίηση K-means, όπως αναφέρθηκε παραπάνω. Το βήμα αυτό περιλαμβάνει μια επαναληπτική διαδικασία για τον εντοπισμό των βέλτιστων αριθμών συστάδων, ακολουθούμενη από τη δημιουργία διακριτών συστάδων. Οι προκύπτουσες συστάδες παρέχουν μια δομημένη άποψη της

συμπεριφοράς των πελατών, επιτρέποντας την ανάλυση. Κατά τη διαδικασία του feature engineering, όπως φαίνεται στον υπολογισμό «Recency», συμβάλλει καθοριστικά στην δημιουργία πολύπλοκων χρονικών προτύπων σε μια συνοπτική και κατατοπιστική μέτρηση, συμβάλλοντας σημαντικά στη συνολική αποτελεσματικότητα της τμηματοποίησης και της ανάλυσης πελατών.

Η διερεύνηση της συμπεριφοράς των πελατών εμπλουτίζεται περαιτέρω μέσω του υπολογισμού των μετρήσεων «Frequency» και «Monetary», ώστε να καταλήξουμε στον τελικό υπολογισμό του RFM σκορ. Η ανάλυση εκτυλίσσεται με παράλληλο τρόπο, με την ομαδοποίηση του συνόλου δεδομένων βάσει μοναδικών αναγνωριστικών πελατών, προσδιορίζεται ο αριθμός των μοναδικών παραγγελιών ανά πελάτη, δημιουργώντας τη μετρική «Frequency». Αυτή η μετρική περικλείει το βαθμό δέσμευσης των πελατών, ρίχνοντας φως στη συνήθη φύση των συναλλαγών τους. Ομοίως, επεκτείνεται στη μεταβλητή «Monetary», όπου συγκεντρώνεται η συνολική αξία πληρωμής ανά πελάτη. Αυτό συμπυκνώνει την οικονομική συνεισφορά κάθε πελάτη στην πλατφόρμα ηλεκτρονικού εμπορίου. Η ενσωμάτωση αυτών των μετρικών, των «Recency», «Frequency» και «Monetary» στο κύριο πλαίσιο δεδομένων διευκολύνει μια ολοκληρωμένη εικόνα της συμπεριφοράς των πελατών.

Παρόμοια με τον υπολογισμό του «Recency», τόσο οι μεταβλητές «Frequency» και «Monetary» υποβάλλονται σε ανάλυση ομαδοποίησης K-means για την αποκάλυψη εγγενών μοτίβων και την παροχή ενός μέσου για την τμηματοποίηση. Η επαναληπτική διαδικασία προσδιορισμού των βέλτιστων αριθμών συστάδων, ακολουθούμενη από τη δημιουργία διακριτών συστάδων, προσθέτει λεπτομέρεια στην ερμηνεία της δέσμευσης των πελατών και του οικονομικού αντίκτυπου. Οι προκύπτουσες συστάδες συχνότητας και νομισματικών συστάδων ενισχύουν την κατανόηση των τμημάτων πελατών, καθοδηγώντας προσαρμοσμένες στρατηγικές και την κατανομή πόρων.

Στη συνέχεια για να αποκτήσουμε μια ολιστική κατανόηση της αξίας του πελάτη, μια ακόμη μεταβλητή που δημιουργείται - ένα κρίσιμο χαρακτηριστικό για την μελέτη της σχέσης πελάτη επιχείρησης, είναι η αξία διάρκειας ζωής του πελάτη (CLV), υπολογίζεται με το συνδυασμό των μετρικών «Monetary», «Frequency» και «Recency». Η CLV προκύπτει μέσω ενός απλού, αλλά ισχυρού τύπου:

$$CLV = (Monetary * Frequency) / Recency$$

Αυτός ο τύπος συμπυκνώνει την οικονομική συνεισφορά κάθε πελάτη (Monetary) και τη συχνότητα εμπλοκής του (Frequency), λαμβάνοντας παράλληλα υπόψη την επαναληπτικότητα των αλληλεπιδράσεών του (Recency). Η προκύπτουσα μετρική CLV χρησιμεύει ως ολοκληρωμένος δείκτης της μακροπρόθεσμης αξίας ενός πελάτη για την επιχείρηση, ενσωματώνοντας τόσο τις συναλλακτικές όσο και τις χρονικές διαστάσεις. Η ενσωμάτωση της CLV στο σύνολο των χαρακτηριστικών ενισχύει το βάθος της τμηματοποίησης των πελατών και παρέχει έναν στρατηγικό φακό για εξατομικευμένες προσπάθειες μάρκετινγκ και κατανομή πόρων. Η επακόλουθη ανάλυση, αξιοποιώντας την CLV, είναι έτοιμη να αποκαλύψει γνώσεις σχετικά με την αφοσίωση των πελατών και να ενημερώσει για στοχευμένες επιχειρηματικές στρατηγικές.

3.4 ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ (DATA MODELING)

Η μοντελοποίηση δεδομένων είναι ένα κρίσιμο βήμα στη ροή της μελέτης, όπου δημιουργείται ένα μαθηματικό μοντέλο για την κατανόηση και την πραγματοποίηση προβλέψεων με βάση τα δεδομένα. Σε αυτό το σενάριο, η μοντελοποίηση δεδομένων μπορεί να χρησιμεύσει για την αποκάλυψη μοτίβων, την πραγματοποίηση προβλέψεων ή την ομαδοποίηση δεδομένων σε ομάδες. Ο στόχος είναι να μετατραπούν οι γνώσεις που προκύπτουν από τα ακατέργαστα δεδομένα σε αξιοποιήσιμη γνώση, η οποία μπορεί να εφαρμοστεί σε διάφορα πρακτικά σενάρια. Η διαδικασία αυτή περιλαμβάνει την επιλογή των κατάλληλων αλγορίθμων, την προετοιμασία των δεδομένων και τη ρύθμιση των μοντέλων για βέλτιστη απόδοση.

Σε αυτή την μελέτη περίπτωσης περιλαμβάνονται αλγόριθμοι μηχανικής μάθησης, όπως ο K-Means, που αναφέρθηκε παραπάνω κατά την δημιουργία των μετρικών «Recency», «Frequency», «Monetary» και που κυρίως χρησιμοποιείται για την ομαδοποίηση. Η ομαδοποίηση βοηθά στον εντοπισμό διακριτών ομάδων στα δεδομένα χωρίς προηγούμενη γνώση της ταυτότητας των ομάδων. Η επιλογή αυτού του αλγορίθμου υποδηλώνει την πρόθεση διερεύνησης υποκείμενων μοτίβων ή τμημάτων στο σύνολο δεδομένων σας.

3.4.1 Διαχωρισμός συνόλου δεδομένων

Η διαδικασία διαχωρισμού του συνόλου δεδομένων σε σύνολα «εκπαίδευσης» (training dataset) και δοκιμής (testing dataset) είναι ζωτικής σημασίας για την αξιολόγηση της απόδοσης ενός μοντέλου. Μια συνήθης προσέγγιση είναι ο

διαχωρισμός σε ένα μεγαλύτερο μέρος (π.χ. 70-80%) για τις ανάγκες της εκπαίδευσης και το υπόλοιπο για δοκιμή. Αυτός ο διαχωρισμός βοηθά στην επικύρωση της απόδοσης του μοντέλου σε αθέατα δεδομένα και χρησιμοποιείται κατά την πρόβλεψη της αξίας των πελατών (CLV).

Για το ίδιο μοντέλο πρόβλεψης χρησιμοποιήθηκε και μια ακόμα τεχνική διαχωρισμού δεδομένων. Ακολουθώντας τη μετατροπή της ημερομηνίας, που αναφέραμε στο Κεφάλαιο 3.1.2, η μελέτη προχωρά στην τμηματοποίηση των δεδομένων των πελατών με βάση την ημερομηνία της τελευταίας τους αγοράς. Αυτή η τμηματοποίηση είναι μια στρατηγική προσέγγιση για την κατηγοριοποίηση των πελατών σύμφωνα με τις πρόσφατες αλληλεπιδράσεις τους με την επιχείρηση. Συγκεκριμένα, δημιουργούνται δύο διακριτά τμήματα πελατών: Πελάτες με αγορές τους τελευταίους 3 μήνες: Αυτό το τμήμα περιλαμβάνει πελάτες των οποίων η τελευταία αγορά εμπίπτει μεταξύ της 1ης Οκτωβρίου 2017 και της 1ης Ιανουαρίου 2018. Αυτή η ομάδα είναι απαραίτητη για την κατανόηση της πρόσφατης συμπεριφοράς και δέσμευσης των πελατών. Πελάτες με αγορές τους τελευταίους 6 μήνες: Αυτή η ομάδα περιλαμβάνει πελάτες που πραγματοποίησαν την τελευταία τους αγορά μεταξύ της 1ης Οκτωβρίου 2017 και της 1ης Απριλίου 2018. Αντιπροσωπεύει ένα ευρύτερο χρονικό πλαίσιο, αποτυπώνοντας ένα πιο εκτεταμένο ιστορικό αλληλεπίδρασης.

Πολύ σημαντική είναι εξίσου και η τεχνική διασταυρούμενης επικύρωσης (Cross-Validation Techniques), όπως η μέθοδος K-fold. Αυτή η τεχνική περιλαμβάνει το διαχωρισμό των δεδομένων σε k υποσύνολα και την επαναληπτική χρήση ενός υποσυνόλου για δοκιμή και των υπολοίπων για εκπαίδευση. Εξασφαλίζει μια πιο ισχυρή αξιολόγηση της απόδοσης του μοντέλου σε διαφορετικά δείγματα δεδομένων. Αυτή η μέθοδος χρησιμοποιείται στην συγκεκριμένη μελέτη κατά την διάρκεια της ανάλυση συναισθήματος των κριτικών που έχουν τοποθετήσει οι πελάτες στις παραγγελίες τους.

3.4.2 Μοντέλο Πρόβλεψης διάρκειας ζωής πελατών (CLV Prediction)

Η μεθοδολογία που παρουσιάζεται σε αυτό το τμήμα αντιπροσωπεύει μια λεπτομερή προσέγγιση για την κατανόηση και την πρόβλεψη της αξίας διάρκειας ζωής του πελάτη (CLV). Ξεκινά από την ειδική προετοιμασία των δεδομένων, την τμηματοποίηση και την οπτικοποίηση, προχωρά μέσω της ομαδοποίησης των πελατών με βάση τα έσοδα και καταλήγει στην εφαρμογή της μηχανικής μάθησης για

την πρόβλεψη τμημάτων CLV. Η χρήση της τμηματοποίησης με βάση το χρόνο, της ανάλυσης εσόδων, της ομαδοποίησης και της μηχανικής μάθησης παρέχει μια ολοκληρωμένη άποψη της αξίας των πελατών, καθιστώντας την μια ισχυρή προσέγγιση για την πρόβλεψη της CLV.

Η διαδικασία αρχίζει με την κωδικοποίηση των κατηγορικών μεταβλητών. Σε αυτό το συγκεκριμένο πλαίσιο, η στήλη "Segment", η οποία κατηγοριοποιεί τους πελάτες, μετατρέπεται σε αριθμητική μορφή μέσω κωδικοποίησης με ένα σημείο. Αυτός ο μετασχηματισμός είναι ζωτικής σημασίας για τις εφαρμογές μηχανικής μάθησης, καθώς οι περισσότεροι αλγόριθμοι έχουν σχεδιαστεί για να εργάζονται με αριθμητικά δεδομένα. Η διαδικασία κωδικοποίησης one-hot δημιουργεί δυαδικές στήλες για κάθε κατηγορία στο αρχικό πεδίο "Segment", μετατρέποντας ουσιαστικά τα ποιοτικά δεδομένα σε μορφή κατάλληλη για ποσοτική ανάλυση.

Μετά την κωδικοποίηση, δημιουργείται ένας πίνακας συσχέτισης. Αυτός ο πίνακας είναι ένα ισχυρό εργαλείο για την αποκάλυψη σχέσεων μεταξύ διαφορετικών μεταβλητών στο σύνολο δεδομένων, τονίζοντας ιδιαίτερα τον τρόπο με τον οποίο κάθε μεταβλητή συσχετίζεται με το "LTVCluster". Το "LTVCluster" είναι ένα κρίσιμο στοιχείο της ανάλυσης και αντιπροσωπεύει διαφορετικά τμήματα της αξίας διάρκειας ζωής των πελατών. Με την εξέταση του πίνακα συσχέτισης, μπορούν να αποκτηθούν γνώσεις σχετικά με το ποιοι παράγοντες συνδέονται πιο έντονα με υψηλότερες ή χαμηλότερες αξίες διάρκειας ζωής των πελατών. Αυτή η ανάλυση δεν είναι μόνο καθοριστική για την κατανόηση της τρέχουσας κατάστασης των αλληλεπιδράσεων με τους πελάτες, αλλά και για την πρόβλεψη της μελλοντικής συμπεριφοράς και αξίας.

Το τελευταίο και πιο εξελιγμένο μέρος της μεθοδολογίας περιλαμβάνει τη χρήση μηχανικής μάθησης για την πρόβλεψη τμημάτων CLV. Αυτό εκτελείται με τη χρήση του PyCaret, μιας αυτοματοποιημένης βιβλιοθήκης μηχανικής μάθησης που βελτιώνει τη διαδικασία επιλογής και συντονισμού μοντέλων. Το αρχικό βήμα σε αυτή τη φάση είναι η ρύθμιση ενός πειράματος ταξινόμησης με το PyCaret, ορίζοντας το "LTVCluster" ως μεταβλητή-στόχο. Το 'LTVCluster' ουσιαστικά κατηγοριοποιεί τους πελάτες σε διαφορετικά επίπεδα αξίας διάρκειας ζωής, καθιστώντας το έργο κατηγορικής πρόβλεψης.

Βασικό στοιχείο αυτής της φάσης είναι η σύγκριση και η επιλογή μοντέλων μηχανικής μάθησης. Το σενάριο χρησιμοποιεί τη λειτουργικότητα του PyCaret για να συγκρίνει διάφορα μοντέλα με βάση την απόδοσή τους, εστιάζοντας ιδιαίτερα στη μετρική AUC. Η AUC είναι μια ευρέως χρησιμοποιούμενη μετρική σε εργασίες

ταξινόμησης, καθώς παρέχει ένα ολοκληρωμένο μέτρο της ικανότητας ενός μοντέλου να διακρίνει μεταξύ διαφορετικών κλάσεων. Στη συνέχεια επιλέγονται τα τρία κορυφαία μοντέλα, όπως καθορίζονται από τις βαθμολογίες AUC, για περαιτέρω βελτίωση.

Η τελειοποίηση αυτών των μοντέλων περιλαμβάνει τη ρύθμιση των υπερπαραμέτρων τους για τη βελτιστοποίηση της απόδοσης, στοχεύοντας συγκεκριμένα στη βελτίωση της AUC. Ο συντονισμός των υπερπαραμέτρων είναι ένα κρίσιμο βήμα στη μηχανική μάθηση, καθώς μπορεί να βελτιώσει σημαντικά την ακρίβεια πρόβλεψης ενός μοντέλου. Με την προσαρμογή των διαφόρων ρυθμίσεων και παραμέτρων των αλγορίθμων, τα μοντέλα είναι καλύτερα εξοπλισμένα για να συλλάβουν τις πολυπλοκότητες και τις αποχρώσεις του συνόλου δεδομένων, οδηγώντας σε ακριβέστερες προβλέψεις των τμημάτων αξίας διάρκειας ζωής των πελατών.

Συνοψίζοντας, αυτό το μέρος της μεθοδολογίας ενσωματώνει προηγμένες τεχνικές επεξεργασίας δεδομένων με μηχανική μάθηση για την πρόβλεψη τμημάτων αξίας διάρκειας ζωής του πελάτη (CLV). Ο μετασχηματισμός των κατηγορικών δεδομένων σε αριθμητική μορφή ανοίγει το δρόμο για αποτελεσματικές εφαρμογές μηχανικής μάθησης, ενώ η ανάλυση συσχέτισης παρέχει πληροφορίες για τους παράγοντες που επηρεάζουν την CLV.

3.4.3 Ανάλυση Συναισθήματος (NLP & Sentiment analysis)

Με σκοπό την πιο επιτυχημένη μέτρηση και πρόβλεψη της πιστότητας πελατών, η έρευνα επεκτάθηκε πέρα από την ανάλυση RFM και CLV, στην ανάλυση του συναισθήματος(sentiment analysis) των πελατών που πραγματοποίησαν μια συναλλαγή με την επιχείρηση και προχώρησαν στην αξιολόγηση και τοποθέτηση κριτικής.

Για την ανάλυση των σχολίων που δημοσίευσαν οι πελάτες, χρησιμοποιήθηκε πρώτα η βιβλιοθήκη nltk, όπως αναφέραμε και στην προεπεξεργασία δεδομένων κατά την διαδικασία stemming, και εν συνεχεία, το CountVectorizer από το sklearn, μια βιβλιοθήκη μηχανικής μάθησης στην Python. Ο κώδικας επικεντρώνεται στην ανάλυση κειμένου, και συγκεκριμένα στη μετατροπή μιας συλλογής εγγράφων κειμένου σε έναν πίνακα με αριθμούς συμβόλων. Η κλάση CountVectorizer χρησιμοποιείται για τη μετατροπή δεδομένων κειμένου σε αριθμητική μορφή που μπορούν να κατανοήσουν τα μοντέλα μηχανικής μάθησης. Οι παράμετροι

max_features=300, min_df=7 και max_df=0.8 χρησιμοποιούνται για τη διαμόρφωση του vectorizer. Αυτό σημαίνει ότι ο vectorizer θα δημιουργήσει έναν πίνακα με μέγιστο αριθμό 300 χαρακτηριστικών (tokens), περιλαμβάνοντας μόνο τις λέξεις που εμφανίζονται σε τουλάχιστον 7 «έγγραφα», αλλά όχι σε περισσότερο από το 80% των εγγράφων.

Αρχικά, ο κώδικας προσθέτει μια πρώτη περιγραφή συναισθήματος στα σχόλια, ταξινομώντας τα ως "αρνητικά" εάν η βαθμολογία τους είναι 1, 2 ή 3, και ως "θετικά" εάν η βαθμολογία τους είναι μεγαλύτερη από 3. Αυτή η βαθμολογία προκύπτει από την αξιολόγηση του ίδιου το πελάτη, καθώς κάθε σχόλιο/κριτική ακολουθείται και από έναν βαθμό που έχει επιδώσει ο πελάτης στην επιχείρηση και την εξυπηρέτησή του από εκείνη. Ο τελικός στόχος όμως είναι η αυτοματοποίηση της διαδικασίας αυτής μέσω ενός πειράματος για την ταξινόμηση των σχολίων με τη χρήση μηχανικής μάθησης, αυτό το πείραμα περιλαμβάνει την δυαδική ταξινόμηση με χρήση ταξινομητών Logistic Regression και Naive Bayes.

Το μοντέλο Naive Bayes βασίζεται στο θεώρημα του Bayes και υποθέτει την ανεξαρτησία μεταξύ των χαρακτηριστικών. Υπολογίζει την πιθανότητα κάθε κατηγορίας για μια δεδομένη είσοδο/εγγραφή και επιλέγει την κατηγορία με την υψηλότερη πιθανότητα. Παρά την απλότητά του και την ισχυρή υπόθεση ανεξαρτησίας, το Naive Bayes μπορεί να αποδώσει σημαντικά καλά σε εργασίες ταξινόμησης κειμένου.

Από την άλλη, η λογιστική παλινδρόμηση(Logistic Regression) είναι ένα γραμμικό μοντέλο που χρησιμοποιεί μια λογιστική συνάρτηση για να μοντελοποιήσει την πιθανότητα μια δεδομένη είσοδος/εγγραφή να ανήκει σε μια συγκεκριμένη κατηγορία. Το μοντέλο αυτό συνήθως προτιμάται για την απλότητα και την ερμηνευσιμότητά του.

Ένα πείραμα δυαδικής ταξινόμησης στη μηχανική μάθηση περιλαμβάνει την εκπαίδευση αλγορίθμων για την κατηγοριοποίηση δεδομένων σε μία από δύο ομάδες. Στο πλαίσιο των αξιολογήσεων πελατών, αυτές οι δύο ομάδες είναι συνήθως τα "θετικά" και τα "αρνητικά" συναισθήματα. Σκοπός αυτού του πειράματος, όπως αναφέραμε παραπάνω, είναι να αυτοματοποιήσει τη διαδικασία ανάλυσης συναισθήματος, η οποία μπορεί να είναι εξαιρετικά πολύτιμη για τις επιχειρήσεις που θέλουν να μετρήσουν την κοινή γνώμη για τα προϊόντα ή τις υπηρεσίες τους γρήγορα και σε κλίμακα. Σε αυτή την μελέτη, πρωταρχικός στόχος είναι να προβλεφθεί με

ακρίβεια αν μια κριτική πελάτη είναι θετική ή αρνητική με βάση το κείμενο της κριτικής.

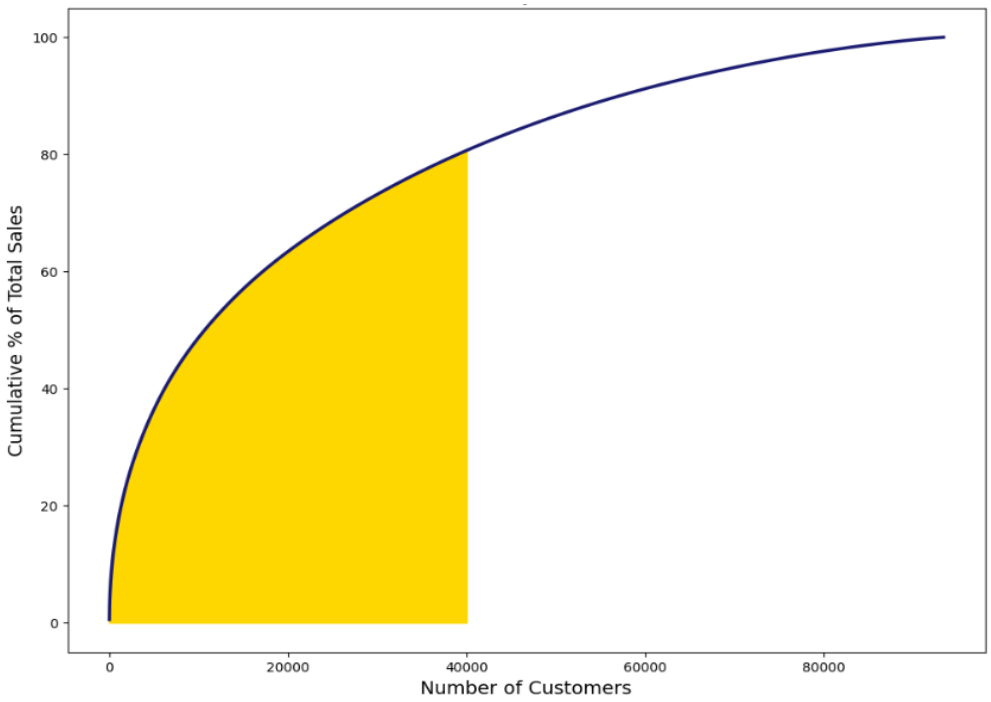
Πριν από την εκτέλεση του πειράματος, τα δεδομένα κειμένου από τις κριτικές των πελατών υποβάλλονται σε προεπεξεργασία, η οποία μπορεί να περιλαμβάνει καθαρισμό (αφαίρεση περιττών χαρακτήρων, διόρθωση τυπογραφικών λαθών), κανονικοποίηση (μετατροπή σε πεζά γράμματα), tokenization (διαχωρισμό του κειμένου σε λέξεις ή tokens), stemming/lemmatization (αναγωγή των λέξεων στη βασική τους μορφή) και αφαίρεση των stop words (κοινές λέξεις όπως "και", "η", κ.λπ., που δεν συμβάλλουν στο συναίσθημα) (3.1.3).

Για την αποτελεσματικότητα των μοντέλων μετά την εφαρμογή τους, υπολογίστηκαν μετρικές απόδοσης κατά τη διάρκεια των φάσεων εκπαίδευσης (με χρήση διασταυρούμενης επικύρωσης/cross-validation) και δοκιμής. Τα μοντέλα αξιολογούνται χρησιμοποιώντας ένα σύνολο μετρήσεων. Η ακρίβεια(Accuracy) μετράει το ποσοστό των συνολικών σωστών προβλέψεων, το precision μετράει την ακρίβεια των θετικών προβλέψεων, η ανάκληση(recall) μετράει το ποσοστό των πραγματικών θετικών προβλέψεων που αναγνωρίστηκαν σωστά και το F1 score παρέχει μια ισορροπία μεταξύ ακρίβειας και ανάκλησης. Αυτές οι μετρικές προκύπτουν από τον πίνακα σύγχυσης, ο οποίος δείχνει τα αληθώς θετικά, τα ψευδώς θετικά, τα αληθώς αρνητικά και τα ψευδώς αρνητικά.

Κεφάλαιο 4: Αποτελέσματα

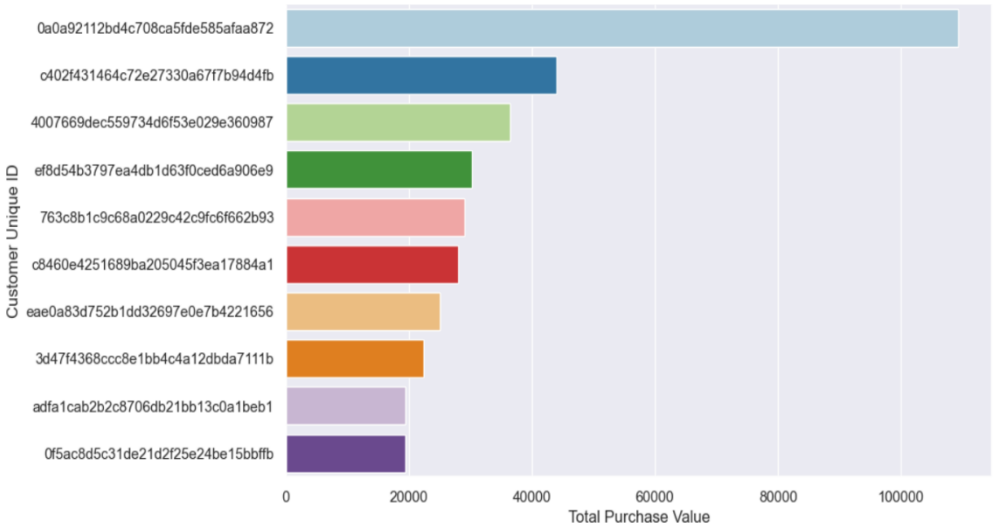
4.1 ΠΕΡΙΓΡΑΦΙΚΑ ΣΤΑΤΙΣΤΙΚΑ ΣΤΟΙΧΕΙΑ

Μια ενδιαφέρουσα μέτρηση είναι η αθροιστική κατανομή των πωλήσεων με βάση τον αριθμό των πελατών. Ο υπολογισμός των συνολικών πληρωμών που πραγματοποιήθηκαν από κάθε μοναδικό πελάτη ταξινομημένο κατά φθίνουσα σειρά αντιπροσωπεύει τους πελάτες που πληρώνουν περισσότερο. Το παρακάτω γραμμικό διάγραμμα (Εικόνα 4.1) προσφέρει μια σαφή εικόνα του πόσο από τις συνολικές πωλήσεις συγκεντρώνεται με κάθε πρόσθετο πελάτη, δείχνοντας ουσιαστικά την κατανομή των συνεισφορών στις πωλήσεις μεταξύ των διαφόρων πελατών.



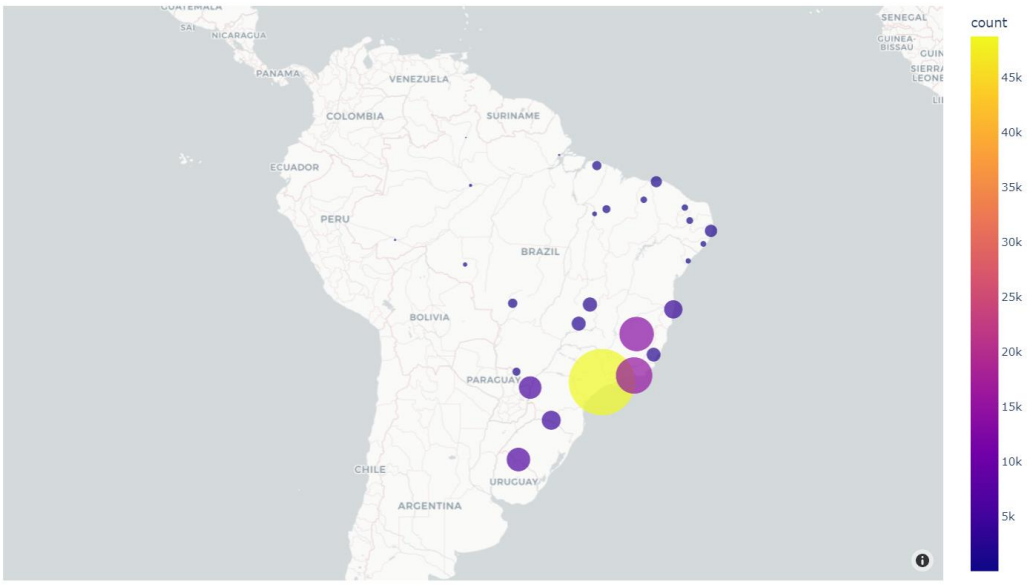
Εικόνα 4.1 Διανομή πωλήσεων: Η επίδραση των κορυφαίων πελατών στα συνολικά έσοδα

Από τη συνολική πελατειακή βάση, οι αρχικοί 40.000 πελάτες -που αποτελούν περίπου το 42% του συνολικού πελατολογίου- συμβάλλουν σημαντικά στις πωλήσεις της εταιρείας, αντιπροσωπεύοντας περίπου το 80% των συνολικών πωλήσεων. Η ομάδα αυτή ασκεί σημαντική επιρροή στα έσοδα της εταιρείας. Στο πλαίσιο απόκτησης πληροφοριών σχετικά με τον πελατολόγιο, είναι επίσης ενδιαφέρον να δούμε τους κορυφαίους πελάτες (Εικόνα 4.2).



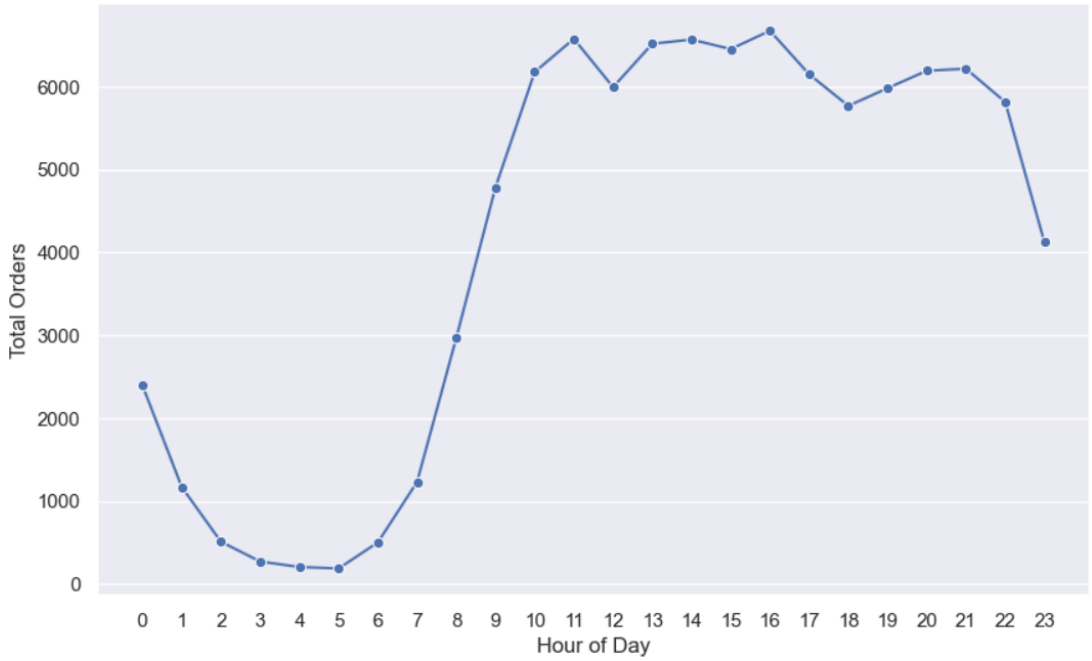
Εικόνα 4.2 Οι 10 Κορυφαίοι πελάτες με βάση την συνολική αξία αγοράς

Σημαντική αναφορά, επίσης, αξίζει και στην κατανομή των παραγγελιών ανά τοποθεσία. Ο χάρτης της Βραζιλίας συμπυκνώνει τη γεωγραφική κατανομή των παραγγελιών (Εικόνα 4.3). Κάθε φυσαλίδα απεικονίζει παραγγελίες, με το μέγεθος και το χρώμα της να αντικατοπτρίζουν τον αριθμό αυτών. Όπως παρατηρείται, οι πωλήσεις συγκεντρώνονται γύρω από την πρωτεύουσα και κεντρική χώρα της Βραζιλίας. Πριν από την ανάπτυξη των μοντέλων που θα μελετήσουμε στην συνέχεια (RFM, CLV, Sentiment analysis), οι εταιρείες χρησιμοποιούσαν συνήθως δημογραφικά προφίλ πελατών για σκοπούς μάρκετινγκ. Ωστόσο, η έρευνα υποδεικνύει ότι τα πιο σύνθετα μοντέλα είναι καλύτεροι προγνωστικοί δείκτες των μελλοντικής αγοραστικής συμπεριφοράς πελατών από ό,τι τα δημογραφικά στοιχεία (Gupta et al., 2006).

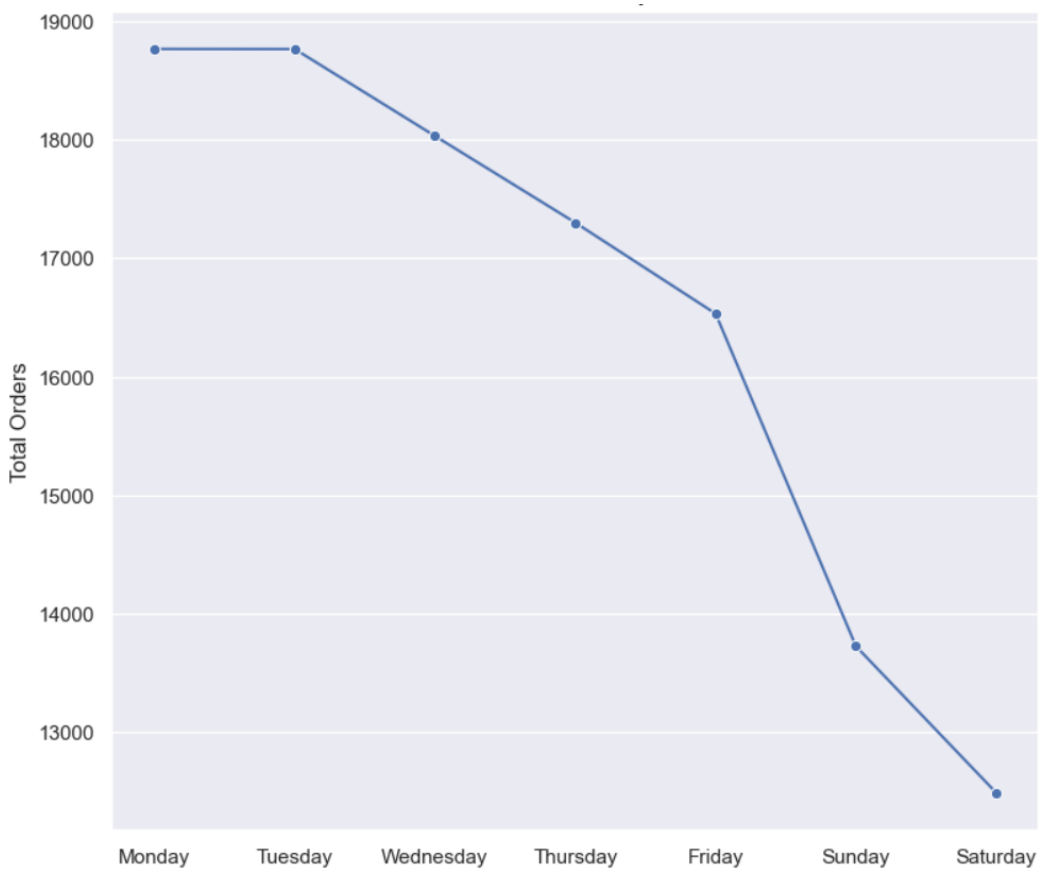


Εικόνα 4.3 Γεωγραφική κατανομή παραγγελιών ανά πόλη

Γύρω από την μελέτη των παραγγελιών έγινε και διερεύνηση του χρόνου τοποθέτησης παραγγελιών στο ηλεκτρονικό κατάστημα, ανά ώρα (Εικόνα 4.4) και ανά ημέρα της εβδομάδας (Εικόνα 4.5).



Εικόνα 4.4 Κατανομή των Παραγγελιών ανά ώρα



Εικόνα 4.5 Κατανομή Παραγγελιών ανά ημέρα της εβδομάδας

Σύμφωνα με τα παραπάνω διαγράμματα οι περίοδοι περισσότερης κίνησης και ταυτόχρονα τοποθέτησης παραγγελιών στην ιστοσελίδα του καταστήματος είναι μετά τις μεσημεριανές ώρες της ημέρας, ενώ με βάση τις ημέρες της εβδομάδας, οι αγορές τείνουν να έχουν φθίνουσα πορεία κατά την διάρκεια της εβδομάδας. Περισσότερες παραγγελίες τοποθετούνται από Δευτέρα έως Παρασκευή, μάλιστα η Κυριακή έχει περίπου 31% λιγότερες παραγγελίες απ' ότι η Δευτέρα κατά μέσο όρο.

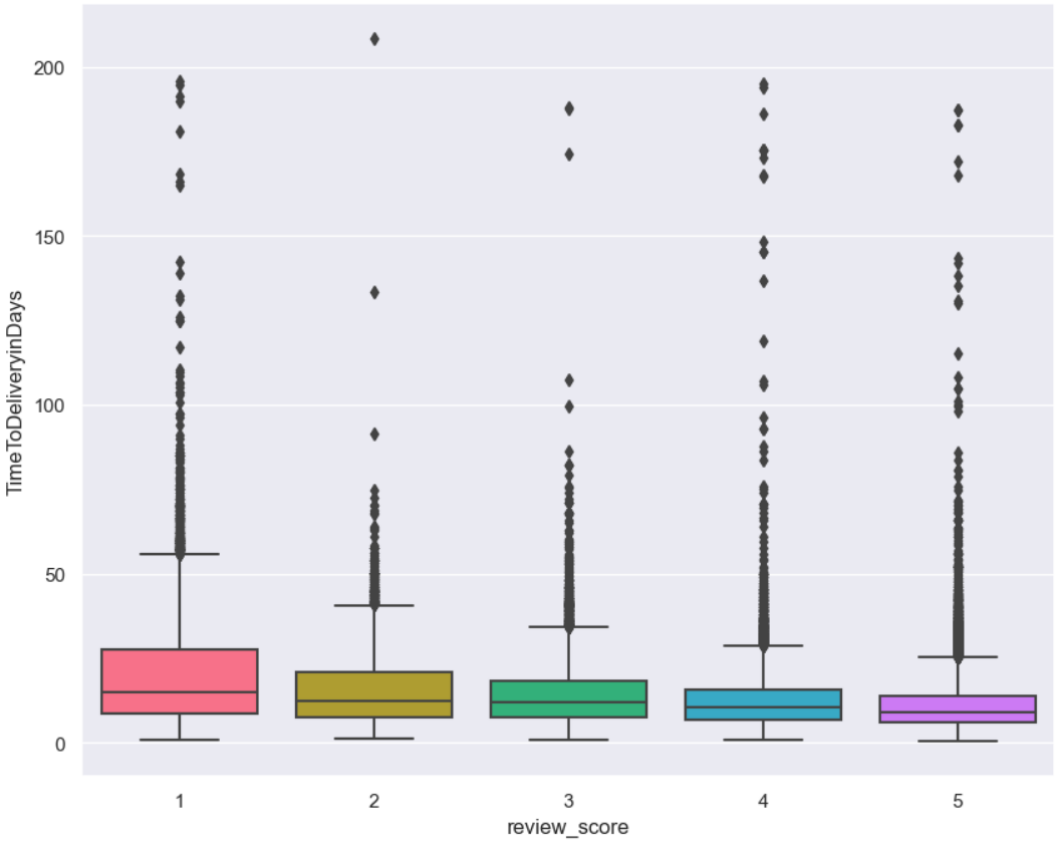
Στη συνέχεια της έρευνας των χαρακτηριστικών των παραγγελιών, ένα πολύ σημαντικό εύρημα σχετικά με την συμπεριφορά πελατών αφορά και τις ακυρώσεις. Με τη βοήθεια των δεδομένων του ηλεκτρονικού καταστήματος μπορούμε να εξετάσουμε και την πιθανή συσχέτιση μεταξύ των μεθόδων πληρωμής και των ακυρώσεων παραγγελιών.

order_status	canceled	delivered	% Canceled	Avg Cancelation Rate
payment_type				
boleto	92	22029	0.417631	0.473456
credit_card	411	83536	0.492003	0.473456
debit_card	6	1623	0.369686	0.473456
voucher	27	6022	0.448356	0.473456

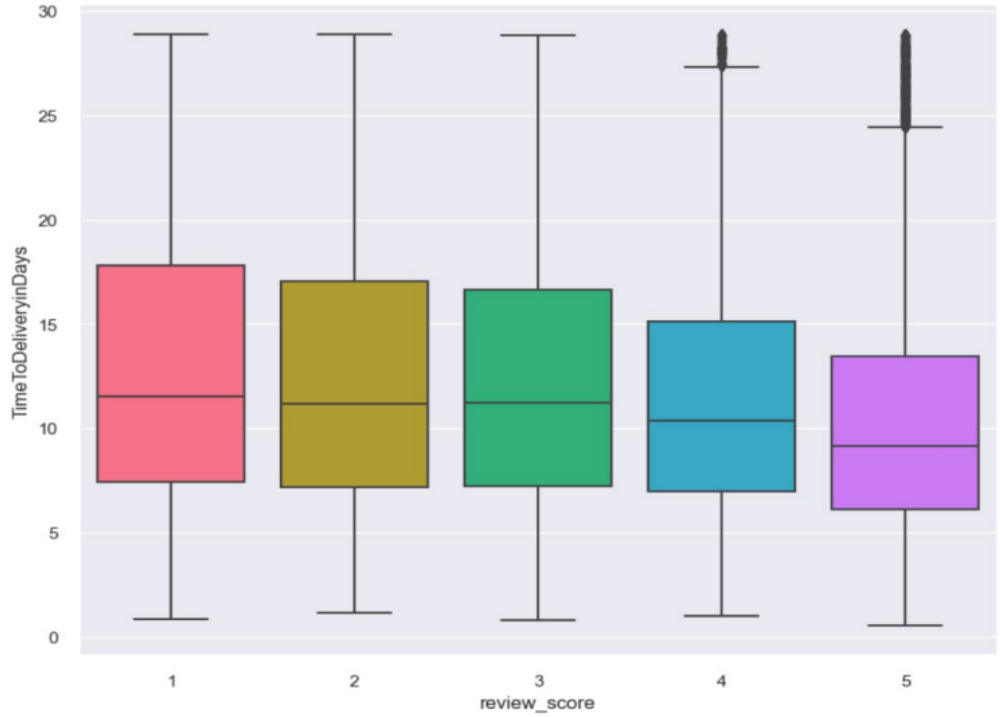
Πίνακας 3 Συσχέτιση μεθόδου πληρωμής και ακύρωση παραγγελίας

Ο παραπάνω πίνακας (Πίνακας 3) δείχνει τα ποσοστά ακύρωσης να είναι σχετικά χαμηλά για όλες τις μεθόδους πληρωμής, γεγονός που υποδηλώνει ότι η πλειονότητα των παραγγελιών παραδίδεται επιτυχώς. Η μέθοδος πληρωμής «credit_card» έχει το υψηλότερο ποσοστό ακύρωσης μεταξύ των αναφερόμενων μεθόδων. Ωστόσο είναι σημαντικό να σημειωθεί ότι το μέσο ποσοστό ακύρωσης είναι το ίδιο για όλες τις μεθόδους πληρωμής.

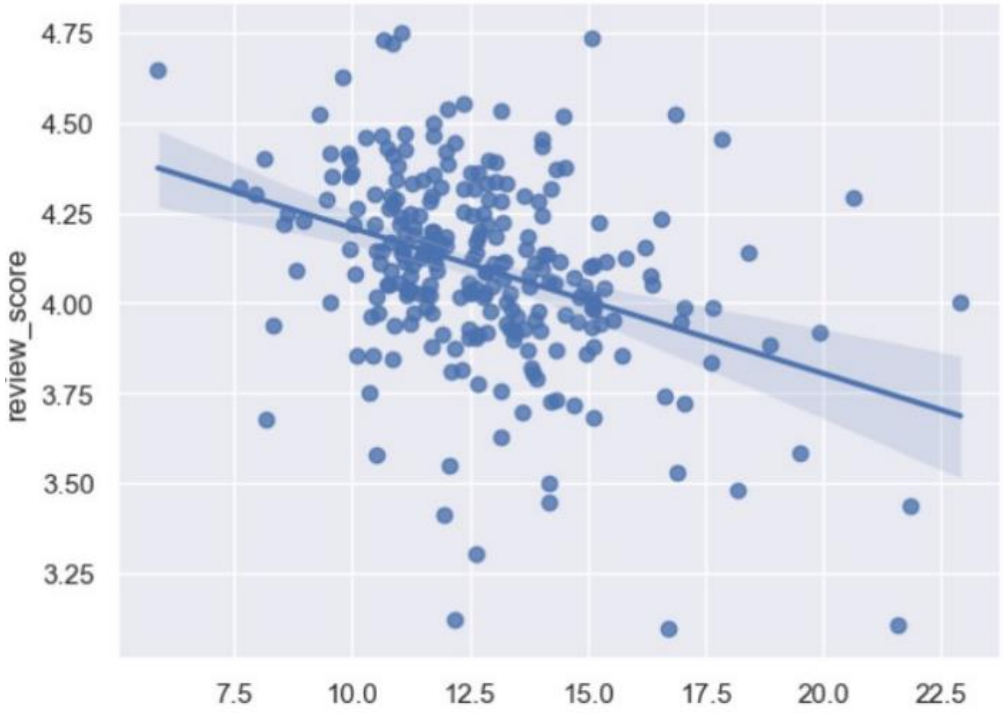
Ακόμη, σημαντικός παράγοντας στην διατήρηση των πελατών είναι η ταχύτητα παράδοσης, ένα παράπονο που συναντάται συχνά σε κριτικές. Το μέτρο αυτό εκτιμήθηκε συνδυαστικά με την βαθμολόγηση των πελατών, κάνοντας ένα πρώτο βήμα για την διερεύνηση του συναισθήματος των πελατών. Όπως παρατηρούμε (Εικόνα 4.6, Εικόνα 4.7), αρνητική συσχέτιση μεταξύ του χρόνου παράδοσης και των βαθμολογιών που σημαίνει ότι όσο μειώνεται ο χρόνος παράδοσης, οι βαθμολογίες αναθεώρησης τείνουν να αυξάνονται. Το διάγραμμα διασποράς παρέχει σαφέστερη ένδειξη της αρνητικής συσχέτισης μεταξύ χρόνου παράδοσης και βαθμολογιών, επειδή δείχνει τα μεμονωμένα σημεία δεδομένων και την κατανομή τους.



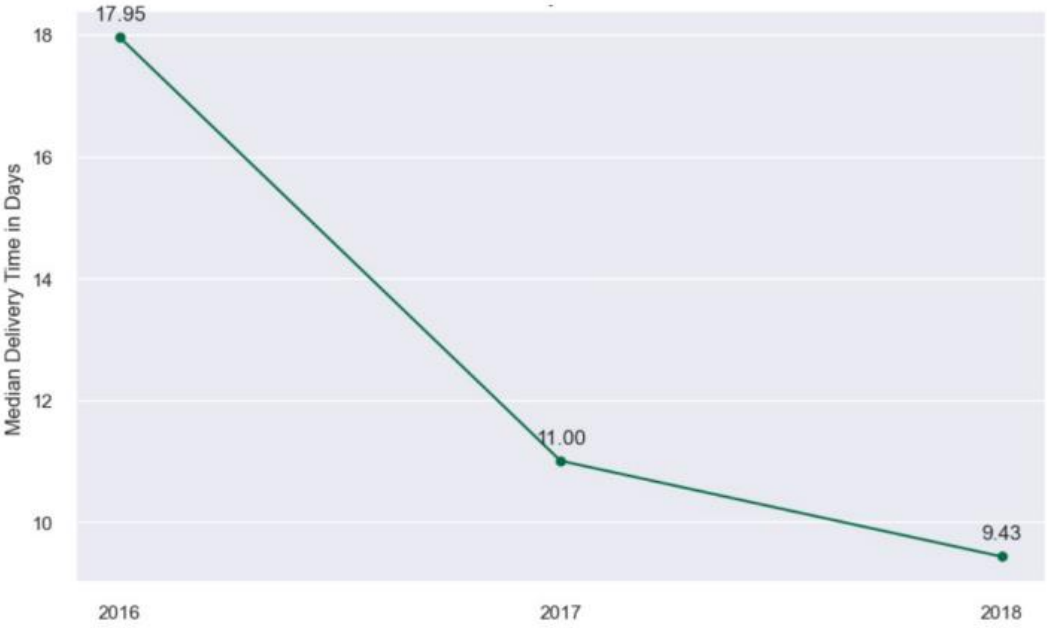
Εικόνα 4.6 Συσχέτιση χρόνου παράδοσης και βαθμολογίας



Εικόνα 4.7 Συσχέτιση χρόνου παράδοσης και βαθμολογίας χωρίς ακραίες τιμές

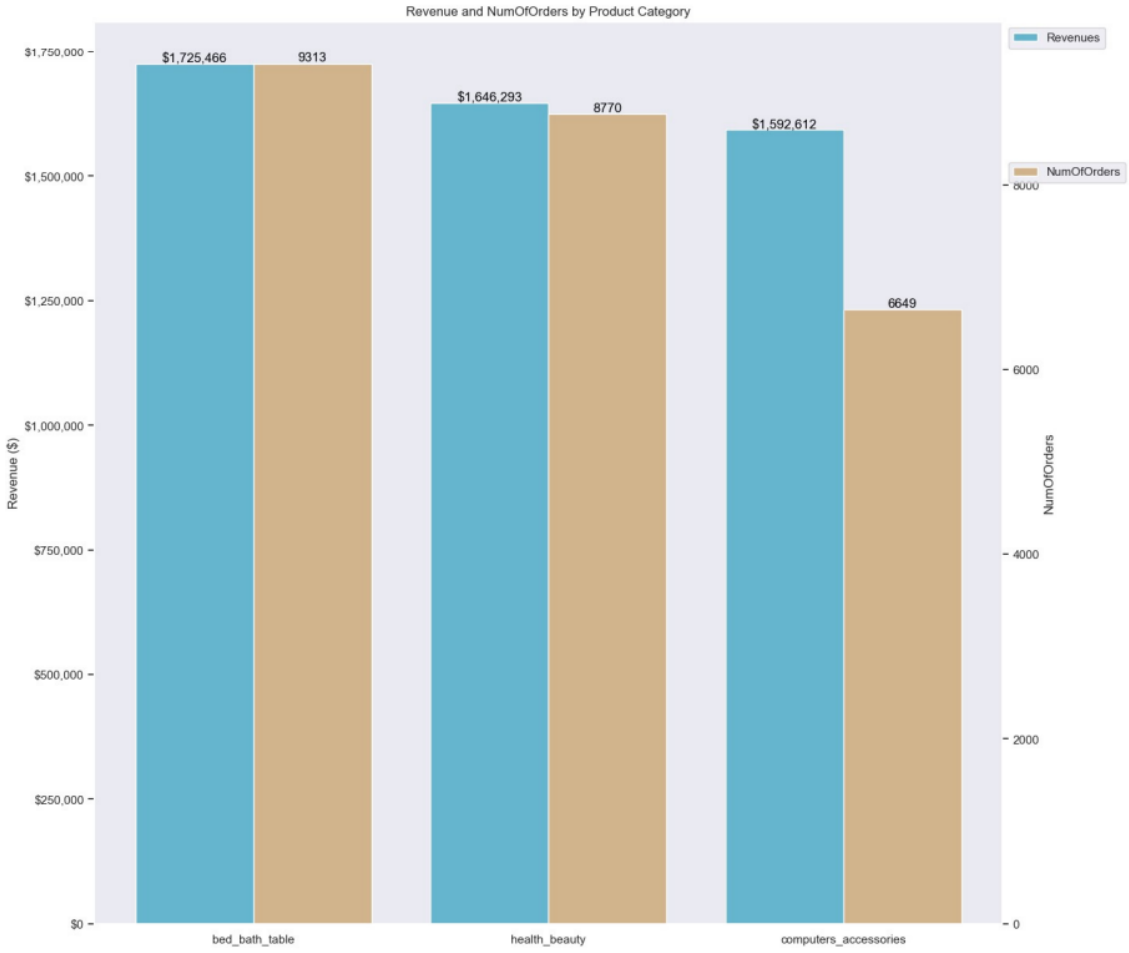


Εικόνα 4.8 Σχέση μεταξύ μέσου χρόνου παράδοσης και μέσης βαθμολογίας



Εικόνα 4.9 Μέσος χρόνος παράδοσης ανά έτος

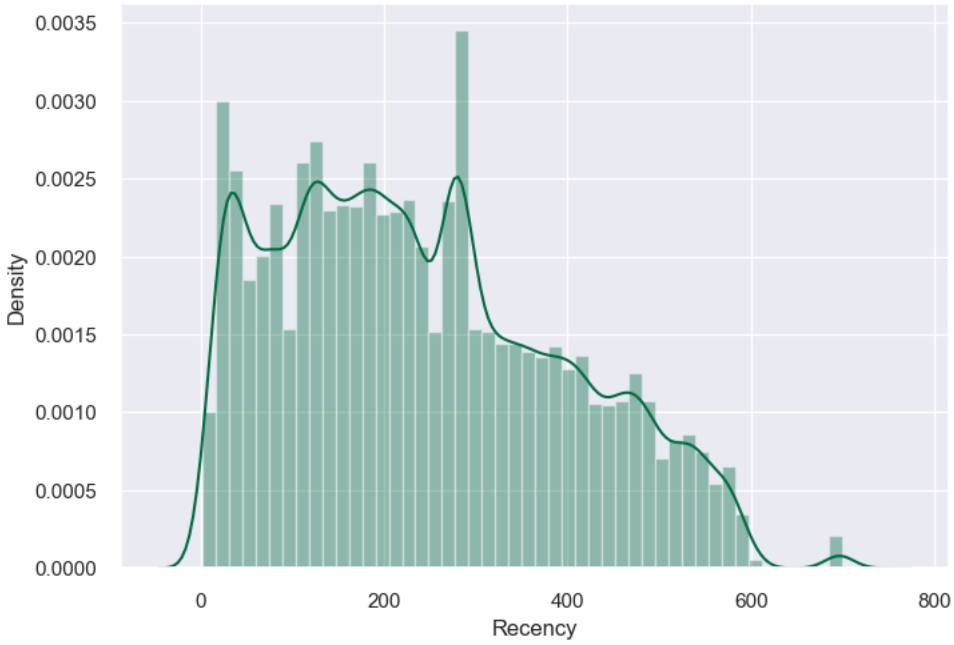
Ολοκληρώνοντας την έρευνα των περιγραφικών στατιστικών, μελετήθηκαν οι κατηγορίες προϊόντων και η επιρροή της κάθε κατηγορίας στο σύνολο των κερδών και το σύνολο των παραγγελιών. Στο παρακάτω διάγραμμα παρατηρούμε ότι οι τρεις καλύτερες κατηγορίες προϊόντων από άποψη πωλήσεων είναι τα έπιπλα μπάνιου, τα προϊόντα ομορφιάς και τα αξεσουάρ ηλεκτρονικών υπολογιστών, κατηγορίες που συγκέντρωσαν περίπου πέντε εκατομμύρια δολάρια κέρδος για την επιχείρηση.



Εικόνα 4.10 Οι καλύτερες κατηγορίες προϊόντων βάσει κέρδους και αριθμού παραγγελιών

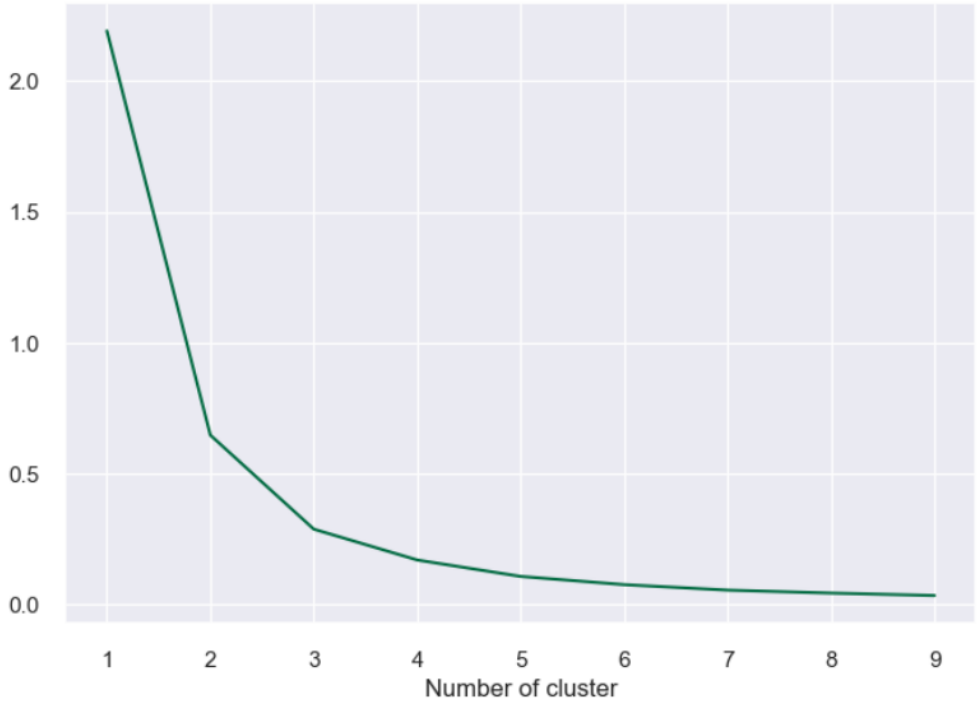
4.2 ΑΠΟΤΕΛΕΣΜΑΤΑ RFM ΑΝΑΛΥΣΗΣ

Το πρώτο βήμα προς την ανάλυση RFM εστιάζει στη μετρική "Recency", η οποία είναι απαραίτητη για την κατανόηση της συμπεριφοράς των πελατών όσον αφορά τις πιο πρόσφατες αλληλεπιδράσεις τους με την επιχείρηση. Η βαθμολογία Recency, παρουσιάζει μια κατανομή που είναι κάπως λοξή προς τα δεξιά (Εικόνα 4.11 Κατανομή Recency score, υποδεικνύοντας ότι ένας σημαντικός αριθμός πελατών έχει πραγματοποιήσει αγορές στο πιο πρόσφατο παρελθόν, με μια ουρά που δείχνει λιγότερους πελάτες με μεγαλύτερα χρονικά διαστήματα από την τελευταία τους αγορά. Τα στατιστικά στοιχεία λοξότητας και κύρτωσης επιβεβαιώνουν την έλλειψη συμμετρίας στην κατανομή, η οποία είναι τυπική για τα δεδομένα συχνότητας, καθώς οι πελάτες συνήθως αποκτώνται κατά τη διάρκεια μιας εκτεταμένης χρονικής περιόδου.



Εικόνα 4.11 Κατανομή Recency score

Οι τιμές αυτές χρησιμοποιούνται στη συνέχεια για την ανάθεση κάθε πελάτη σε μια συστάδα, χρησιμοποιώντας τον αλγόριθμο ομαδοποίησης KMeans. Ο αλγόριθμος έχει οριστεί να κατηγοριοποιεί τους πελάτες σε συστάδες, με τον αριθμό των συστάδων να καθορίζεται παρατηρώντας τον "αγκώνα" στο γράφημα του αθροίσματος των τετραγώνων εντός της συστάδας σε σχέση με τον αριθμό των συστάδων. Το γράφημα αυτό υποδεικνύει τον βέλτιστο αριθμό συστάδων για την τμηματοποίηση.



Εικόνα 4.12 Elbow method για την ομαδοποίηση με βάση το Recency score

Μόλις οι πελάτες χωριστούν σε ομάδες, οι πληροφορίες που προκύπτουν από τις περιγραφές των ομάδων αποκαλύπτουν τα χαρακτηριστικά κάθε ομάδας. Πριν από την παραγγελία, η συστάδα 0, για παράδειγμα, έχει πελάτες οι οποίοι, κατά μέσο όρο, πραγματοποίησαν μια αγορά πριν από περίπου 66,89 ημέρες, υποδεικνύοντας μια σχετικά πρόσφατη εμπλοκή με την επιχείρηση. Καθώς αυξάνεται ο αριθμός της συστάδας, αυξάνεται και ο μέσος όρος της παλαιότητας, αναδεικνύοντας ομάδες πελατών των οποίων η τελευταία αγορά ήταν πιο μακριά στο παρελθόν. Για παράδειγμα, η συστάδα 1 έχει μέση αναδρομικότητα περίπου 320,13 ημέρες, γεγονός που υποδηλώνει ότι αυτοί οι πελάτες είναι λιγότερο πρόσφατοι.

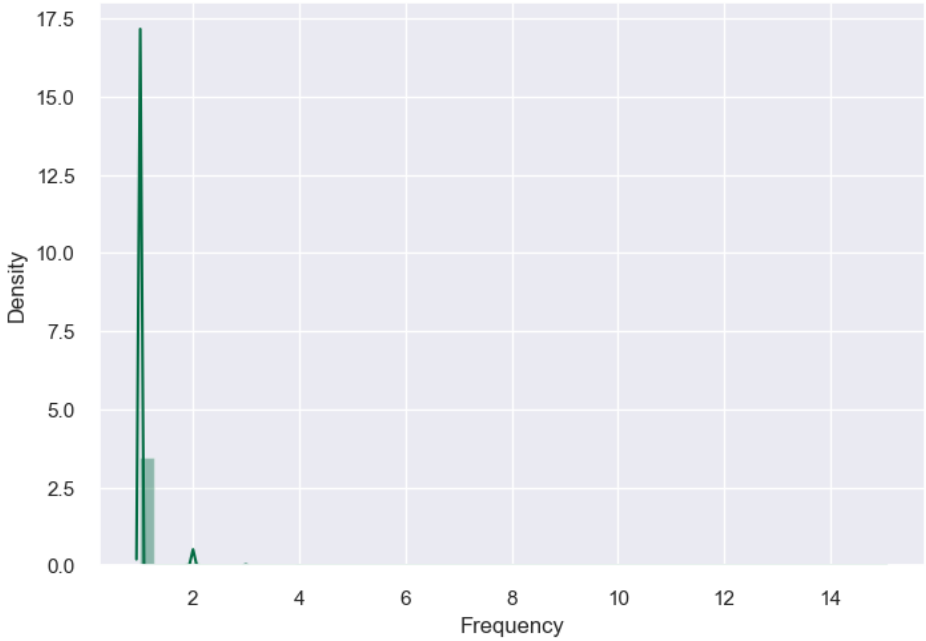
	count	mean	std	min	25%	50%	75%	max
RecencyCluster								
0	25519.0	66.894.314	35.896.129	0.0	34.0	67.0	100.0	127.0
1	24428.0	320.126.944	43.809.601	255.0	282.0	313.0	357.0	405.0
2	27057.0	188.386.074	35.282.423	128.0	158.0	188.0	219.0	254.0
3	16392.0	490.349.317	59.099.382	406.0	444.0	481.0	532.0	729.0

Πίνακας 4 Περιγραφικά στατιστικά των συστάδων του Recency

Το τελικό μέρος της ανάλυσης της μετρικής «Recency», περιλαμβάνει την αναδιάταξη των συστάδων ώστε να διασφαλιστεί ότι οι υψηλότεροι αριθμοί συστάδων αντιστοιχούν σε πιο πρόσφατες αγορές. Αυτή η ταξινόμηση επιτρέπει τη διαισθητική ερμηνεία των συστάδων, όπου, για παράδειγμα, ένας πελάτης στη συστάδα 3 έχει πραγματοποιήσει μια αγορά πιο πρόσφατα από εκείνους στις συστάδες 0 έως 2. Αυτή η αναδιάταξη παρέχει μια σαφέστερη εικόνα της δέσμευσης των πελατών, επιτρέποντας στην επιχείρηση να προσαρμόσει κατάλληλα τις στρατηγικές μάρκετινγκ. Οι πελάτες στις υψηλότερες συστάδες θα μπορούσαν να αποτελέσουν στόχο για πρωτοβουλίες δέσμευσης, ώστε να αξιοποιηθούν οι πρόσφατες αλληλεπιδράσεις τους, ενώ οι πελάτες στις χαμηλότερες συστάδες θα μπορούσαν να προσεγγιστούν με εκστρατείες επανενεργοποίησης. Αυτές οι γνώσεις είναι ζωτικής σημασίας για την ανάπτυξη στοχευμένων στρατηγικών μάρκετινγκ και τη διατήρηση των πελατειακών σχέσεων. Οι συστάδες επαναφοράς μπορούν να καθοδηγήσουν τον τρόπο με τον οποίο η επιχείρηση δίνει προτεραιότητα στις προσπάθειες δέσμευσης πελατών, εστιάζοντας σε εκείνους που είναι πιο πιθανό να ανταποκριθούν με βάση την πρόσφατη δραστηριότητά τους.

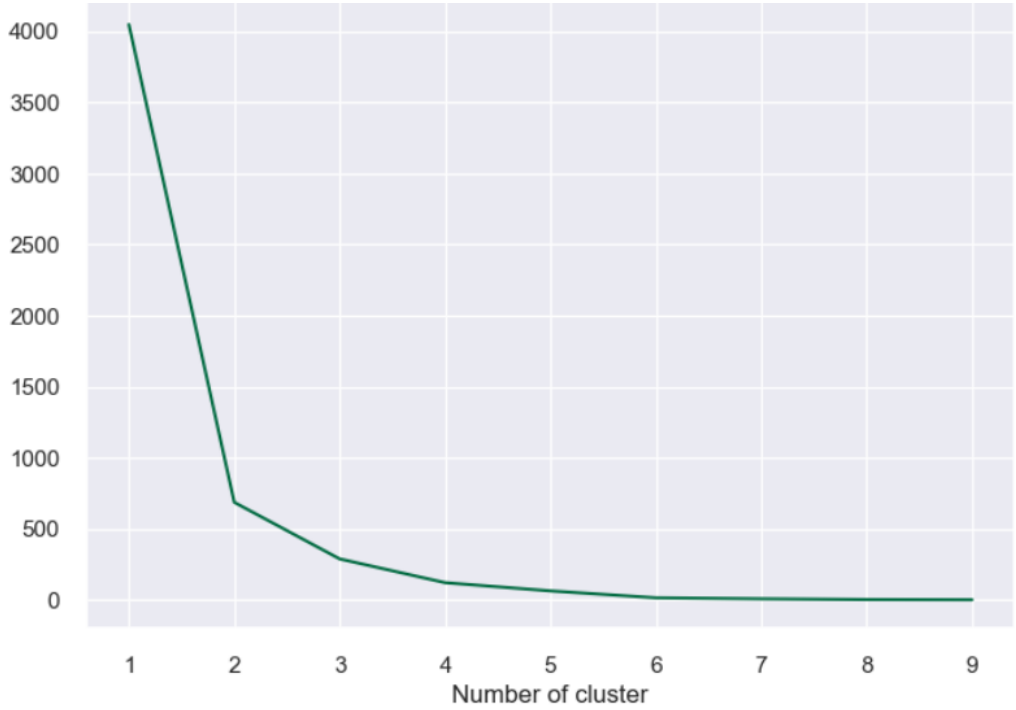
Συνεχίζοντας με τον υπολογισμό της μετρικής «Frequency», η ανάλυση μετρά τον αριθμό των μοναδικών παραγγελιών που πραγματοποιεί κάθε πελάτης. Η κατανομή αυτής της βαθμολογίας είναι εξαιρετικά λοξή, με μια κορυφή στο κατώτερο

άκρο, γεγονός που υποδηλώνει ότι η πλειονότητα των πελατών πραγματοποιεί αγορές σπάνια.



Εικόνα 4.13 Κατανομή Frequency score

Η επόμενη φάση περιλαμβάνει την ομαδοποίηση των πελατών με βάση αυτές τις τιμές συχνότητας, σε συστάδες που αντικατοπτρίζουν τα διαφορετικά επίπεδα αγοραστικής δραστηριότητας.



Εικόνα 4.14 Elbow method για την ομαδοποίηση με βάση το Frequency score

Αφού καθοριστεί ο κατάλληλος αριθμός συστάδων, εφαρμόζεται ο αλγόριθμος KMeans για την κατάτμηση των πελατών σε τέσσερις διαφορετικές ομάδες. Η τμηματοποίηση αποκαλύπτει σημαντικές πληροφορίες για τα αγοραστικά πρότυπα των πελατών. Για παράδειγμα, η συντριπτική πλειονότητα των πελατών ανήκει στη συστάδα 0, υποδεικνύοντας ότι έχουν πραγματοποιήσει μόνο μία αγορά. Αυτή η συστάδα έχει μέση βαθμολογία συχνότητας ακριβώς 1, γεγονός που αντανακλά μια εφάπαξ αγοραστική συμπεριφορά. Καθώς προχωράμε σε υψηλότερες συστάδες, η συχνότητα αυξάνεται, αν και με μικρότερο αριθμό πελατών σε κάθε επόμενη συστάδα. Η συστάδα 1, με μέσο σκορ συχνότητας 2, καταγράφει όσους έχουν αγοράσει δύο φορές. Η μέση τιμή συχνότητας της συστάδας 2, περίπου 3,22, υποδηλώνει ότι αυτοί οι πελάτες έχουν ασχοληθεί με την επιχείρηση πιο συχνά, αν και εξακολουθούν να έχουν χαμηλό αριθμό συναλλαγών. Η τελευταία συστάδα, η συστάδα 3, αν και πολύ μικρή σε μέγεθος, έχει μέση συχνότητα περίπου 7,88, αντιπροσωπεύοντας τους πιο πιστούς πελάτες που έχουν αγοράσει πολλές φορές.

	count	mean	std	min	25%	50%	75%	max
FrequencyCluster								
0	90589.0	1.000.000	0.000000	1.0	1.0	1.0	1.0	1.0
1	2581.0	2.000.000	0.000000	2.0	2.0	2.0	2.0	2.0
2	218.0	3.220.183	0.505414	3.0	3.0	3.0	3.0	5.0
3	8.0	7.875.000	3.044.316	6.0	6.0	7.0	7.5	15.0

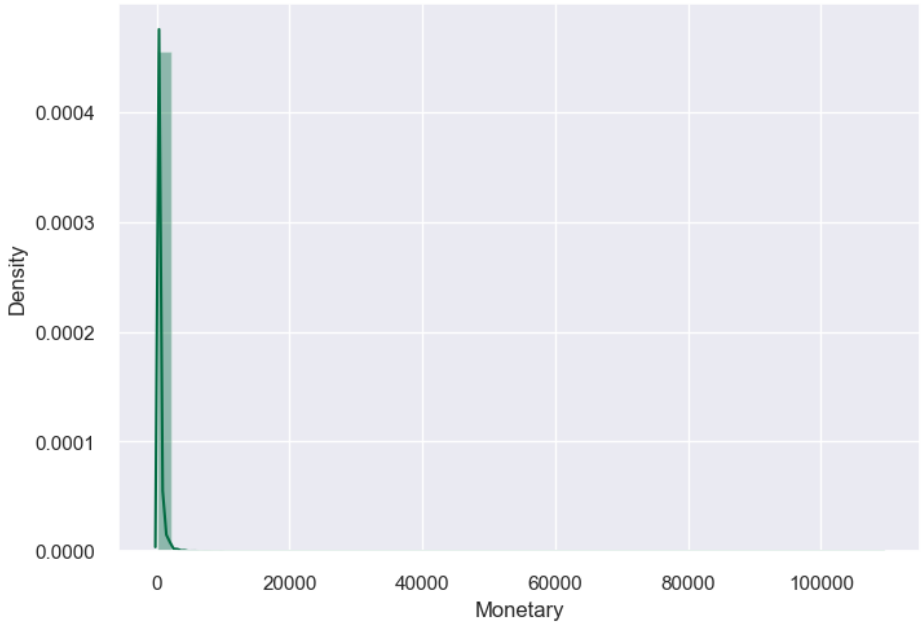
Πίνακας 5 Περιγραφικά στατιστικά των συστάδων του Frequency

Αυτές οι ομάδες συχνοτήτων είναι ενδεικτικές του διαφορετικού βαθμού αφοσίωσης των πελατών, με τη συντριπτική πλειοψηφία να είναι αγοραστές που αγοράζουν μία φορά, ένα μικρότερο τμήμα να ψωνίζει περιστασιακά και μια μικρή ομάδα να επιδεικνύει ισχυρή αφοσίωση μέσω των επαναλαμβανόμενων αγορών τους. Αυτή η διαστρωμάτωση επιτρέπει στην επιχείρηση να προσαρμόσει αποτελεσματικά τις στρατηγικές δέσμευσής της - για παράδειγμα, δίνοντας κίνητρα στους πελάτες των χαμηλότερων συστάδων να κάνουν πρόσθετες αγορές και αναγνωρίζοντας και ανταμείβοντας εκείνους που βρίσκονται στην υψηλότερη συστάδα για την αφοσίωσή τους.

Τέλος, υπολογίζεται η μετρική «Monetary» που αξιολογεί το συνολικό ποσό των χρημάτων που ξοδεύουν οι πελάτες, παρέχοντας πληροφορίες για την αξία που προσφέρει κάθε πελάτης στην επιχείρηση. Αρχικά, μετράται η συνολική χρηματική αξία που έχει συνεισφέρει κάθε πελάτης, η οποία είναι το άθροισμα των αξιών

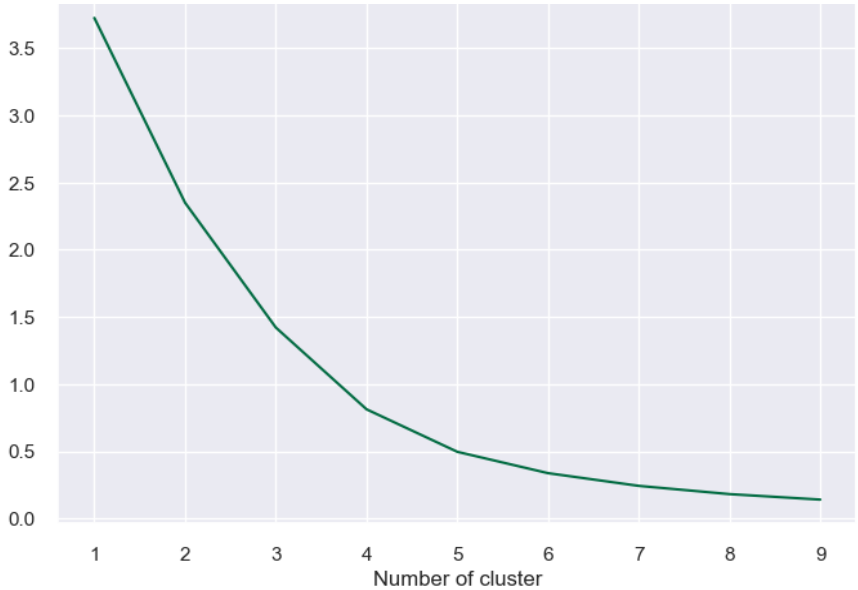
πληρωμής σε όλες τις συναλλαγές του. Ο αριθμός αυτός χρησιμεύει ως άμεσος δείκτης της οικονομικής αξίας του πελάτη για την επιχείρηση.

Η βαθμολογία Monetary, παρουσιάζει επίσης μια εξαιρετικά λοξή κατανομή, με τους περισσότερους πελάτες να συνεισφέρουν μικρότερη χρηματική αξία και λίγους πελάτες να συνεισφέρουν σημαντικά μεγαλύτερη. Αυτό το μοτίβο αντικατοπτρίζει συχνά την αρχή Pareto ή τον κανόνα 80/20, όπου ένα μικρό κλάσμα πελατών αντιπροσωπεύει ένα μεγάλο μέρος των εσόδων.



Εικόνα 4.15 Κατανομή Monetary score

Στη συνέχεια, οι πελάτες ομαδοποιούνται χρησιμοποιώντας τον αλγόριθμο ομαδοποίησης KMeans, παρόμοια με τα προηγούμενα βήματα.



Εικόνα 4.16 Elbow method για την ομαδοποίηση με βάση το Monetary score

	count	mean	std	min	25%	50%	75%	max
MonetaryCluster								
0	91850.0	170.337.573	182.074.594	9.59	63270	111020	195.93	1279.10
1	1514.0	2.387.275.634	1.430.423.876	1280.52	1500240	1881485	2631.32	9859.88
2	31.0	18.076.726.129	8.142.027.413	10999.26	11813005	14963640	19399.65	44048.00
3	1.0	109.312.640.000	NaN	109312.64	109312640	109312640	109312.64	109312.64

Πίνακας 6 Περιγραφικά στατιστικά των συστάδων του Monetary

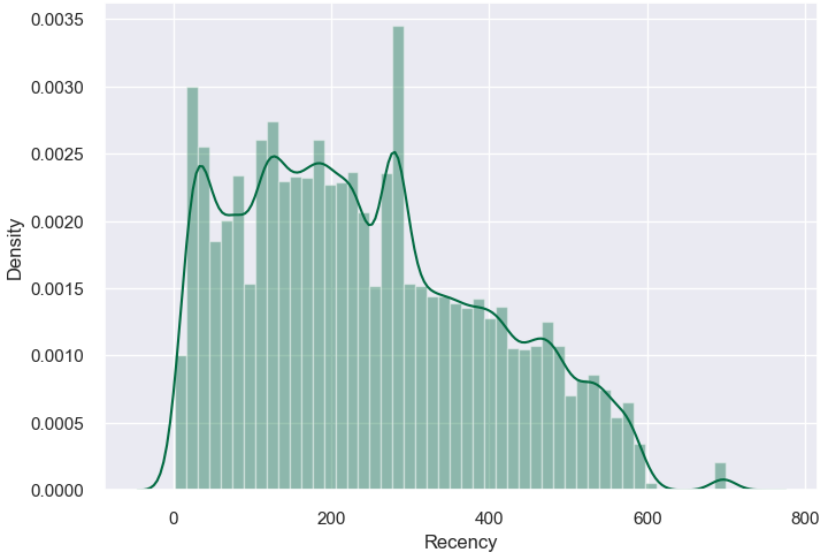
Τα αποτελέσματα δείχνουν ότι η μεγαλύτερη συστάδα, η συστάδα 0, αποτελείται από πελάτες με χαμηλότερη μέση δαπάνη, υποδηλώνοντας το τμήμα που συνεισφέρει μέτρια από άποψη χρηματικής αξίας. Η μέση δαπάνη σε αυτή τη συστάδα είναι περίπου 170,34 δολάρια, γεγονός που υποδηλώνει ότι πρόκειται για συναλλαγές χαμηλής αξίας κατά μέσο όρο. Ανεβαίνοντας προς τα πάνω στις συστάδες, η μέση χρηματική αξία αυξάνεται σημαντικά. Η συστάδα 1 παρουσιάζει υψηλότερη μέση δαπάνη περίπου 2387,28 δολάρια, υποδηλώνοντας μια ομάδα πελατών που συμβάλλουν πιο ουσιαστικά στα έσοδα της εταιρείας. Η συστάδα 2, με μέση δαπάνη περίπου 18076,72 δολάρια, αντιπροσωπεύει μια ακόμη μικρότερη ομάδα πελατών που πιθανότατα περιλαμβάνει «premium» αγοραστές ή όσους πραγματοποιούν αγορές υψηλής αξίας. Τέλος, η συστάδα 3, αν και με ένα μόνο σημείο δεδομένων, παρουσιάζει μια εξαιρετική μέση δαπάνη 109312,64 δολάρια, η οποία θα μπορούσε να αντιπροσωπεύει έναν ακραίο πελάτη ή έναν πελάτη με εξαιρετικά υψηλή αξία συναλλαγών.

Οι νομισματικές ομάδες επιτρέπουν στην επιχείρηση να εντοπίσει ποιοι πελάτες παράγουν τα περισσότερα έσοδα και να χαράξει ανάλογη στρατηγική. Στους πελάτες υψηλής αξίας θα μπορούσαν να παρέχονται εξατομικευμένες υπηρεσίες ή ανταμοιβές επιβράβευσης για να διατηρηθεί η δέσμευσή τους, ενώ θα μπορούσαν επίσης να αναπτυχθούν στρατηγικές για την αύξηση των δαπανών των πελατών χαμηλότερης αξίας. Αυτή η κλιμακωτή κατανόηση της χρηματικής συνεισφοράς είναι ζωτικής σημασίας για την κατανομή των πόρων, την ιεράρχηση των τμημάτων πελατών και την προσαρμογή των προσπαθειών μάρκετινγκ για τη μεγιστοποίηση της κερδοφορίας.

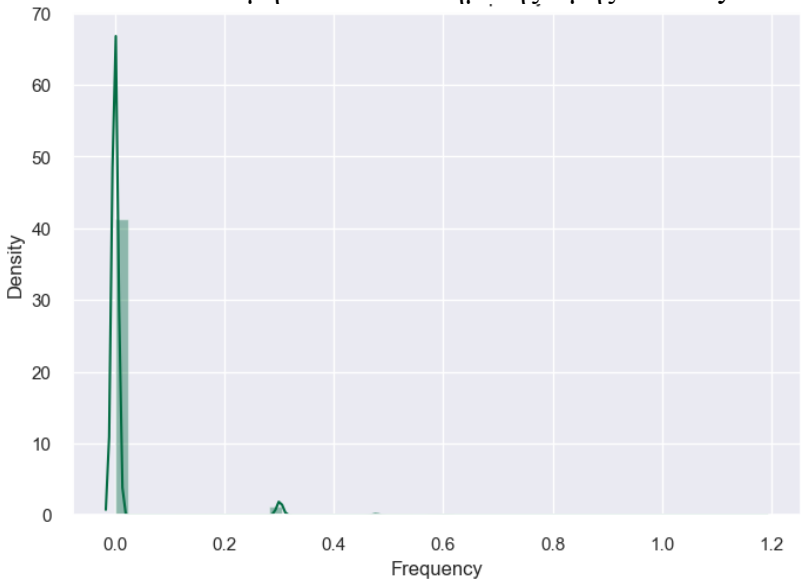
Μετά την διερεύνηση των βαθμολογιών Recency, Frequency, Monetary και την τμηματοποίηση του δείγματος με καθεμία από αυτές τις μετρικές, το τελευταίο βήμα στην ανάλυση του μοντέλου RFM είναι ο συνδυασμός αυτών των τριών μετρικών σε

ένα τελικό RFM score και η τμηματοποίηση του δείγματος με βάση αυτή την συνολική βαθμολογία.

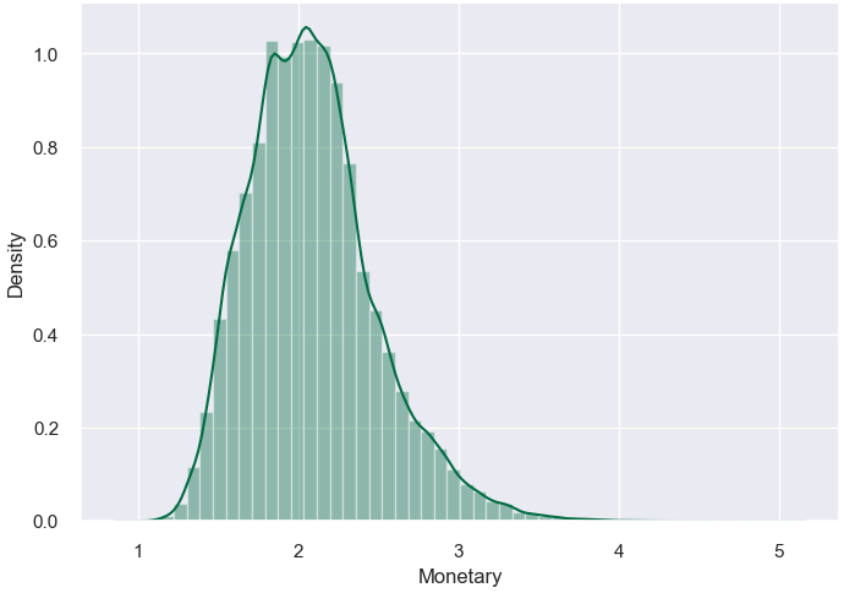
Όπως διερευνήσαμε παραπάνω, η λοξότητα στην κατανομή των βαθμολογιών RFM απαιτεί συχνά μετασχηματισμό για την κανονικοποίηση των δεδομένων, καθιστώντας τα πιο κατάλληλα για αλγορίθμους ομαδοποίησης όπως ο KMeans. Ο μετασχηματισμός λογαρίθμου είναι μια κοινή προσέγγιση για τη μείωση της λοξότητας, το αντίστοιχο βήμα έχει ακολουθηθεί σε πολλές αντίστοιχες μελέτες στο παρελθόν (Gustriansyah et al., 2019), σε αυτή την μελέτη διερευνήσαμε την κατανομή χρησιμοποιώντας ένα ιστόγραμμα που επικαλύπτεται από μια εκτίμηση πυκνότητας πυρήνα (KDE), μια άλλη αντίστοιχα αποτελεσματική μέθοδος είναι και το διάγραμμα διασποράς 3D, που ακολούθησαν οι Gustriansyah κ.α..



Εικόνα 4.17 Κατανομή κανονικοποιημένης τιμής Recency score



Εικόνα 4.18 Κατανομή κανονικοποιημένης τιμής Frequency score



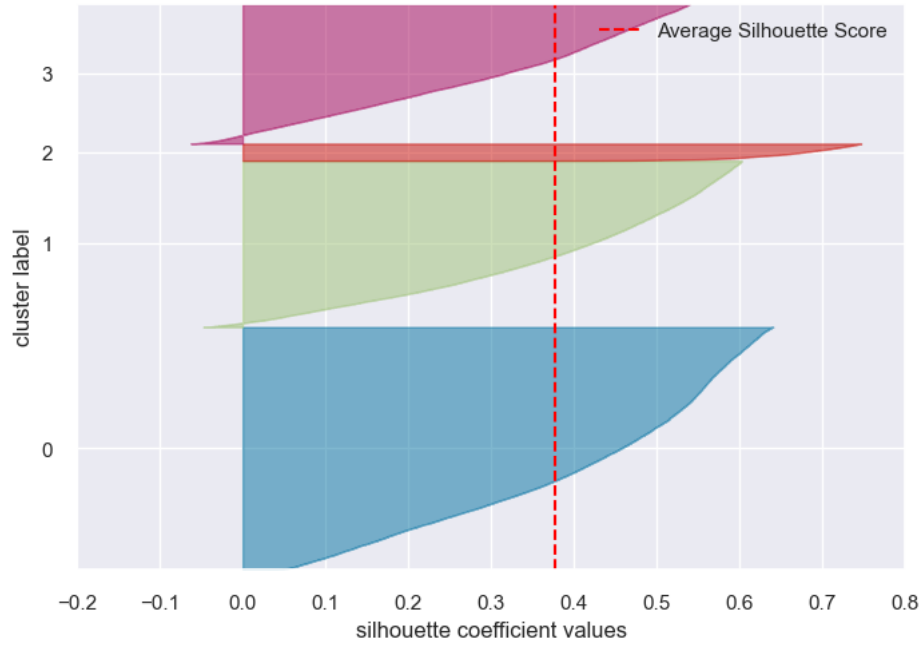
Εικόνα 4.19 Κατανομή κανονικοποιημένης τιμής Monetary score

Στο τελικό σύνολο δεδομένων RFM, ο συνδυασμός των βαθμολογιών Recency, Frequency και Monetary για κάθε πελάτη δίνει μια πολυδιάστατη εικόνα της συμπεριφοράς των πελατών. Οι πελάτες με χαμηλές βαθμολογίες Recency (πιο πρόσφατες αγορές), υψηλές βαθμολογίες Frequency (πιο συχνές αγορές) και υψηλές βαθμολογίες Monetary (υψηλότερες δαπάνες) θεωρούνται συνήθως οι πιο πολύτιμοι. Αντίθετα, οι πελάτες με υψηλές βαθμολογίες Recency, χαμηλές βαθμολογίες Frequency και χαμηλές νομισματικές βαθμολογίες μπορεί να θεωρηθούν ότι διατρέχουν κίνδυνο ή ότι έχουν παραγραφεί.

Με την εξέταση αυτών των κατανομών πριν και μετά τον μετασχηματισμό σε λογάριθμο, η επιχείρηση μπορεί να κατανοήσει καλύτερα τους φυσικούς λογαρίθμους της συμπεριφοράς των πελατών και να λάβει πιο τεκμηριωμένες αποφάσεις σχετικά με τις στρατηγικές μάρκετινγκ, την τμηματοποίηση των πελατών και την κατανομή των πόρων. Τα μετασχηματισμένα δεδομένα είναι επίσης πιο πρόσφορα για ομαδοποίηση και άλλες τεχνικές μηχανικής μάθησης που μπορούν να τμηματοποιήσουν περαιτέρω την πελατειακή βάση σε ομάδες με νόημα για στοχευμένες εκστρατείες μάρκετινγκ και διαχείριση πελατειακών σχέσεων.

Επομένως, το πρώτο βήμα για τον υπολογισμό του RFM score είναι η κανονικοποίηση των τιμών γύρω από μια μέση τιμή μηδέν και μια τυπική απόκλιση ένα. Αυτό είναι ένα κρίσιμο βήμα για την προετοιμασία των δεδομένων για την ομαδοποίηση, καθώς το KMeans είναι ευαίσθητο στην κλίμακα των δεδομένων. Τα χαρακτηριστικά γνωρίσματα πρέπει να βρίσκονται στην ίδια κλίμακα, ώστε το καθένα να συμβάλλει εξίσου στους υπολογισμούς της απόστασης.

Έπειτα, η ανάλυση σιλουέτας όπως παρουσιάζεται στην επόμενη εικόνα, αξιολογεί τη ποιότητα των συστάδων που σχηματίζονται από τον αλγόριθμο KMeans. Ο συντελεστής σιλουέτας μετρά πόσο παρόμοιο είναι ένα αντικείμενο με τη δική του συστάδα (cohesion) σε σύγκριση με άλλες συστάδες (separation). Το διάγραμμα σιλουέτας εμφανίζει μια μέτρηση του πόσο κοντά βρίσκεται κάθε σημείο μιας συστάδα στα σημεία των άλλων κοντινών συστάδων. Αυτός ο συντελεστής κυμαίνεται από -1 έως +1, όπου μια υψηλή τιμή δείχνει ότι το αντικείμενο ταιριάζει καλά με τη δική του συστάδα και ελάχιστα με τις γειτονικές συστάδες. Εν προκειμένω, το διάγραμμα σιλουέτας υποδεικνύει την παρουσία τεσσάρων συστάδων (όπως υποδηλώνουν τα τέσσερα διαφορετικά χρώματα), και η μέση βαθμολογία σιλουέτας (σημειωμένη με την κόκκινη διακεκομμένη γραμμή) παρέχει μια ένδειξη της συνολικής προσαρμογής. Εάν η βαθμολογία είναι πιο κοντά στο +1, αυτό υποδηλώνει καλή προσαρμογή σε συστάδες.



Εικόνα 4.20 Υπολογισμός Silhouette score για την RFM ομαδοποίηση

Αυτό που μπορούμε να συμπεράνουμε από το διάγραμμα σιλουέτας είναι ότι δεν έχουν όλες οι συστάδες το ίδιο επίπεδο συνοχής και διαχωρισμού. Το πλάτος των διαγραμμάτων σιλουέτας για κάθε συστάδα δεν είναι ομοιόμορφο, γεγονός που υποδηλώνει ποικιλομορφία στο πόσο καλά ορίζεται κάθε συστάδα. Η μέση βαθμολογία σιλουέτας, η οποία φαίνεται να πλησιάζει το 0,4, υποδηλώνει μια δίκαιη και σχετικά ισχυρή, δομή ομαδοποίησης.

Συγκρίνοντας το παραπάνω διάγραμμα, αποτέλεσμα της έρευνάς μας με άλλες μελέτες της βιβλιογραφίας, παρατηρούμε ότι βρίσκεται πάνω από το μέσο όρο και

θεωρείται αποτελεσματική ομαδοποίηση. Σε μια μελέτη (Christy et al., 2021), όπου εξετάστηκαν συγκριτικά τρεις μέθοδοι συσταδοποίησης των RFM σκορ, K-means, RM K-means και Fuzzy C-means, ο κλασσικός αλγόριθμος K-means είχε μέσο πλάτος σιλουέτας 0.38, ενώ ο Fuzzy C-means μόλις 0.06.

Με βάση το συνδυασμό των μετρικών προχωρήσαμε στην τμηματοποίηση του δείγματος σε 4 ομάδες. Με την βοήθεια μιας προσαρμοσμένης συνάρτησης που αθροίζει τη μέση τιμή του «Recency», τη τη μέση τιμή του «Frequency» και τη μέση τιμή και τον αριθμό της «Monetary» τιμής για κάθε συστάδα. Αυτή η σύνοψη δίνει τις ακόλουθες πληροφορίες για τις τελικές συστάδες RFM:

- 0: αντιπροσωπεύει τη μεγαλύτερη ομάδα πελατών που είναι μέτρια πρόσφατοι, λιγότερο συχνοί και ξοδεύουν τα λιγότερα κατά μέσο όρο. (count:40090)
- 1: πελάτες που είναι οι λιγότερο πρόσφατοι, έχουν παρόμοια χαμηλή συχνότητα, αλλά ξοδεύουν περισσότερα κατά μέσο όρο από εκείνους της συστάδας 0. (count:27455)
- 2: είναι πιο πρόσφατοι και πιο συχνοί από εκείνους των συστάδων 0 και 1 και ξοδεύουν πολύ περισσότερα κατά μέσο όρο. (count:2807)
- 3: αν και δεν είναι οι πιο πρόσφατοι καταναλωτές, έχουν την υψηλότερη μέση δαπάνη, υποδεικνύοντας ότι οι πελάτες αυτοί πραγματοποιούν σημαντικές αγορές, ακόμη και αν το κάνουν λιγότερο συχνά. (count:23044)

	Recency	Frequency	Monetary	
	mean	mean	mean	count
Cluster				
0	146.0	1.0	81.0	40090
1	426.0	1.0	124.0	27455
2	226.0	2.0	488.0	2807
3	195.0	1.0	516.0	23044

Πίνακας 7 Χαρακτηριστικά RFM ομαδοποίησης δείγματος

Τέλος, ένα ακόμη βήμα για την ενίσχυση της RFM ανάλυσης, είναι ο υπολογισμός ενός «OverallScore» με το άθροισμα των επιμέρους βαθμολογιών συστάδων RFM. Αυτή η βαθμολογία χρησιμοποιείται στη συνέχεια για την ομαδοποίηση των δεδομένων και τον υπολογισμό του μέσου όρου των αρχικών μετρικών RFM για κάθε "OverallScore". Το αποτέλεσμα στον παρακάτω πίνακα παρουσιάζει τις μέσες τιμές Recency, Frequency και Monetary για κάθε "OverallScore", οι οποίες μπορούν να παρέχουν πληροφορίες σχετικά με τη

συμπεριφορά των πελατών σε κάθε συνδυασμένη ομάδα αποτελεσμάτων. Ο πίνακας δίνει τις ακόλουθες πληροφορίες:

- Οι πελάτες με OverallScore = 0 έχουν τον υψηλότερο μέσο όρο Recency (πιο πρόσφατα), τον χαμηλότερο μέσο όρο Frequency (λιγότερο συχνά) και σχετικά χαμηλότερο μέσο όρο Monetary (χαμηλότερη κατανάλωση),
- καθώς το OverallScore αυξάνεται, ο μέσος όρος Recency μειώνεται (λιγότερο πρόσφατα), ο μέσος όρος Frequency αυξάνεται ελαφρώς (πιο συχνά) και ο μέσος όρος Monetary αυξάνεται επίσης (υψηλότερη κατανάλωση),
- ενώ, το υψηλότερο OverallScore αντιπροσωπεύει τους πελάτες που είναι οι πιο πρόσφατοι, οι πιο συχνοί και οι πιο δαπανηροί.

Έπειτα, με βάση το 'OverallScore' υπολογίσαμε μια νέα μεταβλητή που ονομάζεται «Segment» και ταξινομεί τους πελάτες σε τμήματα 'Low-Value', 'Mid-Value' και 'High-Value' ανάλογα με το 'OverallScore' τους.

	Recency	Frequency	Monetary
OverallScore			
0	490.418.062	1.000.000	163.808.291
1	324.114.339	1.014.298	191.818.130
2	193.851.622	1.026.075	201.251.499
3	72.919.377	1.034.334	208.862.058
4	79.366.878	1.726.624	1.331.961.276
5	78.198.198	2.801.802	2.448.332.162
6	84.714.286	4.476.190	2.460.022.381
7	43.666.667	5.666.667	10.924.123.333

Πίνακας 8 Χαρακτηριστικά διαχωρισμού με βάση το Overall Score

Ωστόσο, εξετάζοντας την βιβλιογραφία, παρατηρούμε ότι στο παρελθόν, έχουν διατυπωθεί και άλλες μέθοδοι υπολογισμού του συνολικού RFM σκορ, σε μία από αυτές μάλιστα (Christy et al., 2021), αντί να επιλέγονται τυχαία τα σημεία εκκίνησης (κεντροειδή) για τις συστάδες, η μέθοδος αυτή υπολογίζει τη διάμεση τιμή κάθε ταξινομημένης λίστας (R', F' και M'). Αυτές οι διάμεσες τιμές χρησιμοποιούνται στη συνέχεια ως τα αρχικά κεντροειδή για τη διαδικασία ομαδοποίησης. Τα πλεονεκτήματα αυτής της προσέγγισης είναι κυρίως οι μειωμένες επαναλήψεις και ο μικρός υπολογιστικός χρόνος. Με την επιλογή αρχικών κεντροειδών που είναι πιο αντιπροσωπευτικά της κατανομής των δεδομένων (μέσω των διαμέσων), ο αλγόριθμος χρειάζεται λιγότερες επαναλήψεις. Αυτό καθιστά τη διαδικασία ταχύτερη σε σύγκριση με την παραδοσιακή μέθοδο όπου τα κεντροειδή επιλέγονται τυχαία. Ενώ, ταυτόχρονα,

οι συστάδες που σχηματίζονται με αυτή τη μέθοδο παρατηρείται ότι είναι πιο ουσιαστικές και κατάλληλες.

Η πολυπλοκότητα αυτού του τροποποιημένου αλγορίθμου K-Means, που ονομάζεται RM K-Means, παραμένει η ίδια με τον παραδοσιακό αλγόριθμο K-Means, η οποία είναι O(n + k + i), όπου n είναι ο αριθμός των σημείων δεδομένων, k είναι ο αριθμός των συστάδων και i είναι ο αριθμός των επαναλήψεων. Ωστόσο, παρά την ίδια θεωρητική πολυπλοκότητα, ο πρακτικός αριθμός των απαιτούμενων επαναλήψεων μειώνεται, γεγονός που ενισχύει την αποδοτικότητα (Christy et al., 2021).

4.3 ΑΞΙΑ ΔΙΑΡΚΕΙΑΣ ΖΩΗΣ ΠΕΛΑΤΗ

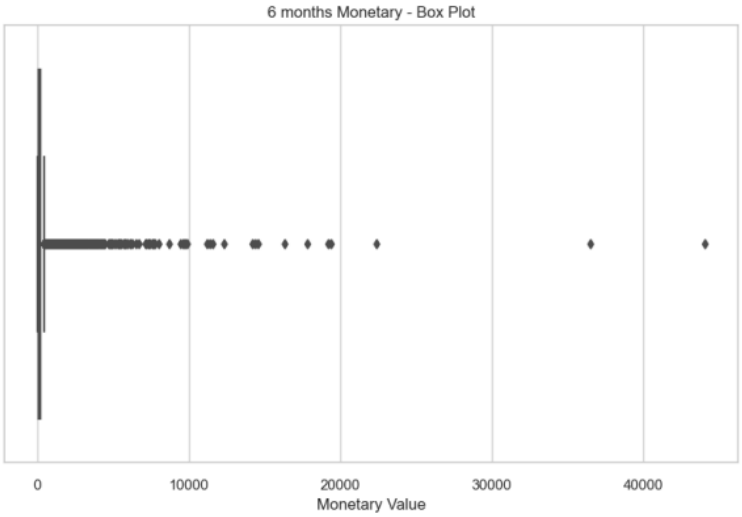
Η αξία διάρκειας ζωής πελάτη - CLV υπολογίζεται χρησιμοποιώντας τον τύπο:

$$CLV=(Monetary \times Frequency) / Recency$$

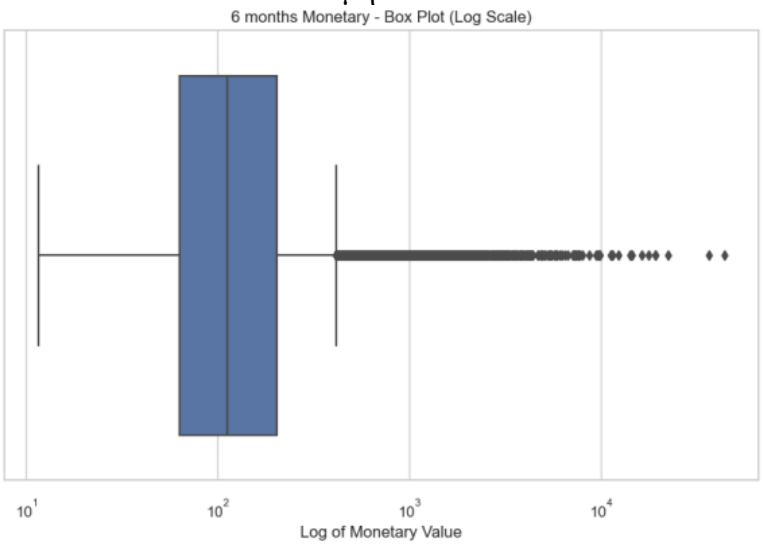
Αυτός ο τύπος υποθέτει ότι η αξία ενός πελάτη είναι υψηλότερη αν ξοδεύει περισσότερα (Monetary), αν το κάνει πιο συχνά (Frequency) και αν το έχει κάνει πιο πρόσφατα (Recency). Οι προκύπτουσες τιμές CLV προστίθενται στη συνέχεια στο σύνολο δεδομένων της ερευνάς. Επομένως το νέο πλαίσιο δεδομένων, το οποίο περιλαμβάνει τις αρχικές τιμές RFM μαζί με το πρόσφατα υπολογισμένο CLV, επιτρέπει την ανάλυση του CLV σε σχέση με τις μετρήσεις RFM.

Πολυάριθμα πρόσφατα ερευνητικά έργα έχουν εξετάσει και αντιπαραβάλει τις επιδόσεις των μοντέλων αξίας διάρκειας ζωής του πελάτη (CLV) με εκείνες των μοντέλων συχνότητας, συχνότητας, νομισματικής (RFM), καταλήγοντας ότι τα μοντέλα CLV υπερτερούν των μοντέλων RFM. Για παράδειγμα, σε μια μελέτη (Reinartz & Kumar, 2003) ανέλυσαν δεδομένα από περίπου 12.000 πελάτες μιας εταιρείας ταχυδρομικών παραγγελιών για μια περίοδο τριών ετών. Τα συμπεράσματά τους έδειξαν ότι το κορυφαίο 30% των πελατών που προσδιορίστηκαν από το μοντέλο CLV απέφερε 33% περισσότερα έσοδα από το κορυφαίο 30% που προσδιορίστηκε από το μοντέλο RFM. Μια άλλη μελέτη (Venkatesan & Kumar, 2004), αξιολόγησε διάφορα μοντέλα που χρησιμοποιούνται για την επιλογή πελατών. Χρησιμοποίησαν δεδομένα από περίπου 2.000 πελάτες που ανήκαν σε μια B2B επιχείρηση. Τα αποτελέσματα έδειξαν ότι το πιο κερδοφόρο 5% των πελατών, όπως προσδιορίστηκε από το μοντέλο CLV, παρήγαγε κέρδη που ήταν 10% έως 50% μεγαλύτερα από εκείνα που επιλέχθηκαν από άλλα μοντέλα, συμπεριλαμβανομένων των μοντέλων RFM και παρελθοντικής αξίας.

Στην δική μας έρευνα του ηλεκτρονικού marketplace, δημιουργήθηκε ένα υποσύνολο πελατών (Κεφάλαιο 3.1.3) που έχουν πραγματοποιήσει αγορές εντός συγκεκριμένων χρονικών πλαισίων (3 μήνες και 6 μήνες) και αθροίζοντας οι χρηματικές τους αξίες επιχειρείται η κατανόηση της συμπεριφοράς τους, εντός αυτών των περιόδων. Παρακάτω βλέπουμε την απεικόνιση της κατανομής των νομισματικών αξιών για τους τελευταίους 6 μήνες (Εικόνα 4.21). Στο διάγραμμα παρατηρούμε ότι υπάρχει σημαντικός αριθμός ακραίων τιμών, οι οποίες προκαλούν την έκταση της κλίμακας. Αυτή η λοξότητα είναι χαρακτηριστική στα νομισματικά δεδομένα, όπου οι περισσότεροι πελάτες μπορεί να ξοδεύουν μικρά ποσά, ενώ λίγοι ξοδεύουν σημαντικά περισσότερο. Για αυτό το λόγο, το Boxplot σε λογαριθμική κλίμακα (Εικόνα 4.22) προσφέρει μια σαφέστερη εικόνα.



Εικόνα 4.21 Boxplot Νομισματικής αξίας πελατών την περίοδο των τελευταίων 6 μηνών



Εικόνα 4.22 Boxplot Λογαριθμικής Νομισματικής αξίας πελατών την περίοδο των τελευταίων 6 μηνών

Το πλεονέκτημα του διαγράμματος σε λογαριθμική κλίμακα, σε σχέση με το προηγούμενο είναι ότι μειώνει τον οπτικό αντίκτυπο των ακραίων τιμών, επιτρέποντάς την απεικόνιση της κατανομής των δεδομένων με μεγαλύτερη σαφήνεια. Σε αυτό το διάγραμμα, το μεγαλύτερο μέρος των δεδομένων βρίσκεται μεταξύ 10^2 και 10^3, πράγμα που σημαίνει ότι οι περισσότερες χρηματικές αξίες είναι μεταξύ 100 και 1.000 δολαρίων. Υπάρχουν ακραίες τιμές μέχρι περίπου 10^4, υποδεικνύοντας ότι υπάρχουν χρηματικές αξίες μέχρι και 10.000 δολάρια, οι οποίες όμως είναι αρκετά σπάνιες σε σύγκριση με τα υπόλοιπα δεδομένα.

Επομένως, με βάση τα παραπάνω συμπεραίνουμε ότι είναι αναγκαία και η αφαίρεση ακραίων τιμών. Μετά την αφαίρεση αυτών, δημιουργήθηκαν ομάδες με βάση αυτές τις τιμές. Παρακάτω (Πίνακας 9) παρατηρούμε τα περιγραφικά στατιστικά στοιχεία για κάθε συστάδα με βάση την CLV.

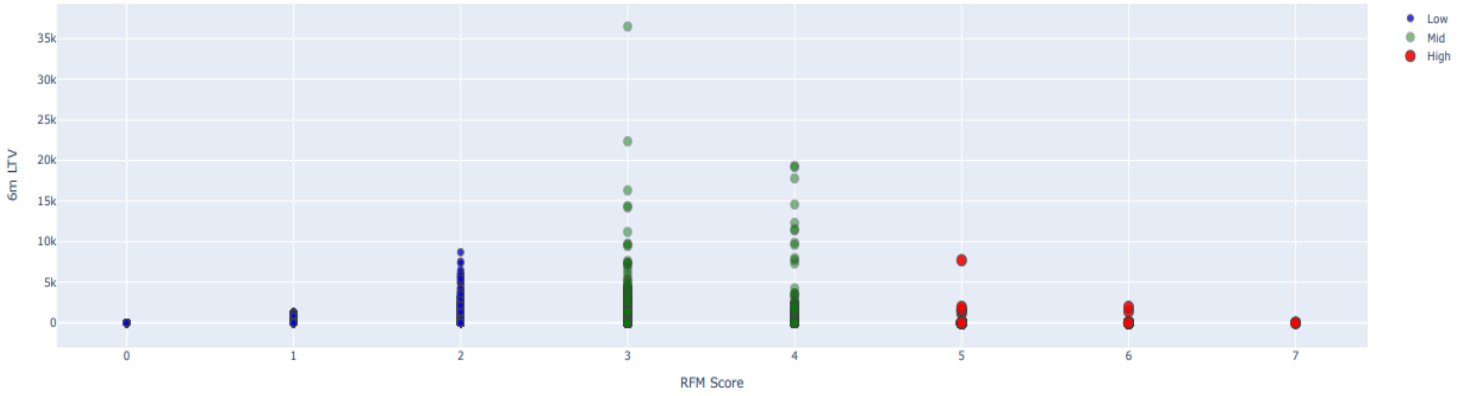
	count	***mean***	***std***	***min***	***0,25***	***0,5***	***0,75***	***max***
LTVCluster								
0	72721.0	12.621.277	25.767.751	0.00	0.00	0.00	0.0000	96.62
1	16674.0	181.040.645	70.490.167	96.63	123.69	162.55	2.189.475	374.01
2	3067.0	567.991.545	148.806.390	374.25	439.53	533.92	6.726.800	939.20

Πίνακας 9 Περιγραφικά στατιστικά ομαδοποίησης με βάση το CLV των τελευταίων 6 μηνών

Η συστάδα 0 έχει τον μεγαλύτερο αριθμό πελατών (72.721), αλλά με τη χαμηλότερη μέση 6μηνη νομισματική τιμή περίπου 12,62 δολάρια. Αυτό υποδηλώνει ότι, ενώ αυτή η συστάδα έχει τους περισσότερους πελάτες, είναι άτομα που ξοδεύουν λιγότερα χρήματα κατά μέσο όρο. Η τυπική απόκλιση είναι σχετικά χαμηλή και η μέγιστη δαπάνη είναι περίπου 96,62 δολαρια, υποδεικνύοντας ότι η ομάδα αυτή είναι αρκετά σταθερή σε χαμηλότερα επίπεδα δαπανών. Η συστάδα 1 περιλαμβάνει σημαντικά λιγότερους πελάτες (16.674) αλλά με πολύ υψηλότερη μέση δαπάνη. Η τυπική απόκλιση είναι υψηλότερη και η μέγιστη δαπάνη φτάνει μέχρι 374,01 δολάρια, γεγονός που υποδηλώνει μεγαλύτερη μεταβλητότητα στις δαπάνες. Οι πελάτες σε αυτή τη συστάδα είναι κατά μέσο όρο πιο πολύτιμοι σε σύγκριση με τη συστάδα 0. Η συστάδα 2 έχει τον μικρότερο αριθμό πελατών (3.067) αλλά την υψηλότερη μέση δαπάνη (περίπου 567,99 δολάρια), με πολύ υψηλή μέγιστη δαπάνη. Η τυπική απόκλιση είναι αρκετά μεγάλη, γεγονός που υποδηλώνει ότι υπάρχει μεγάλο εύρος συμπεριφορών δαπανών εντός αυτής της ομάδας. Αυτή η ομάδα αντιπροσωπεύει τους πιο πολύτιμους πελάτες από την άποψη της 6μηνης νομισματικής αξίας. Για τη

συστάδα 1 και τη συστάδα 2, οι μέσες τιμές (τεταρτημόριο 0,5) και οι τιμές του τεταρτημορίου 0,75 είναι προοδευτικά υψηλότερες, επιβεβαιώνοντας ότι αυτές οι συστάδες περιέχουν πελάτες που ξοδεύουν συχνότερα ή/και σε μεγαλύτερα ποσά.

Στο παρακάτω διάγραμμα διασποράς χρησιμοποιείται για να δείξει τη σχέση μεταξύ του Overall Score με βάση τις τιμές RFM που υπολογίσαμε προηγουμένως (Κεφάλαιο 4.2) και της αξία διάρκειας ζωής πελάτη κατά την χρονική περίοδο των έξι μηνών.



Εικόνα 4.23 Διάγραμμα διασποράς της CLV των τελευταίων 6 μηνών και του OverallScore/rfm πελατών

Οι τιμές που βρίσκονται στον άξονα x του γραφήματος (0, 1, 2, 3, 4, 5, 6, 7) αντιπροσωπεύουν τις μεμονωμένες τιμές του OverallScore (Πίνακας 8). Οι τιμές 0 έως 7 είναι διακριτές βαθμολογίες(OverallScore) που προέκυψαν από την RFM ανάλυση και που έχουν υπολογιστεί για διαφορετικούς πελάτες. Κάθε κουκκίδα στο γράφημα αντιπροσωπεύει έναν πελάτη ή μια ομάδα πελατών με την ίδια βαθμολογία RFM, ενώ ο χρωματισμός των κουκκίδων συμβολίζει την ομάδα με βάση την αξία διάρκειας ζωής πελάτη(CLV).

Αναλυτικότερα, οι πελάτες με **βαθμολογία 0** ανήκουν στο τμήμα "Χαμηλής Αξίας". Η 6μηνη Αξία Διάρκειας Ζωής τους (LTV) είναι συγκεντρωμένη στο κάτω μέρος της κλίμακας, υποδεικνύοντας ελάχιστη οικονομική συνεισφορά στην επιχείρηση κατά τη διάρκεια αυτής της περιόδου. Αυτό υποδηλώνει ότι οι πελάτες αυτοί είτε δεν έχουν πραγματοποιήσει αγορές πρόσφατα, είτε, αν έχουν πραγματοποιήσει, οι αγορές αυτές ήταν χαμηλής συχνότητας και χρηματικής αξίας. Οι στρατηγικές για την ομάδα αυτή θα μπορούσαν να περιλαμβάνουν εκστρατείες επαναπροσέγγισης ή ανάλυση για την κατανόηση της χαμηλής δέσμευσής τους.

Όπως και η βαθμολογία 0, η **βαθμολογία 1** ανήκει επίσης στο τμήμα "Χαμηλής Αξίας". Αυτοί οι πελάτες παρουσιάζουν ελαφρώς υψηλότερο LTV από εκείνους με

βαθμολογία 0, αλλά εξακολουθούν να παραμένουν στο χαμηλότερο άκρο του φάσματος LTV. Αυτή η βαθμολογία υποδηλώνει μια μικρή βελτίωση είτε στην επαναληπτικότητα, είτε στη συχνότητα, είτε στη χρηματική αξία σε σχέση με τη βαθμολογία 0, αλλά εξακολουθεί να σημαίνει περιορισμένη δέσμευση ή δαπάνη.

Οι πελάτες με **βαθμολογία 2** παραμένουν στο τμήμα "Χαμηλής Αξίας", αλλά παρουσιάζουν ευρύτερη κατανομή όσον αφορά το LTV. Ορισμένοι από αυτούς τους πελάτες έχουν σημαντικά υψηλότερο LTV σε σύγκριση με τις βαθμολογίες 0 και 1, γεγονός που υποδηλώνει ότι μπορεί να υπάρχει δυνατότητα προώθησης αυτών των πελατών προς τμήματα υψηλότερης αξίας μέσω στοχευμένων προσπαθειών μάρκετινγκ.

Η **βαθμολογία 3** αρχίζει να διαφαίνεται η μετάβαση από τα τμήματα "Χαμηλής Αξίας" στα τμήματα "Μεσαίας Αξίας". Η κατανομή του LTV κατανέμεται σε ένα υψηλότερο εύρος, υποδεικνύοντας μεγαλύτερη μεταβλητότητα στην αξία του πελάτη. Αυτό θα μπορούσε να σημαίνει ότι ορισμένοι πελάτες βρίσκονται στο κατώφλι του να γίνουν πιο πολύτιμοι και, με τα κατάλληλα κίνητρα, θα μπορούσαν ενδεχομένως να συμβάλουν πιο σημαντικά στα έσοδα.

Στην επόμενη κλίματα - **βαθμολογία 4**, οι περισσότεροι πελάτες ανήκουν στο τμήμα "Μεσαίας Αξίας". Η κατανομή του LTV είναι πιο πυκνή σε υψηλότερο εύρος από τις προηγούμενες βαθμολογίες, υποδεικνύοντας ότι αυτοί οι πελάτες είναι πιο αφοσιωμένοι και έχουν μεγαλύτερη αξία για την επιχείρηση. Αυτή η ομάδα είναι ζωτικής σημασίας, καθώς έχει τη δυνατότητα είτε να γίνει πελάτης υψηλής αξίας είτε να πέσει σε χαμηλότερο τμήμα αν δεν διατηρηθεί σωστά.

Η **βαθμολογία 5** είναι ενδιαφέρουσα, καθώς περιέχει τόσο πελάτες "μεσαίας αξίας" όσο και πελάτες "Υψηλής Αξίας", αλλά με αξιοσημείωτο χάσμα μεταξύ του LTV αυτών των δύο ομάδων. Αυτό το χάσμα θα μπορούσε να υποδεικνύει ένα όριο ή συγκεκριμένα χαρακτηριστικά που διαχωρίζουν τους πελάτες με υψηλές δαπάνες από τους πελάτες μεσαίας κατηγορίας. Η επιχείρηση θα μπορούσε να αναλύσει τις συμπεριφορές που οδηγούν τους πελάτες να γίνουν "Υψηλής Αξίας" και να αξιοποιήσει αυτές τις πληροφορίες για την αναβάθμιση των πελατών "Μεσαίας Αξίας".

Οι **βαθμολογίες 6 και 7** περιλαμβάνουν αποκλειστικά πελάτες "Υψηλής Αξίας", με το LTV να φτάνει στα ανώτερα επίπεδα της κλίμακας. Αυτοί οι πελάτες είναι οι πιο πολύτιμοι για την επιχείρηση, παρουσιάζοντας πιθανώς υψηλά επίπεδα συχνότητας, επανάληψης και χρηματικής αξίας στις συναλλαγές τους. Οι στρατηγικές

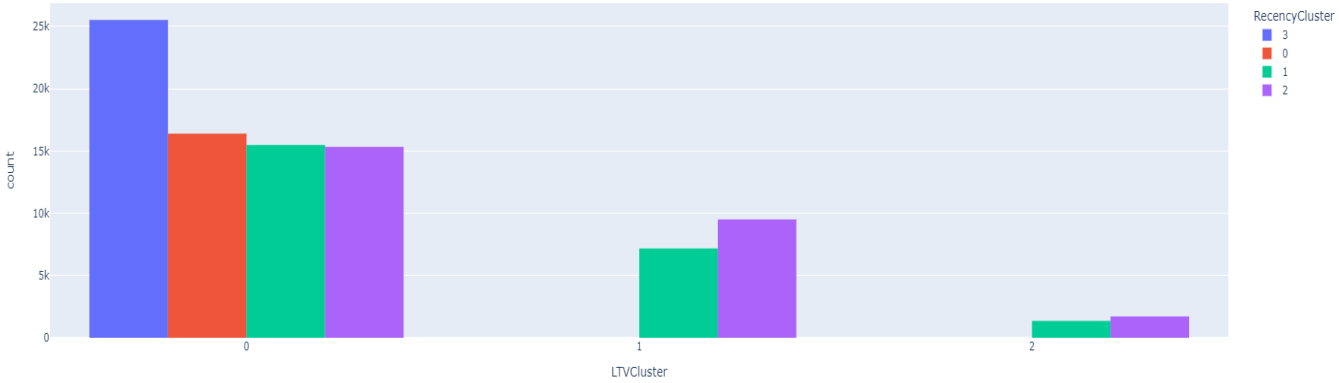
διατήρησης είναι κρίσιμες για αυτό το τμήμα, ώστε να διασφαλιστεί ότι θα συνεχίσουν την αγοραστική τους συμπεριφορά. Επιπλέον, αυτοί οι πελάτες θα μπορούσαν να στοχευθούν για «premium» προσφορές ή προγράμματα πιστότητας για να διατηρήσουν τη δέσμευσή τους.

Παράλληλα εξετάστηκε και η κατανομή των συστάδων με βάση τις ξεχωριστές μετρικές της RFM ανάλυσης, σε διαφορετικές ομάδες LTV (Lifetime Value). Αρχικά οι διαφορετικές ομάδες LTV με βάσης τις συστάδες του Recency score, σε διαφορετικές ομάδες LTV (Εικόνα 4.24):

• Το LTVCluster 0 έχει σημαντικό αριθμό πελατών άνω των 25 χιλιάδων στην πιο πρόσφατη συστάδα αλληλεπίδρασης (RecencyCluster=3), υποδεικνύοντας πρόσφατη δέσμευση από μια μεγάλη πελατειακή βάση.

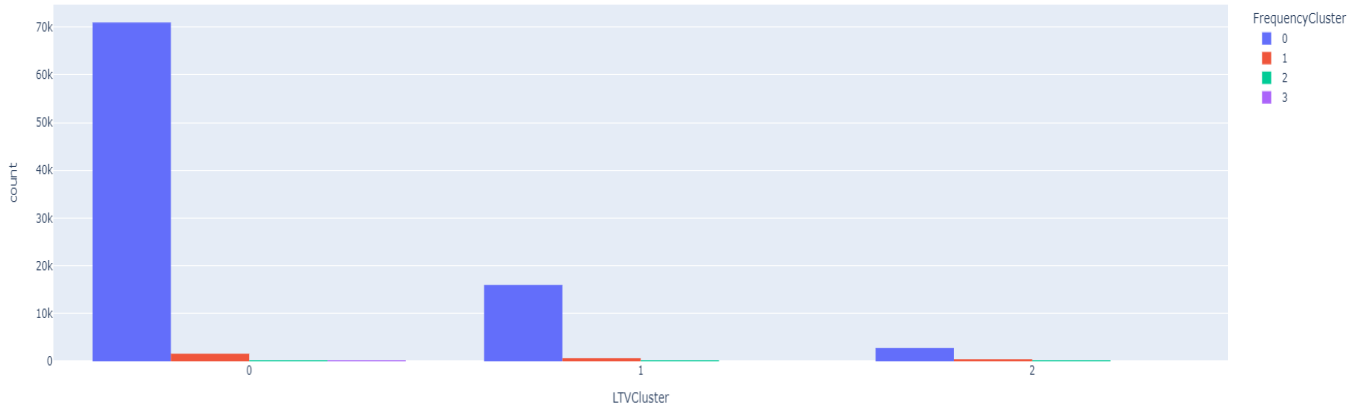
• Το LTVCluster 1 και 2 παρουσιάζουν συνολικά λιγότερους πελάτες, με το LTVCluster 1 να έχει περίπου 10 χιλιάδες πελάτες στο RecencyCluster=2, υποδεικνύοντας κάποια δέσμευση.

• Το LTVCluster 2 έχει τον μικρότερο αριθμό πελατών, μόνο μερικές εκατοντάδες, που κατανέμονται στα RecencyClusters 1 και 2, γεγονός που υποδηλώνει χαμηλή δέσμευση.



Εικόνα 4.24 Recency Cluster per LVTCluster

Συμπερασματικά, οι μετρήσεις μειώνονται με την αύξηση του αριθμού LTVCluster, γεγονός που θα μπορούσε να σημαίνει ότι οι πελάτες χαμηλότερης αξίας εμπλέκονται πιο πρόσφατα σε σύγκριση με τα τμήματα υψηλότερης αξίας. Η κατανομή υποδηλώνει ότι οι στρατηγικές πρόσφατης δέσμευσης μπορεί να είναι πιο αποτελεσματικές με πελάτες χαμηλότερου LTV και ότι μπορεί να απαιτούνται διαφορετικές προσεγγίσεις για την επαναπροσέγγιση πελατών υψηλότερης αξίας αλλά λιγότερο πρόσφατα ενεργών.

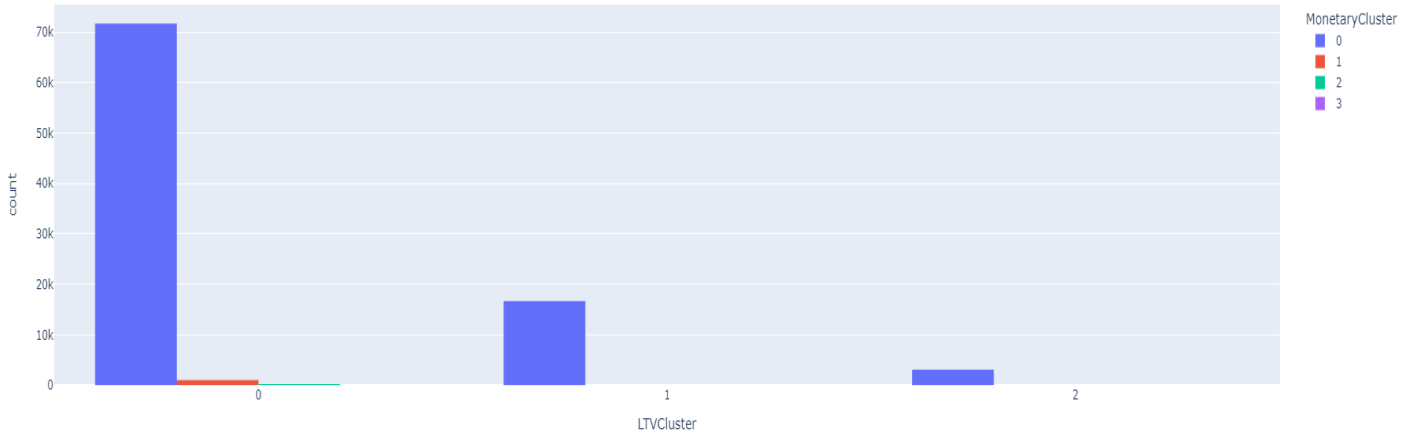


Εικόνα 4.25 Frequency Cluster per LVTCluster

Στο επόμενο διάγραμμα (Εικόνα 4.25), παρατηρούμε την κατανομή των συστάδων με βάση τις ομάδες της μετρικής Frequency.

- LTVCluster 0: Αυτή η συστάδα έχει τη πλειοψηφία των πελατών με FrequencyCluster 0, υποδεικνύοντας ένα μεγάλο τμήμα πελατών με τη χαμηλότερη συχνότητα αγορών.
- LTVCluster 1: Υπάρχει μια αξιοσημείωτη πτώση στον αριθμό των πελατών από το LTVCluster 0 στο 1. Η κυρίαρχη συστάδα συχνότητας εδώ εξακολουθεί να είναι 0, αλλά ο αριθμός είναι πολύ χαμηλότερος.
- LTVCluster 2: Αυτή η συστάδα έχει περισσότερο ομοιόμορφη κατανομή μεταξύ των συστάδων συχνότητας, αλλά όλες έχουν χαμηλό αριθμό.

Το συνολικό μοτίβο υποδηλώνει ότι οι περισσότεροι πελάτες πραγματοποιούν αγορές σπάνια (FrequencyCluster 0), ιδίως στις ομάδες με χαμηλότερο LTV. Καθώς το LTVCluster αυξάνεται, ο αριθμός των πελατών μειώνεται, αλλά υπάρχει μια πιο ομοιόμορφη κατανομή στις διάφορες συστάδες συχνότητας, αν και εξακολουθεί να είναι χαμηλός σε απόλυτους αριθμούς. Αυτές οι γνώσεις μπορεί να είναι ενδεικτικές μιας πελατειακής βάσης που δεν είναι πολύ αφοσιωμένη σε επαναλαμβανόμενες αγορές ή μιας βάσης που έχει σημαντικό ποσοστό αγοραστών που αγοράζουν μία φορά, ιδίως στα τμήματα με χαμηλότερο LTV. Η πτώση των μετρήσεων από τη συστάδα LTVC 0 στη συστάδα LTVC 2 υποδηλώνει ότι υπάρχουν λιγότεροι πελάτες με υψηλότερη αξία διάρκειας ζωής. Οι πληροφορίες αυτές θα μπορούσαν να είναι πολύτιμες για την ανάπτυξη στοχευμένων στρατηγικών μάρκετινγκ ή προγραμμάτων διατήρησης πελατών.



Εικόνα 4.26 Monetary Cluster Counts per LVTCluster

Αναφορικά με την κατανομή κατανομή των συστάδων με βάση τις ομάδες του Monetary score, σε διαφορετικές ομάδες LTV:

- Η συστάδα LTV 0 έχει μια υπεροχή πελατών στη νομισματική συστάδα 0, γεγονός που υποδηλώνει ότι ένα σημαντικό μέρος των πελατών στο χαμηλότερο κλιμάκιο LTV συνεισφέρει χαμηλή νομισματική αξία.
- Η συστάδα LTV 1 παρουσιάζει μικρότερο, αλλά σημαντικό αριθμό πελατών στη νομισματική συστάδα 0, γεγονός που υποδηλώνει ότι οι πελάτες αυτοί συνεισφέρουν μεσαία αξία.
- Η συστάδα LTV 2, η οποία μπορεί να αντιπροσωπεύει την υψηλότερη βαθμίδα LTV, έχει σημαντικά μικρότερο αριθμό στη νομισματική συστάδα 0, γεγονός που αντανακλά ότι λίγοι πελάτες υψηλής αξίας έχουν χαμηλή νομισματική συνεισφορά.

Παρατηρούμε ότι υπάρχουν ελάχιστοι έως μηδενικοί πελάτες στις υψηλότερες νομισματικές ομάδες σε όλες τις συστάδες LTV, γεγονός που θα μπορούσε να σημαίνει ότι λίγοι πελάτες πραγματοποιούν συναλλαγές υψηλής αξίας ή ότι οι συναλλαγές αυτές δεν είναι συχνές.

Αυτό ο συνδυασμός, RFM ανάλυσης και CLV είναι συχνή μορφή μελέτης, με σκοπό την εξομάλυνση των αδυναμιών της RFM ανάλυσης. Οι Fader et al., 2005 αντιμετώπισαν τον περιορισμό των μοντέλων RFM δείχνοντας πώς οι μεταβλητές RFM μπορούν να ενσωματωθούν σε ένα μοντέλο CLV. Στην μελέτη τους, απέδειξαν ότι οι μετρικές RFM είναι "επαρκή στατιστικά στοιχεία" για το μοντέλο CLV που πρότειναν. Αυτό σημαίνει ότι οι τιμές RFM περιέχουν αρκετές πληροφορίες για το μοντέλο ώστε να προβλέψει την CLV χωρίς την ανάγκη πρόσθετων δεδομένων, όπως την υπολογίσαμε και στην δική μας περίπτωση. Μια ενδιαφέρουσα έννοια που

εισήγαγε η προσέγγισή τους είναι οι καμπύλες iso-CLV. Αυτές οι καμπύλες αντιπροσωπεύουν συνδυασμούς τιμών R, F και M που αναμένεται να αποδώσουν το ίδιο CLV. Αυτό θα μπορούσε να βοηθήσει τις επιχειρήσεις να κατανοήσουν πώς διαφορετικές συμπεριφορές πελατών (όσον αφορά την επαναληπτικότητα, τη συχνότητα και τη χρηματική αξία) μπορούν να οδηγήσουν στην ίδια συνολική αξία για την εταιρεία (Gupta et al., 2006).

Ολοκληρώνοντας την μελέτη της τιμής CLV, επιχειρήθηκε η πρόβλεψη της LTVCluster, δηλαδή των συστάδων με βάση την CLV, χρησιμοποιώντας τις μεταβλητές που έχουν υπολογισθεί. Για να καταφέρουμε να βρούμε ποια από όλες τις μεταβλητές που δημιουργήθηκαν κατά το Feature engineering, RFM Cluster, Segment levels, Recency score, Recency Cluster, Frequency score, Frequency Cluster, Monetary score, Monetary Cluster, είναι πιο σημαντική ή σημαντικές ώστε να προβλέψουν επιτυχώς την LTVCluster, διερευνήθηκαν οι κορυφαίες συσχετιζόμενες μεταβλητές με αυτήν.

Columns	**corr.**
LTVCluster	1.000.000
m6_Monetary	0.893692
Cluster	0.397477
Segment_Low-Value	0.274628
Segment_Mid-Value	0.273529
Monetary	0.082281
FrequencyCluster	0.068306
Frequency	0.063721
MonetaryCluster	0.051039
RecencyCluster	0.043847
OverallScore	0.035950
CLV	0.020525
Segment_High-Value	0.018218
Recency	0.005629

Πίνακας 10 Κορυφαία συσχετιζόμενες μεταβλητές με την LTVCluster

Με βάση τα παραπάνω στοιχεία (Πίνακας 10), επιλέχθηκαν οι μεταβλητές Cluster που αντικατοπτρίζουν τις συστάδες RFM και το OverallScore - μια σύνθετη βαθμολογία που προκύπτει από τρεις βασικές μετρήσεις τμηματοποίησης πελατών: Recency Cluster, Frequency Cluster, and Monetary Cluster. Τα αποτελέσματα από τα μοντέλα μηχανικής μάθησης για την πρόβλεψη της κατηγοριοποίησης της αξίας διάρκειας ζωής των πελατών, όπως φαίνονται στο Πίνακα, δείχνουν μια ποικιλία

μοντέλων που αξιολογούνται με βάση μετρήσεις όπως η ακρίβεια, η AUC (περιοχή κάτω από την καμπύλη), η ανάκληση, η ακρίβεια (Prec.), το αποτέλεσμα F1, το Kappa, ο συντελεστής συσχέτισης MCC (Matthews Correlation Coefficient) και ο χρόνος εκπαίδευσης (TT σε δευτερόλεπτα).

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
dt	*Decision Tree Classifier*	0.8393	0.8885	0.8393	0.8047	0.8212	0.4877	0.4939	0.0400
rf	*Random Forest Classifier*	0.8393	0.8885	0.8393	0.8047	0.8212	0.4877	0.4939	0.7520
et	*Extra Trees Classifier*	0.8393	0.8885	0.8393	0.8047	0.8212	0.4877	0.4939	0.5810
lightgbm	*Light Gradient Boosting Machine*	0.8393	0.8885	0.8393	0.8047	0.8212	0.4877	0.4939	27.020
gbc	*Gradient Boosting Classifier*	0.8380	0.8884	0.8380	0.8044	0.8205	0.4869	0.4924	63.260
qda	*Quadratic Discriminant Analysis*	0.7790	0.8438	0.7790	0.7046	0.7332	0.3331	0.3643	0.0980
nb	*Naive Bayes*	0.7790	0.8420	0.7790	0.7046	0.7332	0.3331	0.3643	0.0360
knn	*K Neighbors Classifier*	0.7442	0.8263	0.7442	0.7430	0.7409	0.2750	0.2796	12.660
ada	*Ada Boost Classifier*	0.8333	0.7998	0.8333	0.7994	0.8156	0.4710	0.4764	14.560
lr	*Logistic Regression*	0.7865	0.7415	0.7865	0.6186	0.6925	0.0000	0.0000	0.2510
lda	*Linear Discriminant Analysis*	0.7865	0.7383	0.7865	0.6186	0.6925	0.0000	0.0000	0.0550
dummy	*Dummy Classifier*	0.7865	0.5000	0.7865	0.6186	0.6925	0.0000	0.0000	0.0480
svm	*SVM - Linear Kernel*	0.7865	0.0000	0.7865	0.6186	0.6925	0.0000	0.0000	0.1870
ridge	*Ridge Classifier*	0.7865	0.0000	0.7865	0.6186	0.6925	0.0000	0.0000	0.0390

Πίνακας 11 Συγκριτική ανάλυση μοντέλων μηχανικής μάθησης για την πρόβλεψη της κατηγοριοποίησης με βάση την CLV

Τα κορυφαία μοντέλα όσον αφορά την AUC, η οποία είναι μια κρίσιμη μετρική για εργασίες ταξινόμησης, καθώς μετρά την ικανότητα του μοντέλου να διακρίνει μεταξύ κλάσεων, είναι τα *Decision Tree Classifier, Random Forest Classifier, Extra Trees Classifier* και *Light Gradient Boosting Machine*, τα οποία εμφανίζουν ταυτόσημες βαθμολογίες 0,8885 AUC και 0,8393 ακρίβεια. Αυτά τα μοντέλα έχουν επίσης τις ίδιες βαθμολογίες για τα Recall, Precision και F1, υποδεικνύοντας πολύ

παρόμοιες επιδόσεις σε αυτές τις μετρήσεις. Αυτά τα μοντέλα δεν υπερέχουν μόνο στην AUC, αλλά παρουσιάζουν επίσης σταθερά υψηλές επιδόσεις σε άλλες μετρικές όπως η ακρίβεια, η ανάκληση, η ακρίβεια και το σκορ F1. Αυτό υποδεικνύει μια ισορροπημένη ικανότητα να προβλέπουν σωστά τόσο τις θετικές όσο και τις αρνητικές κλάσεις.

Όπως παρατηρούμε, υπάρχει σημαντική διαφορά στο χρόνο εκπαίδευσης (TT) μεταξύ των κορυφαίων μοντέλων. Το *Light Gradient Boosting Machine* έχει αξιοσημείωτα υψηλό χρόνο εκπαίδευσης 27.020 δευτερόλεπτα, ο οποίος είναι σημαντικά υψηλότερος από τα άλλα μοντέλα. Αυτό μπορεί να αποτελέσει μια σκέψη, εάν ο χρόνος εκπαίδευσης του μοντέλου αποτελεί περιορισμό. Άλλα μοντέλα, όπως ο ταξινομητής *Gradient Boosting*, η *Quadratic Discriminant Analysis*, το *Naive Bayes* και ο ταξινομητής *K-Neighbors*, εμφανίζουν χαμηλότερα αποτελέσματα AUC, υποδεικνύοντας μικρότερη ικανότητα διάκρισης μεταξύ κλάσεων σε σύγκριση με τα κορυφαία μοντέλα. Επιπλέον, ο ταξινομητής *Ada Boost Classifier*, η *λογιστική παλινδρόμηση*, η *γραμμική διακριτική ανάλυση*, ο *ταξινομητής Dummy Classifier*, το *SVM - Linear Kernel* και ο *ταξινομητής Ridge Classifier* παρουσιάζουν μια παρουσιάζουν μέτρια ανταγωνιστικές βαθμολογίες AUC (π.χ. *Gradient Boosting Classifier* με 0,8884 και *Ada Boost Classifier* με 0,7998), ενώ άλλα, όπως τα *Logistic Regression, Linear Discriminant Analysis, Dummy Classifier, SVM - Linear Kernel και Ridge Classifier*, παρουσιάζουν σημαντικά χαμηλότερες ή ακόμη και μηδενικές βαθμολογίες AUC (τα τρία τελευταία μοντέλα έχουν βαθμολογίες AUC 0, υποδεικνύοντας ότι δεν υπάρχει διακριτική ικανότητα μεταξύ των κλάσεων). Αυτή η διαφοροποίηση αναδεικνύει τη σημασία της επιλογής μοντέλου με βάση τα συγκεκριμένα χαρακτηριστικά και τις απαιτήσεις του συνόλου δεδομένων και του έργου πρόβλεψης.

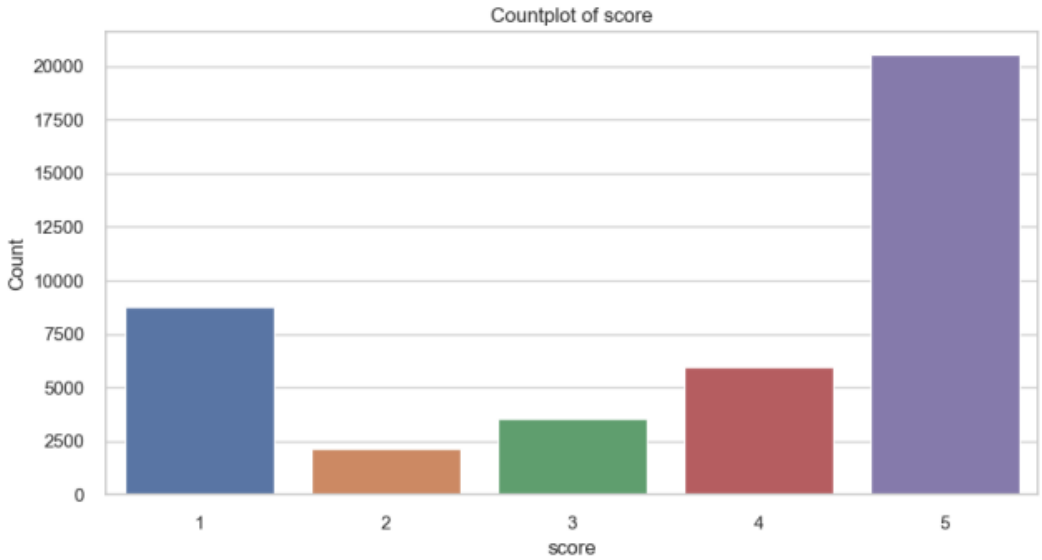
Υπάρχει μια αξιοσημείωτη αντιστάθμιση μεταξύ της αποτελεσματικότητας του μοντέλου (όπως μετράται με την AUC και άλλες μετρήσεις) και της αποδοτικότητάς του (χρόνος εκπαίδευσης). Ενώ η *Light Gradient Boosting Machine* και ο *Gradient Boosting Classifier* συγκαταλέγονται μεταξύ των κορυφαίων επιδόσεων όσον αφορά την ακρίβεια πρόβλεψης, οι χρόνοι εκπαίδευσής τους είναι σημαντικά μεγαλύτεροι, γεγονός που μπορεί να μην είναι κατάλληλο για σενάρια που απαιτούν γρήγορη επανεκπαίδευση ή ανάπτυξη του μοντέλου.

Για πρακτικές εφαρμογές, η επιλογή μεταξύ αυτών των μοντέλων θα εξισορροπούσε την προβλεπτική απόδοση με την υπολογιστική αποδοτικότητα. Οι

ταξινομητές *Decision Tree, Random Forest* και *Extra Trees* αναδεικνύονται ως ισχυροί υποψήφιοι δεδομένης της υψηλής AUC και των σχετικά χαμηλότερων χρόνων εκπαίδευσης σε σύγκριση με τη *Light Gradient Boosting Machine*. Ωστόσο, το συγκεκριμένο πλαίσιο στο οποίο θα αναπτυχθεί το μοντέλο (π.χ. η διαθεσιμότητα υπολογιστικών πόρων, η ανάγκη για προβλέψεις σε πραγματικό χρόνο κ.λπ.) θα πρέπει να καθοδηγήσει την τελική επιλογή. Συνοψίζοντας, η ανάλυση προτείνει μια διαφοροποιημένη προσέγγιση για την επιλογή ενός μοντέλου μηχανικής μάθησης για την πρόβλεψη της αξίας διάρκειας ζωής του πελάτη, λαμβάνοντας υπόψη τόσο την ικανότητα του μοντέλου να προβλέπει με ακρίβεια τα αποτελέσματα όσο και τις υπολογιστικές απαιτήσεις του.

4.4 ΑΝΑΛΥΣΗ ΣΥΝΑΙΣΘΗΜΑΤΟΣ

Σαν πρώτο βήμα για την ανάλυση του συναισθήματος και της διάθεσης των πελατών προς την επιχείρηση, είναι η διερεύνηση των βαθμολογιών που έχουν τοποθετήσει οι πελάτες για την επιχείρηση. Το διάγραμμα απεικονίζει την κατανομή των εγγραφών του συνόλου δεδομένων σχετικά με τις αξιολογήσεις των πελατών και την μεταβλητή «score», μια μεταβλητή που κυμαίνεται από το 1 έως το 5.



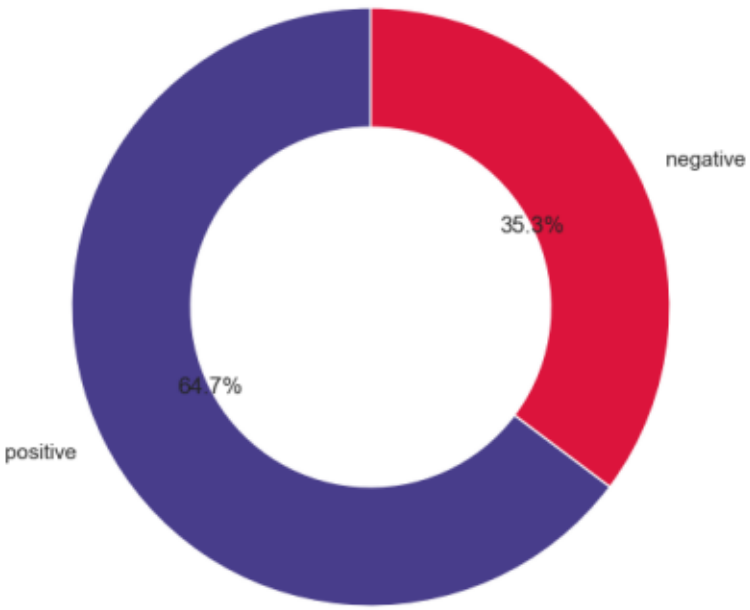
Εικόνα 4.27 Κατανομή βαθμολογιών του συνόλου δεδομένων

Από το παραπάνω ραβδόγραμμα, μπορούμε να δούμε ότι η βαθμολογία 5 έχει την υψηλότερη μέτρηση, σημαντικά μεγαλύτερη από τις μετρήσεις για τις βαθμολογίες 1 έως 4. Η βαθμολογία 1 έχει τη δεύτερη υψηλότερη μέτρηση, ενώ οι βαθμολογίες "2", 3 και 4 έχουν προοδευτικά λιγότερες μετρήσεις. Αυτό θα μπορούσε να υποδηλώνει ότι οι χρήστες είναι πιο πιθανό να δώσουν ακραίες βαθμολογίες (πολύ υψηλές ή πολύ

χαμηλές) από ό,τι μέτριες βαθμολογίες, ένα φαινόμενο γνωστό ως πόλωση ή διπολική κατανομή στις κριτικές. Μια τέτοια κατανομή θα μπορούσε να σημαίνει ότι υπάρχει έντονη διάσταση απόψεων μεταξύ των χρηστών, με μεγάλο αριθμό χρηστών να είναι είτε πολύ ικανοποιημένοι (βαθμολογία 5), είτε πολύ δυσαρεστημένοι (βαθμολογία 1) και λιγότερους χρήστες να αισθάνονται μέτρια ικανοποιημένοι ή ουδέτεροι.

Αυτού του είδους οι παρατηρήσεις είναι ιδιαίτερα χρήσιμες στην ανάλυση συναισθήματος, στην έρευνα ικανοποίησης πελατών και στην ανάλυση της αγοράς. Βοηθά στην κατανόηση της συμπεριφοράς των πελατών και μπορεί να είναι καθοριστικές στη λήψη αποφάσεων για βελτιώσεις προϊόντων ή στην κατανόηση της τοποθέτησης στην αγορά. Συνολικά, το διάγραμμα παρέχει μια σαφή οπτική αναπαράσταση της συχνότητας των διαφορετικών βαθμολογιών στο σύνολο δεδομένων, υποδεικνύοντας τον τρόπο με τον οποίο οι χρήστες αξιολογούν ό,τι αξιολογείται (προϊόντα, υπηρεσίες κ.λπ.).

Στη συνέχεια, με σκοπό την κατηγοριοποίηση των αξιολογήσεων των πελατών σε δύο μόνο κατηγορίες, θετικές ή αρνητικές, με την βοήθεια μιας συνάρτησης ταξινομείται κάθε σχόλιο ως "αρνητικό", εάν η βαθμολογία του είναι 1, 2 ή 3, και ως "θετικό", εάν η βαθμολογία του είναι μεγαλύτερη από 3 (η οποία, δεδομένης της τυπικής κλίμακας αξιολόγησης, θα ήταν συνήθως 4 ή 5). Μετά από αυτή την κατηγοριοποίηση μπορούμε πλέον να κάνουμε ένα περεταίρω βήμα προς την ανάλυση συναισθήματος.



Εικόνα 4.28 Κατανομή συναισθήματος των κριτικών

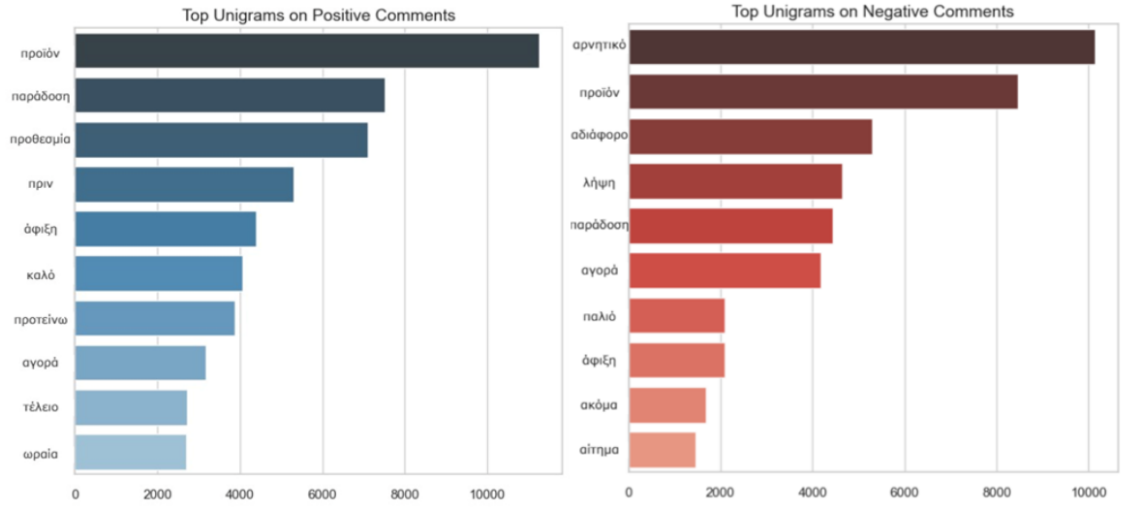
Αυτό που παρατηρείται από τη παραπάνω κατηγοριοποίηση (Εικόνα 4.28 Κατανομή συναισθήματος των κριτικών δείχνει ότι η πλειοψηφία των σχολίων είναι θετικά, γεγονός που υποδηλώνει ότι οι πελάτες είναι γενικά ικανοποιημένοι με το προϊόν, την υπηρεσία ή την εμπειρία. Για μια επιχείρηση, αυτό θα μπορούσε να σημαίνει ότι αυτό που προσφέρει είναι επιθυμητό από τους περισσότερους πελάτες της, που αποτελεί καλό σημάδι ικανοποίησης των πελατών και θα μπορούσε να συσχετιστεί με την αφοσίωση των πελατών και τη θετική διαφήμιση από στόμα σε στόμα.

Από την άλλη μια σημαντική μειοψηφία των σχολίων είναι αρνητικά. Αυτό είναι ένα σημαντικό ποσοστό και δείχνει ότι υπάρχουν τομείς στους οποίους η επιχείρηση θα μπορούσε να βελτιωθεί. Για την επιχείρηση, αυτό υπογραμμίζει τη σημασία της αντιμετώπισης αυτών των αρνητικών εμπειριών για τη βελτίωση της συνολικής ικανοποίησης των πελατών, ενδεχομένως με τη διερεύνηση των αιτιών της δυσαρέσκειας και την εφαρμογή αλλαγών για την αντιμετώπισή τους. Από επιχειρηματική σκοπιά, τα συμπεράσματα από αυτή την οπτικοποίηση θα είναι:

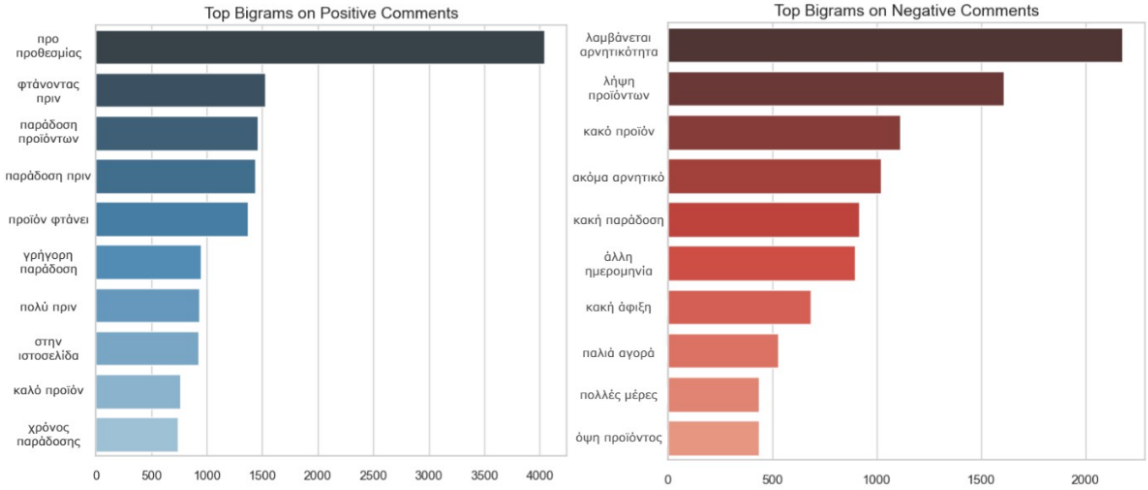
- **Δυνατά σημεία**: Το μεγαλύτερο θετικό τμήμα αναδεικνύει τα δυνατά σημεία των επιχειρηματικών προσφορών που έχουν μεγάλη απήχηση στην πελατειακή βάση.
- **Ευκαιρίες βελτίωσης**: Το αρνητικό τμήμα υποδεικνύει τομείς πιθανής βελτίωσης. Η επιχείρηση θα πρέπει να διερευνήσει τους λόγους πίσω από τα αρνητικά συναισθήματα και να λάβει διορθωτικά μέτρα.
- **Θέση στην αγορά**: Η αναλογία των θετικών προς τα αρνητικά σχόλια μπορεί επίσης να παρέχει έναν πρόχειρο δείκτη της θέσης της επιχείρησης στην αγορά σε σύγκριση με τους ανταγωνιστές.

Συνολικά, ενώ το κλίμα είναι περισσότερο θετικό παρά αρνητικό, η παρουσία ενός αξιοσημείωτου αριθμού αρνητικών σχολίων υποδηλώνει ότι υπάρχει περιθώριο για την επιχείρηση να αναπτυχθεί και να βελτιωθεί. Η αντιμετώπιση αυτών που οδηγούν σε αρνητικά συναισθήματα θα μπορούσε να βοηθήσει στη μετατροπή των δυσαρεστημένων πελατών σε ικανοποιημένους, ενισχύοντας τη φήμη της επιχείρησης και ενδεχομένως τις οικονομικές της επιδόσεις. Έπειτα είναι ιδιαίτερα ενδιαφέρον να εξετάσουμε τα σχόλια σε συνδυασμό με την κατηγορία συναισθήματος όπως αυτή ορίστηκε παραπάνω. Τα παρακάτω διαγράμματα εμφανίζουν τα πιο συνηθισμένα unigrams (μεμονωμένες λέξεις), bigrams (συνδυασμοί δύο λέξεων) και trigrams (συνδυασμοί τριών λέξεων). Ο οριζόντιος προσανατολισμός των ράβδων διευκολύνει

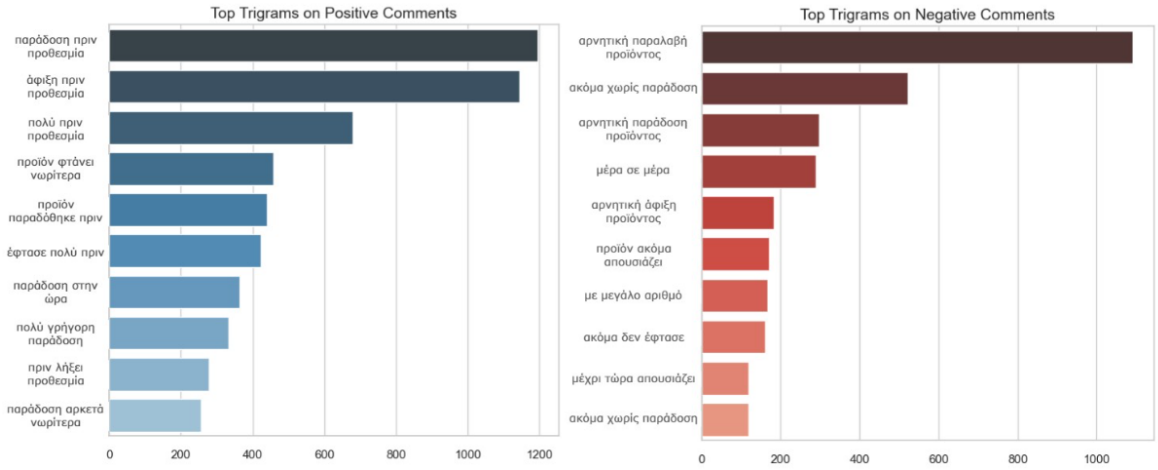
την ανάγνωση και τα χρώματα (μπλε για τα θετικά, κόκκινο για τα αρνητικά) βοηθούν στη γρήγορη διάκριση μεταξύ των συναισθημάτων. Η καταμέτρηση στον άξονα x ποσοτικοποιεί πόσο συχνά εμφανίζεται κάθε n-gram.



Εικόνα 4.29 Unigrams Θετικών & Αρνητικών σχολίων



Εικόνα 4.30 Bigrams Θετικών & Αρνητικών σχολίων



Εικόνα 4.31 Trigrams Θετικών & Αρνητικών σχολίων

Η παρουσία αυτών των n-grams σε θετικά ή αρνητικά σχόλια μπορεί να δώσει σε μια επιχείρηση πληροφορίες σχετικά με το τι αναφέρουν συχνά οι πελάτες όταν είναι ευχαριστημένοι ή δυσαρεστημένοι. Αυτό μπορεί να βοηθήσει στον εντοπισμό των δυνατών σημείων που πρέπει να αξιοποιηθούν και των ζητημάτων που πρέπει να αντιμετωπιστούν. Για παράδειγμα, εάν ένα συγκεκριμένο χαρακτηριστικό του προϊόντος αναφέρεται συχνά σε αρνητικά σχόλια, μπορεί να χρειάζεται βελτίωση. Αντίθετα, αν η άριστη εξυπηρέτηση πελατών είναι ένα επαναλαμβανόμενο θέμα στα θετικά σχόλια, μπορεί να είναι ένα δυνατό σημείο που πρέπει να τονιστεί στις προσπάθειες μάρκετινγκ.

Ολοκληρώνοντας την ανάλυση συναισθήματος έγινε η προσπάθεια με την βοήθεια μηχανικής μάθησης, η δυαδική ταξινόμηση των σχολίων αυτών κάθε αυτών με την χρήση των ταξινομητών Logistic Regression και Naive Bayes. Ο παρακάτω πίνακας περιλαμβάνει τις μετρήσεις επιδόσεων και για τα δύο μοντέλα, οι οποίες μετρήθηκαν χρησιμοποιώντας την διασταυρούμενη επικύρωση(cross-validation) κατά τη διάρκεια των φάσεων εκπαίδευση(training phase) και δοκιμής(testing phase).

Model	approach	acc	precision	recall	f1	auc	total time
LogisticRegression	Training set	0.8863	0.9224	0.9225	0.9225	0.9436	3.656
LogisticRegression	Test set	0.8856	0.9241	0.9204	0.9223	0.9446	0.025
Naive Bayes	Training set	0.8357	0.9459	0.8231	0.8802	0.8846	6.234
Naive Bayes	Test set	0.8333	0.9474	0.8195	0.8788	0.8874	0.110

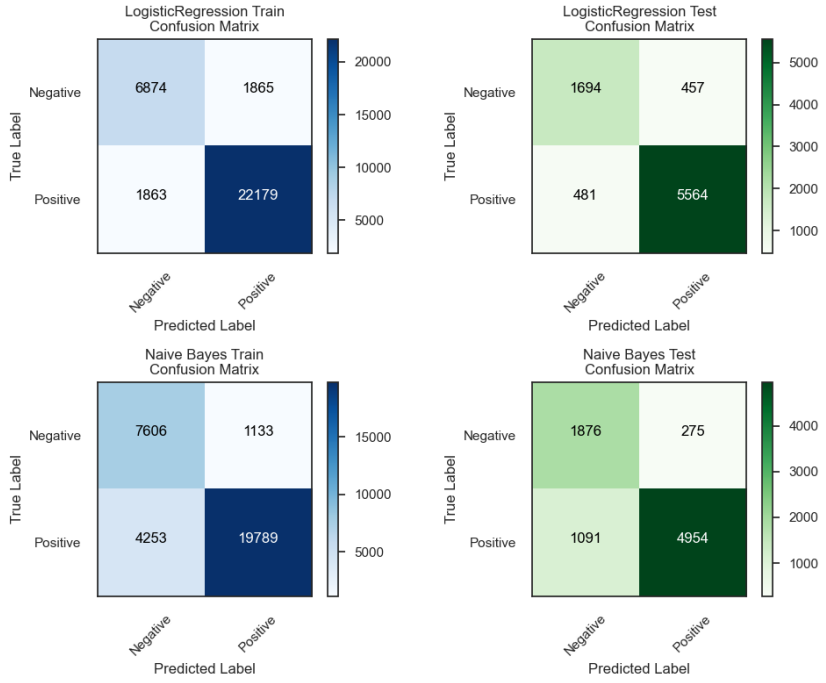
Πίνακας 12 Μετρικές αξιολόγησης των μοντέλων ταξινόμησης

Απόδοση μοντέλου Logistic Regression: Κατά τη διάρκεια εκπαίδευσης μοντέλου(training) με 5-fold cross-validation, η λογιστική παλινδρόμηση πέτυχε ακρίβεια 0,8863, ακρίβεια 0,9224, ανάκληση 0,9225, βαθμολογία F1 0,9436 και AUC 0,3656. Στο σύνολο δεδομένων δοκιμής(testing), η λογιστική παλινδρόμηση πέτυχε ακρίβεια 0,8856, ακρίβεια 0,9241, ανάκληση 0,9204, βαθμολογία F1 0,9446 και πολύ μικρό συνολικό χρόνο 0,025 δευτερόλεπτα.

Τα αποτελέσματα παρέχουν μια συγκριτική άποψη των μετρικών απόδοσης για τους ταξινομητές Logistic Regression και Naive Bayes. Κατά την εξέταση της συνολικής ακρίβειας, η λογιστική παλινδρόμηση παρουσιάζει ένα μικρό πλεονέκτημα έναντι του Naive Bayes. Η ακρίβεια(AUC) κατά την εκπαίδευση για τη Logistic Regression είναι 0,8863, η οποία είναι οριακά υψηλότερη από την ακρίβεια του Naive Bayes που είναι 0,8357. Αυτή η τάση συνεχίζεται στα αποτελέσματα της δοκιμής, όπου η Logistic Regression υπερτερεί και πάλι έναντι της Naive Bayes με ακρίβεια 0,8856 έναντι 0,8333.

Το precision του Naive Bayes είναι αξιοσημείωτο, με υψηλές τιμές τόσο στη φάση της εκπαίδευσης όσο και στη φάση της δοκιμής (0,9459 και 0,9474, αντίστοιχα), υποδεικνύοντας την αποτελεσματικότητά του. Ωστόσο, η λογιστική παλινδρόμηση δεν βρίσκεται πολύ πίσω, παρουσιάζοντας επίσης ανταγωνιστικές τιμές (0,9224 στην εκπαίδευση και 0,9241 στη δοκιμή). Ενώ η βαθμολογία F1, η οποία εξισορροπεί το συμβιβασμό μεταξύ ακρίβειας και ανάκλησης, είναι ελαφρώς υψηλότερη για τη λογιστική παλινδρόμηση, υποδεικνύοντας μια πιο ισορροπημένη απόδοση. Τελικώς, όσον αφορά την υπολογιστική αποδοτικότητα, και τα δύο μοντέλα παρουσιάζουν γρήγορους χρόνους εκτέλεσης στη φάση δοκιμής, αλλά η λογιστική παλινδρόμηση είναι ιδιαίτερα γρήγορη με συνολικό χρόνο μόλις 0,025 δευτερόλεπτα.

Λαμβάνοντας υπόψη αυτές τις μετρήσεις απόδοσης και εξετάζοντας τις πρακτικές εφαρμογές των μοντέλων, η λογιστική παλινδρόμηση φαίνεται να είναι το προτιμότερο μοντέλο για ανάπτυξη. Η καλύτερη ισορροπία της ακρίβειας, της ακρίβειας, της ανάκλησης και της ταχύτητας το καθιστά ισχυρό υποψήφιο για σενάρια όπου απαιτούνται γρήγορες και αξιόπιστες προβλέψεις. Η συνέπεια της απόδοσής του από την εκπαίδευση στη δοκιμή υποδηλώνει ότι το μοντέλο είναι πιθανό να γενικεύει καλά σε αθέατα δεδομένα, γεγονός που αποτελεί ζωτικό χαρακτηριστικό για ένα στιβαρό μοντέλο μηχανικής μάθησης.



Εικόνα 4.32 Confusion Matrices μοντέλων ταξινόμησης

Η παραπάνω εικόνα (Εικόνα 4.32) δείχνει πίνακες «σύγχυσης» (confusion matrices) για τους ταξινομητές Logistic Regression και Naive Bayes, με ξεχωριστούς

πίνακες για τα δεδομένα εκπαίδευσης και δοκιμής ενισχύοντας την αξιολόγηση της απόδοσης των μοντέλων ταξινόμησης. Σκοπός είναι να παρατηρηθεί ο αριθμός των σωστών και λανθασμένων προβλέψεων σε σύγκριση με την πραγματική τους διάσταση, κατανοώντας όχι μόνο της συνολικής ακρίβειας αλλά και των τύπων σφαλμάτων που κάνει το μοντέλο.

Για το μοντέλο **λογιστικής παλινδρόμησης**: 1) Στη φάση εκπαίδευσης, προέβλεψε σωστά 22.179 θετικές περιπτώσεις και 6.874 αρνητικές περιπτώσεις. Ωστόσο, ταξινόμησε εσφαλμένα 1.865 περιπτώσεις ως αρνητικές που στην πραγματικότητα ήταν θετικές και 1.863 περιπτώσεις ως θετικές που στην πραγματικότητα ήταν αρνητικές. 2) Στη φάση δοκιμής, το μοντέλο λογιστικής παλινδρόμησης προέβλεψε σωστά 5.564 θετικές περιπτώσεις και 1.694 αρνητικές περιπτώσεις. Ο αριθμός των ψευδώς θετικών (περιπτώσεις που ταξινομήθηκαν εσφαλμένα ως θετικές) ήταν 457 και ο αριθμός των ψευδώς αρνητικών (περιπτώσεις που ταξινομήθηκαν εσφαλμένα ως αρνητικές) ήταν 481.

Για το μοντέλο **Naive Bayes**: 1) Κατά τη διάρκεια της εκπαίδευσης, προέβλεψε σωστά 19.789 θετικές περιπτώσεις και 7.606 αρνητικές περιπτώσεις. Ο αριθμός των ψευδώς θετικών ήταν 1.133 και ο αριθμός των ψευδώς αρνητικών ήταν πολύ υψηλότερος, 4.253. 2) Στα δεδομένα δοκιμής, το μοντέλο Naive Bayes ταξινόμησε σωστά 4.954 θετικές περιπτώσεις και 1.876 αρνητικές περιπτώσεις. Ταξινόμησε λανθασμένα 275 περιπτώσεις ως θετικές που στην πραγματικότητα ήταν αρνητικές, ενώ ο αριθμός των ψευδώς αρνητικών ήταν σημαντικά υψηλότερος σε 1.091.

Τα συμπεράσματα από αυτούς τους πίνακες σύγχυσης είναι: Λογιστική παλινδρόμηση: Αυτό το μοντέλο παρουσιάζει μια πιο ισορροπημένη απόδοση όσον αφορά τα ψευδώς θετικά και τα ψευδώς αρνητικά αποτελέσματα. Φαίνεται να έχει ένα πιο συντηρητικό πρότυπο πρόβλεψης, με λιγότερες περιπτώσεις που ταξινομούνται ως θετικές σε σύγκριση με το Naive Bayes. Naive Bayes: Αυτό το μοντέλο έχει την τάση να προβλέπει περισσότερα ψευδώς αρνητικά, ιδίως στη φάση δοκιμής. Ωστόσο, έχει λιγότερα ψευδώς θετικά αποτελέσματα, γεγονός που υποδηλώνει ότι όταν προβλέπει μια περίπτωση ως αρνητική, είναι αρκετά αξιόπιστο.

Και για τα δύο μοντέλα, ο αριθμός των ψευδώς αρνητικών(false negatives) είναι υψηλότερος στα δεδομένα εκπαίδευσης σε σύγκριση με τα δεδομένα δοκιμής, γεγονός που θα μπορούσε να υποδηλώνει ότι τα μοντέλα είναι πιο προσεκτικά με την ανάθεση της θετικής κλάσης. Αυτό μπορεί να είναι ένα επιθυμητό χαρακτηριστικό σε ορισμένες

εφαρμογές όπου τα ψευδώς θετικά(false positives) αποτελέσματα είναι πιο δαπανηρά από τα ψευδώς αρνητικά.

Συγκρίνοντας τα αποτελέσματα της ανάλυσης μας με άλλες μελέτες από την βιβλιογραφία, παρατηρείται ότι η απόδοση των μοντέλων μας είναι ιδιαίτερα υψηλή. Σε άλλη μελέτη σύγκρισης αποδοτικότητας διάφορων μοντέλων μηχανικής μάθησης με σκοπό την ανάλυση συναισθήματος (Abate & Rashid, 2024), όπου χρησιμοποιήθηκαν τα μοντέλα της Λογιστικής Παλινδρόμησης και Naive Bayes, η απόδοσή τους ήταν 61% και 86,53% αντίσοιχα, ενώ στην δική μας μελέτη η απόδοσή του είναι 88,63% και 83,53% αντίστοιχα.

Στο πλαίσιο πρακτικής εφαρμογής, η επιλογή μεταξύ αυτών των μοντέλων θα εξαρτιόταν από το συγκεκριμένο κόστος που σχετίζεται με τα ψευδώς θετικά και τα ψευδώς αρνητικά αποτελέσματα. Για παράδειγμα, εάν η αρνητική κατηγορία αντιπροσωπεύει μια κρίσιμη αποτυχία ή μια κατάσταση υψηλού κινδύνου, θα μπορούσε να προτιμηθεί ένα χαμηλότερο ποσοστό ψευδώς αρνητικών, ακόμη και με το κόστος περισσότερων ψευδώς θετικών. Από την άλλη πλευρά, εάν ένα ψευδώς θετικό αποτέλεσμα είναι πιο δαπανηρό ή πιο επικίνδυνο, μπορεί να προτιμηθεί ένα μοντέλο με λιγότερα ψευδώς θετικά αποτελέσματα.

Και τα δύο μοντέλα πρέπει να συντονίζονται με βάση τους συγκεκριμένους συμβιβασμούς και τα κόστη που σχετίζονται με την επιχείρηση. Επιπλέον, θα πρέπει να λαμβάνονται υπόψη πρόσθετες μετρήσεις επιδόσεων όπως η ακρίβεια, η ανάκληση και η βαθμολογία F1 μαζί με τον πίνακα σύγχυσης, ώστε να υπάρχει μια ολοκληρωμένη κατανόηση της απόδοσης του μοντέλου. Συνδυάζοντας τις πληροφορίες τόσο από τον πίνακα μετρικών επιδόσεων όσο και από τους πίνακες σύγχυσης, μπορούμε να σχηματίσουμε μια ολοκληρωμένη άποψη για το πώς αποδίδουν οι ταξινομητές Logistic Regression και Naive Bayes στο πλαίσιο της ανάλυσης του συναισθήματος των κριτικών πελατών.

Το μοντέλο λογιστικής παλινδρόμησης παρουσιάζει υψηλές βαθμολογίες ακρίβειας, ανάκλησης και F1, οι οποίες είναι ενδεικτικές ενός ισορροπημένου και ισχυρού μοντέλου. Η βαθμολογία ακρίβειας από τον προηγούμενο πίνακα υποδηλώνει ότι όταν το μοντέλο προβλέπει μια κριτική ως θετική, είναι σωστό περίπου στο 92% των περιπτώσεων. Η βαθμολογία ανάκλησης δείχνει ότι αναγνωρίζει σωστά το 92% όλων των θετικών κριτικών. Η βαθμολογία F1, η οποία είναι ο αρμονικός μέσος όρος της ακρίβειας και της ανάκλησης, επιβεβαιώνει τη δύναμη του μοντέλου στην εξισορρόπηση αυτών των μετρήσεων. Οι πίνακες σύγχυσης το ενισχύουν αυτό,

δείχνοντας μια σχετικά ομοιόμορφη κατανομή ψευδώς θετικών και ψευδώς αρνητικών αποτελεσμάτων. Σε επιχειρηματικό πλαίσιο, αυτό υποδηλώνει ότι το μοντέλο λογιστικής παλινδρόμησης παρέχει μια αξιόπιστη εκτίμηση του κλίματος των πελατών, με μια ισορροπημένη προσέγγιση και στους δύο τύπους σφαλμάτων.

Ο ταξινομητής Naive Bayes, ενώ έχει υψηλή ακρίβεια, έχει αισθητά χαμηλότερη ανάκληση σε σύγκριση με τη λογιστική παλινδρόμηση. Αυτό σημαίνει ότι ενώ οι προβλέψεις που κάνει για τις θετικές κριτικές είναι πολύ πιθανό να είναι σωστές, χάνει σημαντικό αριθμό θετικών κριτικών, χαρακτηρίζοντάς τες ως αρνητικές (ψευδώς αρνητικές). Οι πίνακες σύγχυσης για το Naive Bayes δείχνουν αυτή τη διαφορά, ιδιαίτερα με υψηλότερο αριθμό ψευδώς αρνητικών. Για τις επιχειρήσεις, αυτό θα μπορούσε να σημαίνει ότι το μοντέλο μπορεί να υποεκφράζει το θετικό συναίσθημα στις κριτικές των πελατών, οδηγώντας ενδεχομένως σε υποτίμηση των θετικών εμπειριών των πελατών.

Για τις στρατηγικές μάρκετινγκ, η λογιστική παλινδρόμηση θα μπορούσε να είναι πιο πολύτιμη, καθώς παρέχει μια ισορροπημένη άποψη, διασφαλίζοντας ότι τα θετικά συναισθήματα καταγράφονται με ακρίβεια, ενώ παράλληλα δεν υπερτονίζεται η θετικότητα. Αυτή η ισορροπία είναι ζωτικής σημασίας για τη δημιουργία εκστρατειών μάρκετινγκ που ανταποκρίνονται στο πραγματικό συναίσθημα της πελατειακής βάσης. Όσον αφορά τη διαχείριση των πελατειακών σχέσεων, ο μικρότερος αριθμός ψευδώς αρνητικών αποτελεσμάτων στη λογιστική παλινδρόμηση θα διασφαλίσει ότι τα θετικά σχόλια των πελατών δεν παραβλέπονται, πράγμα που είναι απαραίτητο για την αναγνώριση και την επιβράβευση της πιστότητας της μάρκας. Σημαίνει επίσης ότι οι αρνητικές κριτικές εντοπίζονται με ακρίβεια, επιτρέποντας στην επιχείρηση να αντιμετωπίσει άμεσα τα ζητήματα αυτά. Το Naive Bayes, με το υψηλότερο ποσοστό ψευδώς αρνητικών κριτικών, μπορεί να είναι πιο χρήσιμο σε περιπτώσεις όπου το κόστος της απώλειας μιας θετικής κριτικής είναι χαμηλότερο από το κόστος του λανθασμένου προσδιορισμού μιας αρνητικής κριτικής ως θετικής. Ωστόσο, αυτό το χαρακτηριστικό μπορεί να μην είναι τόσο επιθυμητό για την ανάλυση του συναισθήματος των πελατών, όπου η ακριβής κατανόηση τόσο των θετικών όσο και των αρνητικών συναισθημάτων είναι σημαντική για την ολοκληρωμένη ανάλυση της ικανοποίησης και της αφοσίωσης των πελατών.

Συνολικά, δεδομένων των μετρήσεων απόδοσης και των πινάκων σύγχυσης, η λογιστική παλινδρόμηση φαίνεται να είναι το καταλληλότερο μοντέλο για την ανάλυση των συναισθημάτων των κριτικών των πελατών για τη λήψη επιχειρηματικών

αποφάσεων. Προσφέρει μια αξιόπιστη και ισορροπημένη προοπτική για το συναίσθημα των πελατών, η οποία είναι ανεκτίμητη για τη διαμόρφωση αποτελεσματικών στρατηγικών μάρκετινγκ και τη λήψη τεκμηριωμένων αποφάσεων διαχείρισης πελατειακών σχέσεων. Ωστόσο, η επιλογή του μοντέλου θα πρέπει πάντα να καθοδηγείται από το συγκεκριμένο επιχειρηματικό πλαίσιο και το σχετικό κόστος των ψευδώς θετικών και αρνητικών αποτελεσμάτων σε αυτό το πλαίσιο.

Στην μελέτη των Abate & Rashid, οι μελετητές κατέληξαν σε μοντέλα βαθιάς μηχανικής μάθησης(deep learning models) λόγω υψηλότερης αποδοτικότητας από τα κλασσικά μοντέλα ML, όπως αυτά που μελετήσαμε παραπάνω. Σε αυτή τη μελέτη, τα μοντέλα ML είχαν συνολικά ποσοστό απόδοσης κατά μέσο όρο 75,89%, ενώ τα μοντέλα deep learning είχαν 83,80%. Για την ενίσχυση των μοντέλων μηχανικής μάθησης όπως το συνιστώνται τα Naive Bayes bigrams που εφαρμόσαμε και παραπάνω (Abate & Rashid, 2024).

Κεφάλαιο 5: Συμπεράσματα

Η έρευνα επικεντρώθηκε κυρίως στη χρήση της ανάλυσης RFM στο πλαίσιο της διαχείρισης των πελατειακών σχέσεων (CRM) και της εκτίμησης της πελατειακής πίστης. Η ανάλυση RFM, η οποία σημαίνει Recency (χρονική εγγύτητα), Frequency (συχνότητα) και Monetary value (χρηματική αξία), εφαρμόστηκε για την κατηγοριοποίηση των πελατών με βάση την αγοραστική τους συμπεριφορά. Αυτή η τμηματοποίηση επέτρεψε τον εντοπισμό πολύτιμων ομάδων πελατών, διευκολύνοντας τις στοχευμένες στρατηγικές μάρκετινγκ. Με μια πρώτη ματιά στις ξεχωριστές αυτές μετρικές διαπιστώνεται ότι οι πελάτες του marketplace έχουν αρκετό χρονικό εύρος στις αγορές τους, που είναι λογικό για τα δεδομένα καταστημάτων καθώς οι πελάτες συνήθως αποκτώνται κατά τη διάρκεια μιας εκτεταμένης χρονικής περιόδου, βέβαια η ομάδα πελατών με τις πιο πρόσφατες αγορές έχει κατά μέσο όρο αγοράσει πριν περίπου 2 μήνες προϊόν από το κατάστημα. Η παρατήρηση αυτή υποδηλώνει πως τους τελευταίους 2 μήνες που καλύπτει το δείγμα μας, δηλαδή κατά τον Αύγουστο και Σεπτέμβριο του 2018, δεν τοποθετήθηκαν ιδιαίτερα πολλές παραγγελίες στο ηλεκτρονικό κατάστημα, δεδομένης της περιόδου αυτό μπορεί να οφείλεται στην χαμηλή τάση εμπορικής κατανάλωσης κατά το τέλος της θερινής περιόδου, είτε στην έλλειψη προωθητικών ενεργειών και καμπανιών από το τμήμα μάρκετινγκ του καταστήματος. Επομένως, η επιχείρηση οφείλει να προσεγγίσει το πελατολόγιό της με εκστρατείες επανενεργοποίησης και να επενδύσει σε καμπάνιες μάρκετινγκ και για την ενδεχόμενη απόκτηση νέων πελατών.

Ταυτόχρονα, στο πλαίσιο της συχνότητας αγορών του πελατολογίου του καταστήματος, όπως παρατηρήθηκε οι πελάτες δεν τείνουν να επαναλαμβάνουν αγορές πολύ συχνά. Οι περισσότεροι τείνουν να πραγματοποιούν μια αγορά και να σταματούν την αλληλεπίδρασή τους με το κατάστημα. Δυστυχώς είναι ελάχιστοι οι πελάτες που στο διάστημα 2 ετών έχουν τοποθετήσει πάνω από 7 παραγγελίες στο κατάστημα. Πρέπει να λαμβάνουμε υπόψιν μας τις κατηγορίες των προϊόντων που η επιχείρηση εμπορεύεται, κατά βάση το κατάστημα εμπορεύεται, ηλεκτρονικές συσκευές, οικιακά προϊόντα, λευκά είδη, παιχνίδια, βιβλία κ.α. Σίγουρα η πλειονότητα των προϊόντων όπως είναι τα ηλεκτρονικά είδη, είναι προϊόντα που δεν τείνει να αγοράζει ένας

καταναλωτής επανειλημμένα κατά τη διάρκεια μικρών χρονικών περιόδων όπως αυτή που εξετάζουμε. Αυτή η παρατήρηση, σίγουρα, ρίχνει φως στην αιτία της τόσο έντονης συσσωρευτικής κατανομής των δεδομένων Frequency. Πάραυτα, η επιχείρηση με βάση τα παραπάνω θα ήταν καλό να στοχεύσει σε πιο καθημερινές διαφημίσεις προϊόντων που είναι συμπληρωματικά αγαθά. Τα συμπληρωματικά αγαθά ορίζονται ως δύο ή και περισσότερα αγαθά που καταναλώνονται μαζί με σκοπό την ικανοποίηση της ίδιας ανάγκης. Για παράδειγμα, η ο φακός και οι μπαταρίες είναι δύο συμπληρωματικά αγαθά. Εάν το κατάστημα πουλά προϊόντα που χρειάζονται μπαταρίες για την λειτουργία τους, η εισαγωγή και διαφήμιση διάφορων ειδών μπαταριών στους χρήστες που αγοράζουν τέτοια αγαθά είναι μια κίνηση παρότρυνσης αγοράς ενός συμπληρωματικού αγαθού, το οποίο μάλιστα καταναλώνεται συχνά και φθείρεται γρηγορότερα από το πρώτο προϊόν, την συσκευή.

Κατά την ανάλυση της χρηματικής αξίας που δαπανούν οι πελάτες στο ηλεκτρονικό κατάστημα αρχικά η παρατήρηση των αποτελεσμάτων μοιάζει να έρχεται σε αντιδιαστολή με τα παραπάνω, γιατί οι περισσότεροι πελάτες κάνουν χαμηλής αξίας αγορές, δεν υπάρχει δηλαδή μεγάλη απόκλιση και μεγάλο εύρος στην κατανομή των χρηματικών αξιών, όπως αντίστοιχα συνέβαινε και στην τιμή Frequency. Ωστόσο, αν παρατηρήσουμε την μέση αξία αγοράς που δαπάνησαν οι πελάτες σε αυτή την κατηγορία, αυτή είναι υψηλότερη των 170 δολαρίων. Αυτό σημαίνει ότι οι πελάτες με τις χαμηλότερες σε τιμή αγορές, περίπου 90.000 στο σύνολο, η αγορά τους ήταν πάνω από 170 δολάρια. Ενώ, όλοι οι υπόλοιποι πελάτες, περίπου 1.550 στο σύνολο ξόδεψαν χιλιάδες δολάρια στις αγορές τους κατά μέσο όρο. Αυτή η παρατήρηση, έρχεται να ενισχύσει με την προηγούμενη υπόθεση. Με σκοπό να αυξήσουμε την συχνότητα των αγορών πρέπει να αυξηθεί το εύρος τιμών των προϊόντων που διαθέτει το κατάστημα για να δοθούν κίνητρα στους πελάτες να τοποθετήσουν περισσότερες από μια παραγγελίες σε τακτά χρονικά διαστήματα.

Τα ευρήματα της μελέτης επεκτάθηκαν με τον συνδυασμό των παραπάνω μετρικών για την εφαρμογή της ανάλυσης RFM με σκοπό να εντοπιστούν μοτίβα συμπεριφοράς για τη βελτίωση της δέσμευσης των πελατών, της διατήρησης και της συνολικής κερδοφορίας.

Τα αποτελέσματα της ανάλυσης RFM παρουσιάζουν μια ολοκληρωμένη εικόνα της συμπεριφοράς των πελατών και της ικανότητας της επιχείρησης να τμηματοποιεί αποτελεσματικά την αγορά της. Με την εξέταση των μετρικών Recency, Frequency και Monetary συνδυαστικά, η ανάλυση χωρίζει τους πελάτες σε διακριτές ομάδες, κάθε μία

από τις οποίες χαρακτηρίζεται από το επίπεδο δέσμευσής τους και τη συμβολή τους στα έσοδα της επιχείρησης. Η μελέτη αποκαλύπτει ότι οι πελάτες είναι βέλτιστο να τμηματοποιούνται σε τέσσερις κύριες ομάδες και για αυτό ακολουθείται αυτή η τμηματοποίηση με βάση τις βαθμολογίες RFM τους. Η μεγαλύτερη ομάδα αποτελείται από πελάτες με μέτρια χρονική εγγύτητα αγορών, χαμηλότερη συχνότητα και χαμηλότερη μέση χρηματική συνεισφορά. Αυτό υποδηλώνει ότι ένα σημαντικό τμήμα της πελατειακής βάσης πραγματοποιεί σπάνια αγορές και δαπανά λιγότερα, αναδεικνύοντας μια ευκαιρία για την επιχείρηση να εμπλέξει αυτούς τους πελάτες πιο αποτελεσματικά για να αυξήσει τη συχνότητα των αγορών τους και τις δαπάνες τους. Όπως αναφέραμε παραπάνω, η επιχείρηση χρειάζεται ένα πελατολόγιο που θα πραγματοποιεί τακτικά αγορές μικρής αξίας.

Από την άλλη μεριά, μια μικρότερη ομάδα πελατών εμφανίζει υψηλές βαθμολογίες συχνότητας, επανάληψης και χρηματικής αξίας, υποδεικνύοντας ένα τμήμα πελατών με μεγάλη αξία, οι οποίοι εμπλέκονται συχνά με την επιχείρηση και συμβάλλουν σημαντικά στα έσοδά της. Αυτοί οι πελάτες είναι πιθανότατα πιστοί και θα μπορούσαν να στοχευθούν με premium υπηρεσίες, προγράμματα πιστότητας ή αποκλειστικές προσφορές για να διατηρήσουν ή να ενισχύσουν τα επίπεδα δέσμευσής τους. Άλλωστε, όπως είδαμε και κατά την διερευνητική ανάλυση των δεδομένων, υπάρχει η παρατήρηση μιας συσωρευτικής κατανομή των πωλήσεων με βάση τον αριθμό των πελατών, όπου σημειώθηκε ότι οι 40.000 κορυφαίοι πελάτες, που αντιπροσωπεύουν περίπου το 42% της συνολικής πελατειακής βάσης, συμβάλλουν περίπου στο 80% των συνολικών πωλήσεων. Οι πελάτες της ομάδας αυτής που κατά την RFM ανάλυση φάνηκε να επηρεάζει σημαντικά τα έσοδα της εταιρείας, πιθανολογείται ότι είναι και η ίδια που αναφέρεται στην αντίστοιχη μέτρηση.

Ταυτόχρονα στην ανάλυση αποκαλύπτεται, επίσης, ένα τμήμα πελατών με υψηλή χρηματική συνεισφορά αλλά χαμηλότερη συχνότητα, γεγονός που υποδηλώνει ότι, ενώ αυτοί οι πελάτες ψωνίζουν λιγότερο συχνά, πραγματοποιούν σημαντικές αγορές όταν το κάνουν. Αυτή η ομάδα μπορεί να περιλαμβάνει premium αγοραστές ή όσους πραγματοποιούν σημαντικές «εφάπαξ» αγορές, αντιπροσωπεύοντας ένα άλλο κρίσιμο τμήμα για στοχευμένες στρατηγικές μάρκετινγκ που θα ενθαρρύνουν πιο συχνές αλληλεπιδράσεις.

Η τμηματοποίηση με βάση τη συνολική βαθμολογία RFM διαιρεί περαιτέρω τους πελάτες σε τμήματα χαμηλής, μεσαίας και υψηλής αξίας, επιτρέποντας στην επιχείρηση να προσαρμόσει ανάλογα τις στρατηγικές μάρκετινγκ και υπηρεσιών της.

Οι πελάτες υψηλής αξίας, που χαρακτηρίζονται από πρόσφατες, συχνές και σημαντικές δαπάνες, αποτελούν τους ιδανικούς στόχους για προσπάθειες διατήρησης και δημιουργίας αφοσίωσης. Αντίθετα, οι στρατηγικές που αποσκοπούν στην αύξηση της δέσμευσης και των δαπανών των τμημάτων χαμηλής και μεσαίας αξίας θα μπορούσαν να επικεντρωθούν σε κίνητρα, εξατομικευμένο μάρκετινγκ και συστάσεις προϊόντων για την αύξηση της αξίας τους με την πάροδο του χρόνου.

Συμπερασματικά, η ανάλυση RFM προσφέρει πολύτιμες πληροφορίες για τη συμπεριφορά των πελατών, αποκαλύπτοντας ευκαιρίες για την ενίσχυση της δέσμευσης σε διάφορα τμήματα. Με την κατανόηση των αποχρώσεων κάθε ομάδας πελατών, η επιχείρηση μπορεί να κατανέμει αποτελεσματικότερα τους πόρους, να προσαρμόζει τις στρατηγικές μάρκετινγκ ώστε να ικανοποιεί τις μοναδικές ανάγκες κάθε τμήματος και τελικά να οδηγεί σε αύξηση των εσόδων και της αφοσίωσης των πελατών.

Στο στάδιο της CLV ανάλυσης παρέχεται μια λεπτομερής ανάλυση της αξίας διάρκειας ζωής του πελάτη στα διάφορα τμήματα πελατών, ενώ στην συνέχεια εστιάζοντας σε μια περίοδο έξι μηνών γίνεται μια περαιτέρω διερεύνηση με σκοπό την λήψη στρατηγικών αποφάσεων μάρκετινγκ. Τα ευρήματα προκύπτουν με βάση τις τιμές Recency, Frequency και Monetary (RFM) που υπολογίστηκαν προηγουμένως. Κατά αυτό τον τρόπο οι ευθυγραμμιζόμενες συστάδες RFM με την CLV του πελάτη συνδυάζονται για την πρόβλεψη της μελλοντικής συμπεριφοράς και της αξίας των πελατών για την εταιρεία.

Η ανάλυση εντόπισε διακριτές ομάδες πελατών με διαφορετικό βαθμό αξίας. Η συστάδα 0, για παράδειγμα, η οποία είναι η μεγαλύτερη σε μέγεθος, περιλαμβάνει πελάτες που μπορεί να έχουν πραγματοποιήσει αγορά μία φορά ή να έχουν δαπανήσει ελάχιστα ποσά, γεγονός που υποδηλώνει τη δυνατότητα για στοχευμένες στρατηγικές μάρκετινγκ με σκοπό την αύξηση της συχνότητας ή της αξίας των αγορών τους. Οι συστάδες 1 και 2, είναι παρόμοιες και παρουσιάζουν προοδευτικά υψηλότερες μέσες τιμές, υποδεικνύοντας πελάτες που ξοδεύουν σημαντικά υψηλότερα ποσά με μέτρια συχνότητα, κατέχοντας έτσι σημαντική αξία για στοχευμένες στρατηγικές διατήρησης και δέσμευσης. Αυτές οι δύο ομάδες είναι πολύ σημαντικές για την επιχείρηση και θα πρέπει να στοχευθούν με στρατηγικές για την ενίσχυση της αφοσίωσης τους, με υπηρεσίες υψηλής ποιότητας, προγράμματα πιστότητας και άλλες τακτικές διατήρησης. Ως τέτοιες ενέργειες ορίζονται συχνά τα προγράμματα επιβράβευσης των καταναλωτών, κατά τα οποία οι πελάτες μετά την πραγματοποίηση μιας αγοράς

κερδίζουν εκπτώσεις/κουπόνια για την επόμενη αγορά τους, ή εντάσσονται σε μια ειδική κατηγορία πελατών που συνήθως ονομάζονται «μέλη» και απολαμβάνουν χαμηλότερες τιμές συγκριτικά με αυτές που προσφέρει το κατάστημα σε νέους πελάτες ή πελάτες χαμηλής αξίας.

Κατά την ανάλυση του διαγράμματος διασποράς μελετήθηκε περαιτέρω η σχέση μεταξύ των βαθμολογιών RFM και της CLV, αποκαλύπτοντας μια εξέλιξη από τα τμήματα πελατών χαμηλής σε τμήματα υψηλής αξίας. Αυτή η ανάλυση υποστηρίζει τη στρατηγική στόχευση των προσπαθειών μάρκετινγκ για την ενίσχυση της αξίας των πελατών σε διάφορα τμήματα, υπογραμμίζοντας την ανάγκη για προσαρμοσμένες προσεγγίσεις για την αποτελεσματική μετακίνηση των πελατών από τμήματα χαμηλότερης σε τμήματα υψηλότερης αξίας. Επιπλέον, διερευνήθηκε η κατανομή των συστάδων με βάση τις επιμέρους μετρήσεις RFM σε διάφορες ομάδες CLV, αναδεικνύοντας ότι οι πρόσφατες στρατηγικές δέσμευσης μπορεί να είναι πιο αποτελεσματικές με πελάτες χαμηλότερης αξίας, ενώ μπορεί να απαιτούνται διαφορετικές προσεγγίσεις για την επαναπροσέγγιση πελατών υψηλότερης αξίας αλλά λιγότερο ενεργών πρόσφατα.

Συνοπτικά, η ανάλυση αποκάλυψε μια εξέλιξη της αξίας των πελατών (CLV) καθώς αυξάνεται η βαθμολογία RFM. Οι χαμηλότερες βαθμολογίες αντιστοιχούν σε χαμηλότερη LTV και βρίσκονται κυρίως στο τμήμα "Χαμηλής Αξίας". Καθώς οι βαθμολογίες αυξάνονται, αυξάνεται και το LTV, με μετατόπιση από τους πελάτες "Χαμηλής Αξίας" στους πελάτες "Μεσαίας Αξίας" και τελικά στους πελάτες "Υψηλής Αξίας". Η οπτική κατανομή του LTV σε κάθε βαθμολογία RFM υποδεικνύει τις δυνατότητες ανάπτυξης και την ανάγκη για διαφοροποιημένες στρατηγικές μάρκετινγκ για κάθε τμήμα.

Εν συνεχεία, εξετάστηκε η αξία διάρκειας ζωής πελάτη συγκριτικά με τις μεταβλητές Recency, Frequency και Monetary ξεχωριστά, προβάλλοντας συγκριτικά τις κατανομές των συστάδων με βάση τη CLV και καθεμία από αυτές τις μεταβλητές. Τα αποτελέσματα αυτής της σύγκρισης, ενίσχυσαν τα συμπεράσματα της RFM ανάλυσης, αναδεικνύοντας την ανάγκη της επιχείρησης να αυξήσει την συχνότητα των αγορών χαμηλής αξίας από το πελατολόγιό της και να ενθαρρύνει τις αγορές υψηλής αξίας διατηρώντας τους πελάτες που κατατάσσονται στις υψηλόβαθμες συστάδες.

Η συγκριτική ανάλυση των μοντέλων μηχανικής μάθησης για την πρόβλεψη της κατηγοριοποίησης CLV υπογραμμίζει την αποτελεσματικότητα ορισμένων μοντέλων

έναντι άλλων. Τα μοντέλα που εμφάνισαν κορυφαίες επιδόσεις, όπως ο ταξινομητής Decision Tree, ο ταξινομητής Random Forest, ο ταξινομητής Extra Trees και η LGBoost Machine, επιδεικνύουν μια σημαντική ακρίβεια πρόβλεψης της μελλοντικής αξίας των πελατών. Ωστόσο, υπάρχει μια σημαντική παρατήρηση που αφορά τη διαφορά μεταξύ της απόδοσης του μοντέλου και του χρόνου εκπαίδευσης, υποδεικνύοντας την ανάγκη εξισορρόπησης της ακρίβειας πρόβλεψης με την υπολογιστική αποδοτικότητα για πρακτικές εφαρμογές.

Εν κατακλείδι, η λεπτομερής ανάλυση CLV υπογραμμίζει τη σημασία της κατανόησης της συμπεριφοράς των πελατών μέσω γνώσεων που βασίζονται σε δεδομένα, επιτρέποντας την ανάπτυξη διαφοροποιημένων στρατηγικών μάρκετινγκ και προγραμμάτων διατήρησης πελατών. Τα αποτελέσματα της προβλεπτικής μοντελοποίησης προσφέρουν μια οδό για την ενίσχυση των επιχειρησιακών στρατηγικών με την ακριβή πρόβλεψη της αξίας διάρκειας ζωής των πελατών, επιτρέποντας έτσι στην επιχείρηση να λαμβάνει τεκμηριωμένες αποφάσεις σχετικά με τη διαχείριση των πελατειακών σχέσεων, την κατανομή των πόρων και τον μακροπρόθεσμο επιχειρηματικό σχεδιασμό. Τελικώς, υπογραμμίζεται η δύναμη της μηχανικής μάθησης ως εργαλείο για την επίτευξη αυτών των γνώσεων, υπό την προϋπόθεση ότι τα μοντέλα επιλέγονται και αξιολογούνται προσεκτικά με βάση τα συγκεκριμένα χαρακτηριστικά και τις απαιτήσεις του συνόλου δεδομένων και του συγκεκριμένου έργου πρόβλεψης.

Στο τελικό βήμα της μελέτης πραγματοποιήθηκε μια σε βάθος ανάλυση συναισθημάτων των κριτικών πελατών, αξιοποιώντας τεχνικές επεξεργασίας φυσικής γλώσσας (NLP) για την κατηγοριοποίηση των συναισθημάτων ως θετικών ή αρνητικών. Αρχικά, παρατηρείται βάση της κατανομής των αξιολογήσεων των πελατών, μια σημαντική πόλωση, κυριαρχεί η υψηλή ικανοποίηση (βαθμολογία 5), ακολουθούμενη από μια σημαντική μειοψηφία που εκφράζει δυσαρέσκεια (βαθμολογία 1). Αυτή η πόλωση υποδηλώνει ότι οι πελάτες τείνουν να δίνουν ακραίες αξιολογήσεις, αντανακλώντας έντονα αντίθετες αντιλήψεις για τα προϊόντα ή τις υπηρεσίες της επιχείρησης.

Η ανάλυση προχώρησε και στην κατηγοριοποίηση των κριτικών σε θετικά ή αρνητικά συναισθήματα με βάση τις αξιολογήσεις τους. Αυτή η κατηγοριοποίηση διευκόλυνε την ανάλυση συναισθήματος, αποκαλύπτοντας ότι η πλειοψηφία των σχολίων είναι θετικά. Αναδεικνύοντας ότι, σε γενικές γραμμές, οι πελάτες είναι ικανοποιημένοι με τις προσφορές της επιχείρησης και υποδεικνύοντας πλεονεκτήματα

που βρίσκουν απήχηση στην πελατειακή βάση. Ωστόσο, η παρουσία αρνητικών σχολίων φανερώνει και τομείς που χρήζουν βελτίωσης, η αντιμετώπιση αυτών των πτυχών για τη βελτίωση της συνολικής ικανοποίησης των πελατών, σίγουρα θα ενισχύσει ακόμα περισσότερο την φήμη της επιχείρησης, το οποίο θα έχει αποτέλεσμα και στα κέρδη αυτής. Η εντελεχής εξέταση των σχολίων, μέσω της ανάλυσης των κοινών διαγραμμάτων των θετικών και αρνητικών κριτικών, προσφέρει χρήσιμες πληροφορίες. Αυτή η ανάλυση βοηθά στον εντοπισμό συγκεκριμένων πτυχών που οι πελάτες αναφέρουν συχνά τόσο σε θετικά όσο και σε αρνητικά πλαίσια, επιτρέποντας στην επιχείρηση να εντοπίσει τους δυνατούς τομείς που πρέπει να αξιοποιήσει και τα ζητήματα που πρέπει να αντιμετωπίσει.

Τέλος, στην ανάλυση συναισθήματος επιχειρήθηκε ξανά η χρήση μηχανικής μάθησης, μέσω χρήσης των ταξινομητών Logistic Regression και Naive Bayes στην κατηγοριοποίηση των σχολίων. Μετά την μελέτη των μέτρων αξιολόγησης αλλά και των πινάκων σύγχυσης για τα δύο μοντέλα, η λογιστική παλινδρόμηση θεωρείται ιδιαίτερα γρήγορη σε χρόνο εκτέλεσης, γεγονός που υποδηλώνει την καταλληλόλητα της για γρήγορη και αξιόπιστη ταξινόμηση συναισθημάτων σε πρακτικές εφαρμογές, αλλά ταυτόχρονα παρουσίασε και την πιο ισορροπημένη απόδοση, υποδηλώνοντας ένα συντηρητικό πρότυπο πρόβλεψης με λιγότερες λανθασμένες ταξινομήσεις.

Συμπερασματικά, η ανάλυση συναισθήματος υπογραμμίζει τους ισχυρούς τομείς της επιχείρησης και τις ευκαιρίες βελτίωσης. Η χρήση μοντέλων μηχανικής μάθησης, ιδίως της λογιστικής παλινδρόμησης, προσφέρει μια αξιόπιστη μέθοδο για την ταξινόμηση των συναισθημάτων των πελατών, παρέχοντας μια βάση για τη λήψη τεκμηριωμένων αποφάσεων στις στρατηγικές μάρκετινγκ και τη διαχείριση των πελατειακών σχέσεων. Η ανάλυση όχι μόνο αποκαλύπτει την τρέχουσα κατάσταση της ικανοποίησης των πελατών, αλλά υποδεικνύει και στρατηγικές κατευθύνσεις για την αξιοποίηση των θετικών ανατροφοδοτήσεων και την αντιμετώπιση των αρνητικών εμπειριών, με απώτερο στόχο τη βελτίωση των προσφορών της επιχείρησης και της θέσης της στην αγορά.

Στο τελικό βήμα της μελέτης πραγματοποιήθηκε μια σε βάθος ανάλυση συναισθημάτων των κριτικών πελατών, αξιοποιώντας τεχνικές επεξεργασίας φυσικής γλώσσας (NLP) για την κατηγοριοποίηση των συναισθημάτων ως θετικών ή αρνητικών. Αρχικά, παρατηρείται βάση της κατανομής των αξιολογήσεων των πελατών, μια σημαντική πόλωση, κυριαρχεί η υψηλή ικανοποίηση (βαθμολογία 5), ακολουθούμενη από μια σημαντική μειοψηφία που εκφράζει δυσαρέσκεια

(βαθμολογία 1). Αυτή η πόλωση υποδηλώνει ότι οι πελάτες τείνουν να δίνουν ακραίες αξιολογήσεις, αντανακλώντας έντονα αντίθετες αντιλήψεις για τα προϊόντα ή τις υπηρεσίες της επιχείρησης.

Η ανάλυση προχώρησε και στην κατηγοριοποίηση των κριτικών σε θετικά ή αρνητικά συναισθήματα με βάση τις αξιολογήσεις τους. Αυτή η κατηγοριοποίηση διευκόλυνε την ανάλυση συναισθήματος, αποκαλύπτοντας ότι η πλειοψηφία των σχολίων είναι θετικά. Αναδεικνύοντας ότι, σε γενικές γραμμές, οι πελάτες είναι ικανοποιημένοι με τις προσφορές της επιχείρησης και υποδεικνύοντας πλεονεκτήματα που βρίσκουν απήχηση στην πελατειακή βάση. Ωστόσο, η παρουσία αρνητικών σχολίων φανερώνει και τομείς που χρήζουν βελτίωσης, η αντιμετώπιση αυτών των πτυχών για τη βελτίωση της συνολικής ικανοποίησης των πελατών, σίγουρα θα ενισχύσει ακόμα περισσότερο την φήμη της επιχείρησης, το οποίο θα έχει αποτέλεσμα και στα κέρδη αυτής. Η εντελεχής εξέταση των σχολίων, μέσω της ανάλυσης των κοινών διαγραμμάτων των θετικών και αρνητικών κριτικών, προσφέρει χρήσιμες πληροφορίες. Αυτή η ανάλυση βοηθά στον εντοπισμό συγκεκριμένων πτυχών που οι πελάτες αναφέρουν συχνά τόσο σε θετικά όσο και σε αρνητικά πλαίσια, επιτρέποντας στην επιχείρηση να εντοπίσει τους δυνατούς τομείς που πρέπει να αξιοποιήσει και τα ζητήματα που πρέπει να αντιμετωπίσει.

Τέλος, στην ανάλυση συναισθήματος επιχειρήθηκε ξανά η χρήση μηχανικής μάθησης, μέσω χρήσης των ταξινομητών Logistic Regression και Naive Bayes στην κατηγοριοποίηση των σχολίων. Μετά την μελέτη των μέτρων αξιολόγησης αλλά και των πινάκων σύγχυσης για τα δύο μοντέλα, η λογιστική παλινδρόμηση θεωρείται ιδιαίτερα γρήγορη σε χρόνο εκτέλεσης, γεγονός που υποδηλώνει την καταλληλόλητα της για γρήγορη και αξιόπιστη ταξινόμηση συναισθημάτων σε πρακτικές εφαρμογές, αλλά ταυτόχρονα παρουσίασε και την πιο ισορροπημένη απόδοση, υποδηλώνοντας ένα συντηρητικό πρότυπο πρόβλεψης με λιγότερες λανθασμένες ταξινομήσεις.

Συμπερασματικά, η ανάλυση συναισθήματος υπογραμμίζει τους ισχυρούς τομείς της επιχείρησης και τις ευκαιρίες βελτίωσης. Η χρήση μοντέλων μηχανικής μάθησης, ιδίως της λογιστικής παλινδρόμησης, προσφέρει μια αξιόπιστη μέθοδο για την ταξινόμηση των συναισθημάτων των πελατών, παρέχοντας μια βάση για τη λήψη τεκμηριωμένων αποφάσεων στις στρατηγικές μάρκετινγκ και τη διαχείριση των πελατειακών σχέσεων. Η ανάλυση όχι μόνο αποκαλύπτει την τρέχουσα κατάσταση της ικανοποίησης των πελατών, αλλά υποδεικνύει και στρατηγικές κατευθύνσεις για την αξιοποίηση των θετικών ανατροφοδοτήσεων και την αντιμετώπιση των

αρνητικών εμπειριών, με απώτερο στόχο τη βελτίωση των προσφορών της επιχείρησης και της θέσης της στην αγορά.

Ο λόγος που επιλέχθηκε αυτή η μελέτη συνδυαστικά με τις μεθόδους κατηγοριοποίησης πελατών σε ομάδες βάση της μελέτης RFM και της τιμής CLV, είναι γιατί η ανάλυση συναισθήματος είναι απαραίτητη για την κατανόηση των ανατροφοδοτήσεων των πελατών χωρίς να εξετάζεται χειροκίνητα κάθε σχόλιο, με σκοπό την κατανόηση της διάθεσης των πελατών απέναντι στην επιχείρηση. Αυτή η αυτοματοποιημένη ανάλυση επιτρέπει στις επιχειρήσεις:

1. Εντοπισμό πληροφοριών σχετικά με το προϊόν: Να κατανοήσουν ποιες πτυχές ενός προϊόντος ή μιας υπηρεσίας αρέσουν ή δεν αρέσουν στους πελάτες.
2. Εξυπηρέτηση πελατών: Γρήγορος εντοπισμός και αντιμετώπιση αρνητικών κριτικών ή παραπόνων.
3. Έρευνα αγοράς: Μετρήστε το κοινό αίσθημα και συγκρίνετέ το με τους ανταγωνιστές.
4. Στοχευμένο μάρκετινγκ: Αναπτύξτε στρατηγικές μάρκετινγκ που να ανταποκρίνονται σε αυτά για τα οποία οι πελάτες είναι ευχαριστημένοι ή δυσαρεστημένοι.

Αυτοματοποιώντας τη διαδικασία με μοντέλα μηχανικής μάθησης όπως η λογιστική παλινδρόμηση και το Naive Bayes, οι εταιρείες μπορούν να επεξεργάζονται αποτελεσματικά μεγάλους όγκους δεδομένων κειμένου, αποκτώντας πληροφορίες που θα ήταν ανέφικτο να συγκεντρωθούν χειροκίνητα λόγω του μεγάλου όγκου των δεδομένων. Αυτό επιτρέπει την ανάλυση συναισθήματος σε πραγματικό χρόνο, η οποία είναι ζωτικής σημασίας στη σημερινή ταχέως εξελισσόμενη αγορά, όπου η κοινή γνώμη μπορεί να αλλάξει γρήγορα. Ταυτόχρονα ο συνδυασμός της ανάλυσης αυτής με τα προηγούμενα μπορεί να δώσει στην επιχείρηση την δυνατότητα εύρεσης της καταλληλότερης μεθόδου προσέγγισης για κάθε πελάτη ξεχωριστά.

Κεφάλαιο 6: Περιορισμοί

Η μελέτη αυτή συνέβαλε σημαντικά στην κατανόηση της συμπεριφοράς των πελατών και της προβλεπτικής δύναμης των μοντέλων μηχανικής μάθησης στην εκτίμηση της αξίας ζωής των πελατών (CLV). Ωστόσο, όπως κάθε έρευνα, έχει τους περιορισμούς της, οι οποίοι ανοίγουν δρόμους για μελλοντική έρευνα.

Πρώτο βήμα για εξέλιξη της έρευνας αυτής θα ήταν η βελτίωση της τμηματοποίησης με βάση την ανάλυση RFM. Το μοντέλο RFM, το οποίο κατηγοριοποιεί τους πελάτες βάσει της συχνότητας, της συχνότητας και της χρηματικής αξίας των αγορών τους, είναι ένα ευρέως υιοθετημένο μοντέλο τμηματοποίησης που θα μπορούσε να βελτιώσει την κατανόηση της συμπεριφοράς των πελατών. Η ανάλυση της σιλουέτας έδειξε ότι, ενώ η τμηματοποίηση παρέχει ένα χρήσιμο πλαίσιο για την κατανόηση της συμπεριφοράς των πελατών, υπάρχει περιθώριο βελτίωσης για την επίτευξη πιο ευδιάκριτου διαχωρισμού και συνοχής εντός κάθε ομάδας. Η μελλοντική έρευνα θα μπορούσε να διερευνήσει μια πιο διαφοροποιημένη τμηματοποίηση με την ενσωμάτωση της ανάλυσης RFM για την επίτευξη σαφέστερης διάκρισης και συνοχής εντός κάθε ομάδας. Η προσέγγιση αυτή θα μπορούσε να βελτιώσει την αποτελεσματικότητα των στοχευμένων εκστρατειών μάρκετινγκ και των στρατηγικών διαχείρισης πελατειακών σχέσεων.

Επεκτείνοντας την ανάλυση στην πρόβλεψη της CLV, θα μπορούσε να εξεταστεί περισσότερο η προβλεψιμότητα της αξίας διάρκειας ζωής πελάτη. Η ακριβής πρόβλεψη της CLV επιτρέπει τη λήψη πιο τεκμηριωμένων αποφάσεων σχετικά με τη διαχείριση των πελατειακών σχέσεων, τις στρατηγικές μάρκετινγκ, την κατανομή των πόρων και τον μακροπρόθεσμο επιχειρηματικό σχεδιασμό. Ενώ η παρούσα μελέτη καταδεικνύει τις δυνατότητες της μηχανικής μάθησης ως ισχυρό εργαλείο για την επίτευξη αυτών των γνώσεων, προϋποθέτει την προσεκτική επιλογή και αξιολόγηση των μοντέλων. Μελλοντικές μελέτες θα μπορούσαν να εμβαθύνουν σε υβριδικά μοντέλα που συνδυάζουν τα πλεονεκτήματα διαφόρων αλγορίθμων για τη βελτίωση της ακρίβειας πρόβλεψης και της επιχειρησιακής αποδοτικότητας.

Ο εντοπισμός των μοντέλων με τις καλύτερες επιδόσεις είναι ένα κρίσιμο βήμα, το οποίο ακολουθείται από την περαιτέρω βελτίωσή τους (εάν είναι απαραίτητο) και

την ανάπτυξή τους για την πρόβλεψη της CLV για υφιστάμενους ή νέους πελάτες. Ο απώτερος στόχος είναι να χρησιμοποιηθούν αυτές οι προβλέψεις για την εφαρμογή επιχειρησιακών στρατηγικών που μπορούν να αναλάβουν δράση, όπως εξατομικευμένες εκστρατείες μάρκετινγκ και βελτιστοποίηση της αξίας διάρκειας ζωής των πελατών. Η έρευνα θα μπορούσε να επικεντρωθεί στη συνεχή βελτίωση αυτών των μοντέλων και στην ενσωμάτωση δεδομένων πραγματικού χρόνου για δυναμικές δυνατότητες πρόβλεψης.

Η ανάλυση έδειξε ότι τα μοντέλα εκπαιδεύτηκαν αποτελεσματικά και είναι ικανά να προβλέπουν ουσιαστικά το CLV. Ωστόσο, η επιλογή του τελικού μοντέλου (ή των τελικών μοντέλων) για ανάπτυξη θα εξαρτηθεί από συγκεκριμένες επιχειρηματικές ανάγκες, συμπεριλαμβανομένης της εξισορρόπησης της ακρίβειας πρόβλεψης με τους διαθέσιμους υπολογιστικούς πόρους. Η μελλοντική έρευνα θα μπορούσε να διερευνήσει οικονομικά αποδοτικές υπολογιστικές προσεγγίσεις ή λύσεις βασισμένες στο υπολογιστικό νέφος για την ενίσχυση της επεκτασιμότητας και της προσβασιμότητας των μοντέλων για επιχειρήσεις διαφορετικών μεγεθών.

Επιπρόσθετα οι τιμές AUC που παρουσιάστηκαν στα αποτελέσματα ήταν ασυνήθιστα χαμηλές και φαίνεται να μην συνάδουν με άλλες μετρήσεις επιδόσεων, γεγονός που θα μπορούσε να υποδηλώνει πιθανό σφάλμα είτε στην αναφορά των εν λόγω τιμών είτε στη μεθοδολογία που χρησιμοποιήθηκε για τον υπολογισμό τους. Αυτή η ασυμφωνία υπογραμμίζει τη σημασία των αυστηρών μεθόδων επικύρωσης στην προγνωστική μοντελοποίηση. Απαιτείται περαιτέρω έρευνα για τη διερεύνηση των αιτιών αυτών των ανωμαλιών και την ανάπτυξη κατευθυντήριων γραμμών για τη διασφάλιση της αξιοπιστίας και της εγκυρότητας των αξιολογήσεων των μοντέλων. Ενώ η παρούσα μελέτη παρείχε πολύτιμες πληροφορίες σχετικά με την εφαρμογή της μηχανικής μάθησης για την ανάλυση της συμπεριφοράς των πελατών και την πρόβλεψη του CLV, οι περιορισμοί αυτοί υπογραμμίζουν την ανάγκη για συνεχή έρευνα. Οι μελλοντικές μελέτες θα πρέπει να στοχεύουν στην αντιμετώπιση αυτών των προκλήσεων, στη διερεύνηση νέων μεθοδολογιών και στη συνεχή βελτίωση των μοντέλων πρόβλεψης ώστε να συμβαδίζουν με την εξελισσόμενη δυναμική της αγοράς και τις τεχνολογικές εξελίξεις.

Η διερεύνηση της ανάλυσης συναισθήματος, όπως παρουσιάζεται στη μελέτη, αποτελεί σημαντικό βήμα προς μια πιο διαφοροποιημένη κατανόηση της αφοσίωσης και της ικανοποίησης των πελατών. Επεκτείνοντας την έρευνα πέρα από τις αναλύσεις RFM και CLV για να συμπεριλάβει την ανάλυση συναισθήματος των κριτικών των

πελατών, η μελέτη εμβαθύνει στο πλούσιο μωσαϊκό των συναισθημάτων και των απόψεων των πελατών. Χρησιμοποιώντας εργαλεία όπως η βιβλιοθήκη nltk για προεπεξεργασία και το CountVectorizer από τη βιβλιοθήκη sklearn για ανάλυση κειμένου, η μεθοδολογία μετατρέπει τα σχόλια κειμένου σε δομημένη μορφή που μπορούν να ερμηνεύσουν τα μοντέλα μηχανικής μάθησης. Η διαδικασία αυτή όχι μόνο κατηγοριοποιεί τα σχόλια σε θετικά ή αρνητικά συναισθήματα με βάση τις σχετικές βαθμολογίες τους, αλλά στοχεύει επίσης στην αυτοματοποίηση της ταξινόμησης με τη χρήση αλγορίθμων μηχανικής μάθησης όπως η λογιστική παλινδρόμηση και ο Naive Bayes. Παρόλα αυτά, η προσέγγιση της ανάλυσης συναισθήματος έχει τους περιορισμούς της. Η εξάρτηση από προκαθορισμένες βιβλιοθήκες και οι προκλήσεις στην ακριβή ερμηνεία της διαφοροποιημένης γλώσσας των κριτικών πελατών αναδεικνύουν την πολυπλοκότητα της ανάλυσης συναισθήματος. Επιπλέον, η δυαδική ταξινόμηση των συναισθημάτων σε θετικά ή αρνητικά μπορεί να υπεραπλουστεύσει το φάσμα των συναισθημάτων των πελατών, παραβλέποντας ενδεχομένως λεπτές ενδείξεις που θα μπορούσαν να προσφέρουν βαθύτερες γνώσεις σχετικά με την ικανοποίηση και την αφοσίωση των πελατών.

Η μελλοντική έρευνα θα μπορούσε να διερευνήσει πιο εξελιγμένες τεχνικές και μοντέλα NLP που να καταγράφουν ένα ευρύτερο φάσμα συναισθημάτων και συναισθηματικών τόνων, παρέχοντας μια πιο λεπτομερή κατανόηση των ανατροφοδοτήσεων των πελατών. Επιπλέον, η ενσωμάτωση πολύγλωσσης ανάλυσης για την εξυπηρέτηση μιας παγκόσμιας πελατειακής βάσης και η χρήση αλγορίθμων μάθησης χωρίς επίβλεψη για τον εντοπισμό αναδυόμενων θεμάτων ή συναισθημάτων στις αξιολογήσεις των πελατών θα μπορούσε να ενισχύσει περαιτέρω την ακρίβεια πρόβλεψης και τη σημασία της ανάλυσης συναισθήματος για την κατανόηση της συμπεριφοράς και της αφοσίωσης των πελατών.

Βιβλιογραφία

Abbey, J. D., Meloy, M. G., Guide, V. D. R., & Atalay, S. (2015). Remanufactured Products in Closed-Loop Supply Chains for Consumer Goods. *Production and Operations Management*, *24*(3), 488–503. https://doi.org/10.1111/poms.12238

Adelaar, T. (2000). Electronic Commerce and the Implications for Market Structure: The Example of the Art and Antiques Trade. *J. Computer-Mediated Communication*, *5*.

Agarwal, R., & Dhingra, S. (2023). Factors influencing cloud service quality and their relationship with customer satisfaction and loyalty. *Heliyon*, *9*, 15177. https://doi.org/10.1016/j.heliyon.2023.e15177

Alam, G. M., Taher, M., & Khalifa, B. (2009). The impact of introducing a business marketing approach to education: A study on private HE in Bangladesh. *Article in AFRICAN JOURNAL OF BUSINESS MANAGEMENT*. https://doi.org/10.5897/AJBM09.202

Alamsyah, A., Prasetyo, P. E., Sunyoto, S., Bintari, S. H., Saputro, D. D., Rohman, S., & Pratama, R. N. (2022). Customer Segmentation Using the Integration of the Recency Frequency Monetary Model and the K-Means Cluster Algorithm. *Scientific Journal of Informatics*, *9*(2), 189–196. https://doi.org/10.15294/sji.v9i2.39437

Al-dweeri, R. M., Obeidat, Z. M., Al-dwiry, M. A., Alshurideh, M. T., & Alhorani, A. M. (2017). The Impact of E-Service Quality and E-Loyalty on Online Shopping: Moderating Effect of E-Satisfaction and E-Trust. *International Journal of Marketing Studies*, *9*(2), 92. https://doi.org/10.5539/ijms.v9n2p92

Almeida, M. da G. M. C., & Coelho, A. F. M. (2019). The Antecedents of Corporate Reputation and Image and Their Impacts on Employee Commitment and Performance: The Moderating Role of CSR. *Corporate Reputation Review*, *22*(1), 10–25. https://doi.org/10.1057/s41299-018-0053-8

Aramburu, I. A., & Pescador, I. G. (2019). The Effects of Corporate Social Responsibility on Customer Loyalty: The Mediating Effect of Reputation in

Cooperative Banks Versus Commercial Banks in the Basque Country. *Journal of Business Ethics*, *154*(3), 701–719. https://doi.org/10.1007/s10551-017-3438-1

Armstrong, M., & Wright, J. (2007). Two-sided Markets, Competitive Bottlenecks and Exclusive Contracts. *Economic Theory*, *32*(2), 353–380. https://doi.org/10.1007/s00199-006-0114-6

Arthur Middleton Hughes. (1994). *Strategic Database Marketing: The Masterplan for Starting and Managing a Profitable, Customer-based Marketing Program* (illustrated). Probus Publishing Company.

Ashton, C. F. (2020). The fillip of the World Wide Web: Rewriting and expanding history. *Collegian*, *27*, 595–599. https://doi.org/10.1016/j.colegn.2020.08.007

Askariazad, M. H., & Babakhani, N. (2015). An application of European Customer Satisfaction Index (ECSI) in business to business (B2B) context. *Journal of Business & Industrial Marketing*, *30*(1), 17–31. https://doi.org/10.1108/JBIM-07-2011-0093

Asnawi, N., & Setyaningsih, N. D. (2020). The Role of DART Key Building Blocks as Customer Co-Creation Determinants in Islamic Banking Services. *Journal of Southwest Jiaotong University*, *55*(6). https://doi.org/10.35741/issn.0258-2724.55.6.38

Badase, P. S., Deshbhratar, G. P., & Bhagat, A. P. (2015). Classification and analysis of clustering algorithms for large datasets. *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 1–5. https://doi.org/10.1109/ICIIECS.2015.7193191

Bhattacherjee, A. (2001). Understanding Information Systems Continuance: An Expectation-Confirmation Model. *MIS Quarterly*, *25*(3), 351. https://doi.org/10.2307/3250921

Birant, D. (n.d.). *Data Mining Using RFM Analysis*. www.intechopen.com

Blattberg, R., C., R., Byung-Do, Kim, Neslin, S., & A., N. (2008). *Database Marketing: Analyzing and Managing Customers*. Springer.

Blut, M. (2016). E-Service Quality: Development of a Hierarchical Model. *Journal of Retailing*, *92*(4), 500–517. https://doi.org/10.1016/J.JRETAI.2016.09.002

Boonlertvanich, K. (2019). Service quality, satisfaction, trust, and loyalty: the moderating role of main-bank and wealth status. *International Journal of Bank Marketing*, *37*(1). https://doi.org/10.1108/IJBM-02-2018-0021

Bowen, J. T., & Mccain, S.-L. C. (2015). Transitioning loyalty programs A commentary on “the relationship between customer loyalty and customer satisfaction.” *International Journal of Contemporary Hospitality Management*, *27*(3), 415–430. https://doi.org/10.1108/IJCHM-07-2014-0368

Cai, X., Cebollada, J., & Cortiñas, M. (2023). Impact of seller- and buyer-created content on product sales in the electronic commerce platform: The role of informativeness, readability, multimedia richness, and extreme valence. *Journal of Retailing and Consumer Services*, *70*, 103141. https://doi.org/10.1016/J.JRETCONSER.2022.103141

Cano, J. A., Allec Londoño-Pineda, A., Campo, E. A., & Fernández, S. A. (2023). Sustainable business models of e-marketplaces: An analysis from the consumer perspective. *Journal of Open Innovation: Technology, Market, and Complexity*, *9*, 2199–8531. https://doi.org/10.1016/j.joitmc.2023.100121

Caruana, A. (2002). Service loyalty: the effects of service quality and the mediating role of customer satisfaction. *European Journal of Marketing*, *36*(7/8), 811–828. https://doi.org/10.1108/03090560210430818

Chen, M. C., Chiu, A. L., & Chang, H. H. (2005). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, *28*(4), 773–781. https://doi.org/10.1016/J.ESWA.2004.12.033

Chen, Y.-H., Wang, X., Wang, Y.-Y., & Tsai, S.-C. (2010). The moderating effect of retailer image on customers’ satisfaction-loyalty link. *2010 7th International Conference on Service Systems and Service Management*, 1–6. https://doi.org/10.1109/ICSSSM.2010.5530163

Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications*, *36*(3 PART 1), 4176–4184. https://doi.org/10.1016/J.ESWA.2008.04.003

Childers, T. L., Carr, C. L., Peck, J., & Carson, S. (2001). Hedonic and utilitarian motivations for online retail shopping behavior. *Journal of Retailing*, *77*(4), 511–535. https://doi.org/10.1016/S0022-4359(01)00056-2

Christy, A. J., 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*(10), 1251–1257. https://doi.org/10.1016/j.jksuci.2018.09.004

Claudio Marcus. (1998). A practical yet meaningful approach to customer segmentation. *JOURNAL OF CONSUMER MARKETING*, *15*(5), 494–504. https://doi.org/10.1108/07363769810235974

Condorelli, D., & Padilla, J. (2020). Harnessing Platform Envelopment in the Digital World. *Journal of Competition Law & Economics*, *16*(2), 143–187. https://doi.org/10.1093/joclec/nhaa006

Cronin, J. J., & Taylor, S. A. (1992). Measuring Service Quality: A Reexamination and Extension. *Journal of Marketing*, *56*(3), 55–68. https://doi.org/10.1177/002224299205600304

Dapena-Baron, M., Gruen, T. W., & Guo, L. (2020). Heart, head, and hand: a tripartite conceptualization, operationalization, and examination of brand loyalty. *Journal of Brand Management*, *27*(3), 355–375. https://doi.org/10.1057/s41262-019-00185-3

Delre, S. A., & Luffarelli, J. (2023). Consumer reviews and product life cycle: On the temporal dynamics of electronic word of mouth on movie box office. *Journal of Business Research*, *156*, 113329. https://doi.org/10.1016/J.JBUSRES.2022.113329

Deng, Z., Zhu, Z., Johanson, M., & Hilmersson, M. (2021). *Rapid internationalization and exit of exporters: The role of digital platforms*. https://doi.org/10.1016/j.ibusrev.2021.101896

Edeh, F. O., Teoh, K. B., Murugan, Y., Kee, D. M. H., Wong, J., Wong, X. S., Maheswaran, Y., & Jacinta, O. B. (2021). Contributing Factors to Apple's Sustainability in Malaysia's Information and Communication Technology Industry. *Asia Pacific Journal of Management and Education*, *4*(2), 74–84. https://doi.org/10.32535/apjme.v4i2.1145

Elwalda, A., Lü, K., & Ali, M. (2016). Perceived derived attributes of online customer reviews. *Computers in Human Behavior*, *56*, 306–319. https://doi.org/10.1016/J.CHB.2015.11.051

Fader, P. S., Hardie, B. G. S., & Berger, P. D. (2004). *Customer-Base Analysis with Discrete-Time Transaction Data*. www.petefader.com

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). RFM and CLV: Using iso-value curves for customer base analysis. In *Journal of Marketing Research* (Vol. 42, Issue 4, pp. 415–430). American Marketing Association. https://doi.org/10.1509/jmkr.2005.42.4.415

Fan, S., Lau, R. Y. K., & Zhao, J. L. (2015). Demystifying Big Data Analytics for Business Intelligence Through the Lens of Marketing Mix. In *Big Data Research* (Vol. 2, Issue 1, pp. 28–32). Elsevier Inc. https://doi.org/10.1016/j.bdr.2015.02.006

Feike, M., & Rösch, J. (2024a). Nuanced but important: A literature-based comparison between B2B and B2C platforms. *Decision Analytics Journal*, *10*, 100383. https://doi.org/10.1016/j.dajour.2023.100383

Feike, M., & Rösch, J. (2024b). Nuanced but important: A literature-based comparison between B2B and B2C platforms. *Decision Analytics Journal*, *10*, 100383. https://doi.org/10.1016/j.dajour.2023.100383

Filieri, R., Acikgoz, F., Ndou, V., & Dwivedi, Y. (2021). Is TripAdvisor still relevant? The influence of review credibility, review usefulness, and ease of use on consumers' continuance intention. *International Journal of Contemporary Hospitality Management*, *33*(1), 199–223. https://doi.org/10.1108/IJCHM-05-2020-0402

Frota Neto, J. Q., Bloemhof, J., & Corbett, C. (2016). Market prices of remanufactured, used and new items: Evidence from eBay. *International Journal of Production Economics*, *171*, 371–380. https://doi.org/10.1016/J.IJPE.2015.02.006

Gajewska, T., Zimon, D., Kaczor, G., & Madzík, P. (2019). The impact of the level of customer satisfaction on the quality of e-commerce services. *International Journal of Productivity and Performance Management*, *69*(4), 666–684. https://doi.org/10.1108/IJPPM-01-2019-0018

Giaglis, G., Klein, S., Robert M., & O'Keefe. (2002). The role of intermediaries in electronic marketplaces: developing a contingency model. *Information Systems Journal*, *12*, 231–246. https://www.researchgate.net/publication/2506738

Gierlich-Joas, M., Schüritz, R., & Hess, T. (2019). *SMEs' Approaches for Digitalization in Platform Ecosystems*. https://www.researchgate.net/publication/339875189

Grover, R., & Vriens, M. (2006). *The Handbook of Marketing Research*. SAGE Publications, Inc. https://doi.org/10.4135/9781412973380

Güçdemir, H., & Selim, H. (2015). Integrating multi-criteria decision making and clustering for business customer segmentation. *Industrial Management & Data Systems*, *115*(6), 1022–1040. https://doi.org/10.1108/IMDS-01-2015-0027

Guimaraes, T., & Paranjape, K. (2014). Testing cloud computing for customer satisfaction and loyalty. *International Journal of Electronic Customer Relationship Management*, *8*(1/2/3), 72–86. https://doi.org/10.1504/IJECRM.2014.066885

Guo, J., Wang, X., & Wu, Y. (2020). Positive emotion bias: Role of emotional content from online customer reviews in purchase decisions. *Journal of Retailing and Consumer Services*, *52*, 101891. https://doi.org/10.1016/j.jretconser.2019.101891

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., & Sriram, S. (2006a). Modeling Customer Lifetime Value. *Journal of Service Research*, *9*(2), 139–155. https://doi.org/10.1177/1094670506293810

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., & Sriram, S. (2006b). Modeling customer lifetime value. In *Journal of Service Research* (Vol. 9, Issue 2, pp. 139–155). https://doi.org/10.1177/1094670506293810

Gupta, S., & Lehmann, D. R. (2003). Customers as assets. *Journal of Interactive Marketing*, *17*(1), 9–24. https://doi.org/10.1002/dir.10045

Gupta, S., Lehmann, D. R., & Stuart, J. A. (2004). Valuing Customers. *Journal of Marketing Research*, *41*(1), 7–18. https://doi.org/10.1509/jmkr.41.1.7.25084

Gustriansyah, R., Suhandi, N., & Antony, F. (2019a). Clustering optimization in RFM analysis based on k-means. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(1), 470–477. https://doi.org/10.11591/ijeecs.v18.i1.pp470-477

Gustriansyah, R., Suhandi, N., & Antony, F. (2019b). Clustering optimization in RFM analysis based on k-means. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(1), 470–477. https://doi.org/10.11591/ijeecs.v18.i1.pp470-477

Hailu, T. T., & Tilahun, T. (2014). Linguistic Localization of Opinion Mining From Amharic Blogs. *International Journal of Information Technology & Computer Sciences*, *3*(1), 2319–9024. https://www.researchgate.net/publication/283568620

Han, H., Yu, J., & Kim, W. (2019). Environmental corporate social responsibility and the strategy to boost the airline's image and customer loyalty intentions. *Journal of Travel & Tourism Marketing*, *36*(3), 371–383. https://doi.org/10.1080/10548408.2018.1557580

He, X., & Li, C. (2017). The Research and Application of Customer Segmentation on E-Commerce Websites. *Proceedings - 2016 International Conference on Digital Home, ICDH 2016*, 203–208. https://doi.org/10.1109/ICDH.2016.050

Helgesen, Ø. (2006). Are Loyal Customers Profitable? Customer Satisfaction, Customer (Action) Loyalty and Customer Profitability at the Individual Level. *Journal of Marketing Management*, *22*(3–4), 245–266. https://doi.org/10.1362/026725706776861226

Hossain, M. A., & Quaddus, M. (2012). Expectation–Confirmation Theory in Information System Research: A Review and Analysis. In *Information Systems Theory* (Vol. 28, pp. 441–469). Springer. https://doi.org/10.1007/978-1-4419-6108-2_21

Hsiao, C.-H. (2018). The effects of post-adoption beliefs on the expectation–confirmation model in an electronics retail setting. *Total Quality Management & Business Excellence*, *29*(7–8), 866–880. https://doi.org/10.1080/14783363.2016.1250621

Hsin Hung Wu, En Chi Chang, & Chiao Fang Lo. (2009). Applying RFM model and K-means method in customer value analysis of an outfitter. *Global Perspective for Competitive Enterprise, Economy and Ecology*, 665–672.

Hsu, S. H. (2008). Developing an index for online customer satisfaction: Adaptation of American Customer Satisfaction Index. *Expert Systems with Applications*, *34*(4), 3033–3042. https://doi.org/10.1016/J.ESWA.2007.06.036

Jap, S. D., Gibson, W., & Zmuda, D. (n.d.). *Winning the new channel war on Amazon and third-party platforms*. https://doi.org/10.1016/j.bushor.2021.04.003

Kandampully, J., & Suhartanto, D. (2000). Customer loyalty in the hotel industry: the role of customer satisfaction and image. *International Journal of Contemporary Hospitality Management*, *12*(6), 346–351. https://doi.org/10.1108/09596110010342559

Kasiri, L. A., Guan Cheng, K. T., Sambasivan, M., & Sidin, S. M. (2017). Integration of standardization and customization: Impact on service quality, customer satisfaction, and loyalty. *Journal of Retailing and Consumer Services*, *35*, 91–97. https://doi.org/10.1016/J.JRETCONSER.2016.11.007

Kassim, N., & Asiah Abdullah, N. (2010). The effect of perceived service quality dimensions on customer satisfaction, trust, and loyalty in e-commerce settings.

Asia Pacific Journal of Marketing and Logistics, *22*(3), 351–371. https://doi.org/10.1108/13555851011062269

Kastouni, M. Z., & Lahcen, A. A. (2020). *Big data analytics in telecommunications: Governance, architecture and use cases*. https://doi.org/10.1016/j.jksuci.2020.11.024

Kaushik, K., Mishra, R., Rana, N. P., & Dwivedi, Y. K. (2018). Exploring reviews and review sequences on e-commerce platform: A study of helpful reviews on Amazon.in. *Journal of Retailing and Consumer Services*, *45*, 21–32. https://doi.org/10.1016/J.JRETCONSER.2018.08.002

Khandai, S., Mathew, J., Yadav, R., Kataria, S., & Kohli, H. (2023). Ensuring brand loyalty for firms practising sustainable marketing: a roadmap. *Society and Business Review*, *18*(3), 219–243. https://doi.org/10.1108/SBR-10-2021-0189

Kim, M. K., Park, M. C., & Jeong, D. H. (2004). The effects of customer satisfaction and switching barrier on customer loyalty in Korean mobile telecommunication services. *Telecommunications Policy*, *28*(2), 145–159. https://doi.org/10.1016/J.TELPOL.2003.12.003

Koen van Gelder. (2023). E-commerce worldwide - statistics & facts. *Statista*.

Kotler, P. (1974). Marketing during Periods of Shortage. *Journal of Marketing*, *38*(3), 20. https://doi.org/10.2307/1249846

Kumar, V. (2007). Customer Lifetime Value – The Path to Profitability. *Foundations and Trends® in Marketing*, *2*(1), 1–96. https://doi.org/10.1561/1700000004

Lee, Y. J., Ha, S., & Johnson, Z. (2019). Antecedents and consequences of flow state in e-commerce. *Journal of Consumer Marketing*, *36*(2), 264–275. https://doi.org/10.1108/JCM-10-2015-1579

Li, P., Wang, C., Wu, J., & Madlenak, R. (2022). An E-commerce Customer Segmentation Method based on RFM Weighted K-means. *Proceedings - 2022 International Conference on Management Engineering, Software Engineering and Service Sciences, ICMSS 2022*, 61–68. https://doi.org/10.1109/ICMSS55574.2022.00017

Li, Y., Chu, X., Tian, D., Feng, J., & Mu, W. (2021). Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. *Applied Soft Computing*, *113*. https://doi.org/10.1016/j.asoc.2021.107924

Liat, C. B., Mansori, S., & Huei, C. T. (2014). The Associations Between Service Quality, Corporate Image, Customer Satisfaction, and Loyalty: Evidence From

the Malaysian Hotel Industry. *Journal of Hospitality Marketing & Management*, *23*(3), 314–326. https://doi.org/10.1080/19368623.2013.796867

Lin, S.-Y. (2010). A review of the application of RFM model. In *Article in AFRICAN JOURNAL OF BUSINESS MANAGEMENT*. https://www.researchgate.net/publication/228399859

Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, C. C., Chu, S. W., Chan, Y. K., & Yu, S. S. (2014). A Modified K-Means Algorithm - Two-Layer K-Means Algorithm. *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 447–450. https://doi.org/10.1109/IIH-MSP.2014.118

Liu, D.-R., & Shih, Y.-Y. (2004). *Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences*. https://doi.org/10.1016/j.jss.2004.08.031

Liu, D.-R., & Shih, Y.-Y. (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value. *Information & Management*, *42*(3), 387–400. https://doi.org/10.1016/j.im.2004.01.008

Liu, F., Lim, E. T. K., Li, H., Tan, C. W., & Cyr, D. (2020). Disentangling utilitarian and hedonic consumption behavior in online shopping: An expectation disconfirmation perspective. *Information & Management*, *57*(3), 103199. https://doi.org/10.1016/J.IM.2019.103199

Liu, W.-K., Lee, Y.-S., & Hung, L.-M. (2017). The interrelationships among service quality, customer satisfaction, and customer loyalty: Examination of the fast-food industry. *Journal of Foodservice Business Research*, *20*(2), 146–162. https://doi.org/10.1080/15378020.2016.1201644

Loginova, O. (2022). Branded websites and marketplace selling: Competing during COVID-19. *Journal of Economic Behavior and Organization*, *203*, 577–592. https://doi.org/10.1016/j.jebo.2022.09.020

Ludwig, S., de Ruyter, K., Friedman, M., Brüggen, E. C., Wetzels, M., & Pfann, G. (2013). More than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates. *Journal of Marketing*, *77*(1), 87–103. https://doi.org/10.1509/jm.11.0560

Lumsden, S.-A., Beldona, S., & Morrison, A. M. (2008). Customer Value in an All-Inclusive Travel Vacation Club: An Application of the RFM Framework. *Journal of Hospitality Marketing & Management*, *16*(3), 270–285. https://doi.org/10.1080/10507050801946858

Luo, R., Zhou, L., Song, Y., & Fan, T. (2022). Evaluating the impact of carbon tax policy on manufacturing and remanufacturing decisions in a closed-loop supply chain. *International Journal of Production Economics*, *245*, 108408. https://doi.org/10.1016/J.IJPE.2022.108408

Ma, H., & Guo, Y. (2010). Customer segmentation study of college students Based on the RFM. *Proceedings of the International Conference on E-Business and E-Government, ICEE 2010*, 3860–3863. https://doi.org/10.1109/ICEE.2010.968

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. *UNIVERSITY OF CALIFORNIA, Los ANGELES*.

Marinkovic, V., & Kalinic, Z. (2017). Antecedents of customer satisfaction in mobile commerce Exploring the moderating effect of customization. *Online Information Review*, *41*(2), 138–154. https://doi.org/10.1108/OIR-11-2015-0364

Mauri, A. G., & Minazzi, R. (2013). Web reviews influence on expectations and purchasing intentions of hotel potential customers. *International Journal of Hospitality Management*, *34*(1), 99–107. https://doi.org/10.1016/J.IJHM.2013.02.012

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. *Journal of Business Research*, *60*(6), 656–662. https://doi.org/10.1016/j.jbusres.2006.06.015

Memon, K. H., & Lee, D. (2017). Generalised fuzzy c-means clustering algorithm with local information. *IET Image Processing*, *11*(1), 1–12. https://doi.org/10.1049/iet-ipr.2016.0282

Miglautsch, J. R. (2000). Thoughts on RFM scoring. *Journal of Database Marketing & Customer Strategy Management*, *8*(1), 67–72. https://doi.org/10.1057/palgrave.jdm.3240019

Minnema, A., Bijmolt, T. H. A., Gensler, S., & Wiesel, T. (2016). To Keep or Not to Keep: Effects of Online Customer Reviews on Product Returns. *Journal of Retailing*, *92*(3), 253–267. https://doi.org/10.1016/J.JRETAI.2016.03.001

Morales, A. C., Amir, O., & Lee, L. (2017). Keeping it real in experimental research-understanding when, where, and how to enhance realism and measure consumer

behavior. *Journal of Consumer Research*, *44*(2), 465–476. https://doi.org/10.1093/jcr/ucx048

Morshedlou, H., & Meybodi, M. R. (2014). Decreasing Impact of SLA Violations:A Proactive Resource Allocation Approachfor Cloud Computing Environments. *IEEE Transactions on Cloud Computing*, *2*(2), 156–167. https://doi.org/10.1109/TCC.2014.2305151

Mosaddegh, A., Albadvi, A., Sepehri, M. M., & Teimourpour, B. (2021). Dynamics of customer segments: A predictor of customer lifetime value. *Expert Systems with Applications*, *172*. https://doi.org/10.1016/j.eswa.2021.114606

Nenonen, S., & Storbacka, K. (2016). *Driving shareholder value with customer asset management: Moving beyond customer lifetime value*. *52*, 140–150. https://doi.org/10.1016/j.indmarman.2015.05.019

O. A. Abbas. (2008). Comparisons between data clustering algorithms. *The International Arab Journal of Information Technology*, *5*, 320–325.

Oghuma, A. P., Libaque-Saenz, C. F., Wong, S. F., & Chang, Y. (2016). An expectation-confirmation model of continuance intention to use mobile instant messaging. *Telematics and Informatics*, *33*(1), 34–47. https://doi.org/10.1016/J.TELE.2015.05.006

Oliver, R. L. (1999). Whence Consumer Loyalty? *Journal of Marketing*, *63*, 33. https://doi.org/10.2307/1252099

Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends® in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Pappas, I. O., Kourouthanassis, P. E., Giannakos, M. N., & Lekakos, G. (2017). The interplay of online shopping motivations and experiential factors on personalized e-commerce: A complexity theory approach. *Telematics and Informatics*, *34*(5), 730–742. https://doi.org/10.1016/J.TELE.2016.08.021

Parasuraman, A., Zeithaml, V. A., & Malhotra, A. (2005). E-S-QUAL: A Multiple-Item Scale for Assessing Electronic Service Quality. *Journal of Service Research*, *7*(3), 213–233. https://doi.org/10.1177/1094670504271156

Peppers, D., Rogers Ph.D., M., & Dorf, B. (1999). Is Your Company Ready for One-to-One Marketing? *Harvard Business Review*, *77*(1), 151–160.

Puspitasari, I., Rusydi, F., Nuzulita, N., & Hsiao, C. S. (2023). Investigating the role of utilitarian and hedonic goals in characterizing customer loyalty in E-

marketplaces. *Heliyon*, *9*(8), e19193. https://doi.org/10.1016/J.HELIYON.2023.E19193

Quelhas Brito, P., Soares, C., Almeida, S., Monte, A., & Byvoet, M. (2015). Customer segmentation in a large database of an online customized fashion business. *Robotics and Computer-Integrated Manufacturing*, *36*(1), 93–100. https://doi.org/10.1016/j.rcim.2014.12.014

Rahim, A. G. (2016). Perceived Service Quality and Customer Loyalty: The Mediating Effect of Passenger Satisfaction in the Nigerian Airline Industry. *International Journal of Management and Economics*, *52*(1), 94–117. https://doi.org/10.1515/ijme-2016-0029

Rauyruen, P., Miller, K. E., & Groth, M. (2009). B2B services: linking service loyalty and brand equity. *Journal of Services Marketing*, *23*(3), 175–186. https://doi.org/10.1108/08876040910955189

Reinartz, W. J., & Kumar, V. (2000). On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing. *Journal of Marketing*, *64*(4), 17–35. https://doi.org/10.1509/jmkg.64.4.17.18077

Rita, P., Oliveira, T., & Farisa, A. (2019). The impact of e-service quality and customer satisfaction on customer behavior in online shopping. *Heliyon*, *5*(10), e02690. https://doi.org/10.1016/J.HELIYON.2019.E02690

Robert C. Blattberg, Gary Getz, & Jacquelyn S. Thomas. (2002). Customer Equity: Building and Managing Relationships as Valuable Assets. *Long Range Planning*, *35*(6), 657–661. https://doi.org/10.1016/S0024-6301(02)00155-3

Rust, R. T., Ambler, T., Carpenter, G. S., Kumar, V., & Srivastava, R. K. (2004). Measuring Marketing Productivity: Current Knowledge and Future Directions. *Journal of Marketing*, *68*(4), 76–89. https://doi.org/10.1509/jmkg.68.4.76.42721

S. Salvador, & P. Chan. (2004). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. *16th IEEE International Conference on Tools with Artificial Intelligence*, 576–584.

Saarijarvi, H., Karjaluoto, H., & Kuusela, H. (2013). Customer relationship management: The evolving role of customer data. *Marketing Intelligence & Planning*, *31*(6), 584–600. https://doi.org/10.1108/MIP-05-2012-0055

Schreieck, M., Wiesche, M., & Krcmar, H. (2019). *Developing an Industrial IoT Platform-Trade-off between Horizontal and Vertical Approaches*. https://www.researchgate.net/publication/331353742

Shafiee, M. M., & Bazargan, N. A. (2018). Behavioral Customer Loyalty in Online Shopping: The Role of E-Service Quality and E-Recovery. *Journal of Theoretical and Applied Electronic Commerce Research*, *13*(1), 26–38. https://doi.org/10.4067/S0718-18762018000100103

Sharma, R., & Jain, V. (2019). CSR, Trust, Brand Loyalty and Brand Equity: Empirical Evidences from Sportswear Industry in the NCR Region of India. *Metamorphosis: A Journal of Management Research*, *18*(1), 57–67. https://doi.org/10.1177/0972622519853158

Sheu, P.-L., & Chang, S.-C. (2022). Relationship of service quality dimensions, customer satisfaction and loyalty in e-commerce: a case study of the Shopee App. *Applied Economics*, *54*(40), 4597–4607. https://doi.org/10.1080/00036846.2021.1980198

Shim, B., Choi, K., & Suh, Y. (2012). CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns. *Expert Systems with Applications*, *39*. https://doi.org/10.1016/j.eswa.2012.01.080

Shreya Tripathi, Aditya Bhardwaj, & Poovammal E. (2018). Approaches to Clustering in Customer Segmentation. *International Journal of Engineering &Technology*, *7*(3), 7–802. https://doi.org/10.14419/ijet.v7i3.12.16505

Sohrabi, B., & Amir, K. (2007). Customer lifetime value (CLV) measurement based on RFM model. *Iranian Accounting & Auditing Review*, *14*.

Srivastava, R. K., Shervani, T. A., & Fahey, L. (1999). Marketing, Business Processes, and Shareholder Value: An Organizationally Embedded View of Marketing Activities and the Discipline of Marketing. *Journal of Marketing*, *63*, 168–179. https://doi.org/10.2307/1252110

Stahl, H. K., Matzler, K., & Hinterhuber, H. H. (n.d.). *Linking customer lifetime value with shareholder value*. https://doi.org/10.1016/S0019-8501(02)00188-8

Stone, B., & Jacobs, R. (1995). *Successful Direct Marketing Methods* (illustrated). McGraw-Hill, 2001.

Subbalakshmi, C., Rama Krishna, G., Rao, K. M., & Rao, V. (2015). A method to find optimum number of clusters based on fuzzy silhouette on dynamic data set.

Procedia Computer Science, *46*, 346–353. https://doi.org/10.1016/j.procs.2015.02.030

Swift, & Ronald S. (2000). *Accelerating Customer Relationship Using CRM and Relationship Technologies*. Upper Saddle River.

Tam, C., Santos, D., & Oliveira, T. (2020). Exploring the influential factors of continuance intention to use mobile Apps: Extending the expectation confirmation model. *Information Systems Frontiers*, *22*(1), 243–257. https://doi.org/10.1007/s10796-018-9864-5

Thomas, J. S. (2001). A methodology for linking customer acquisition to customer retention. *Journal of Marketing Research*, *2*(38), 262–268.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 417–424. https://doi.org/10.3115/1073083.1073153

Vagn, P., Ann, F., & Clarke, H. (2001). *Business to Business Market Segmentation*. *30*, 473–486. https://doi.org/https://doi.org/10.1016/S0019-8501(99)00103-0

van Lierop, D., & El-Geneidy, A. (2016). Enjoying loyalty: The relationship between service quality, customer satisfaction, and behavioral intentions in public transit. *Research in Transportation Economics*, *59*, 50–59. https://doi.org/10.1016/j.retrec.2016.04.001

Van Nguyen, T., Zhou, L., Chong, A. Y. L., Li, B., & Pu, X. (2020). Predicting customer demand for remanufactured products: A data-mining approach. *European Journal of Operational Research*, *281*(3), 543–558. https://doi.org/10.1016/J.EJOR.2019.08.015

Wang, L., Wang, Z., Wang, X., & Zhao, Y. (2021). Available online 27. *Electronic Commerce Research and Applications*, *49*, 1567–4223. https://doi.org/10.1016/j.elerap.2021.101080

Wang, Y., Kim, J., & Kim, J. (2021). The financial impact of online customer reviews in the restaurant industry: A moderating effect of brand equity. *International Journal of Hospitality Management*, *95*, 102895. https://doi.org/10.1016/J.IJHM.2021.102895

Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, *80*(6), 97–121. https://doi.org/10.1509/jm.15.0413

Wei, J.-T., Lin, S.-Y., & Wu, H.-H. (2010). A review of the application of RFM model. *African Journal of Business Management*, *4*(19), 4199–4206. http://www.academicjournals.org/AJBM

Wigand, R. T. (2003). *Electronic Commerce: Definition, Theory, and Context* (pp. 489–503). https://doi.org/10.1080/019722497129241

Wu, W.-Y., Ke, C.-C., & Nguyen, P.-T. (2018). Online Shopping Behavior in Electronic Commerce: An Integrative Model from Utilitarian and Hedonic Perspective. *International Journal of Entrepreneurship*, *22*(3). https://www.researchgate.net/publication/328569982

Xiong, C., Hua, Z., Lv, K., & Li, X. (2016). An Improved K-means Text Clustering Algorithm by Optimizing Initial Cluster Centers. *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*, 265–268. https://doi.org/10.1109/CCBD.2016.059

Yoseph, F., Ahamed Hassain Malim, N. H., Heikkilä, M., Brezulianu, A., Geman, O., & Paskhal Rostam, N. A. (2020). The impact of big data market segmentation using data mining and clustering techniques. *Journal of Intelligent and Fuzzy Systems*, *38*(5), 6159–6173. https://doi.org/10.3233/JIFS-179698

Zahrotun, L. (2017). Implementation of data mining technique for customer relationship management (CRM) on online shop tokodiapers.com with fuzzy c-means clustering. *2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 299–303. https://doi.org/10.1109/ICITISEE.2017.8285515

Zeithaml, V. A., Berry, L. L., & Parasuraman, A. (1996). The Behavioral Consequences of Service Quality. *Journal of Marketing*, *60*(2), 31. https://doi.org/10.2307/1251929

Zhai, M., Wang, X., & Zhao, X. (2024). The importance of online customer reviews characteristics on remanufactured product sales: Evidence from the mobile phone market on Amazon.com. *Journal of Retailing and Consumer Services*, *77*, 103677. https://doi.org/10.1016/J.JRETCONSER.2023.103677

Zhao, X. (Roy), Wang, L., Guo, X., & Law, R. (2015). The influence of online reviews to online hotel booking intentions. *International Journal of Contemporary Hospitality Management*, *27*(6), 1343–1364. https://doi.org/10.1108/IJCHM-12-2013-0542

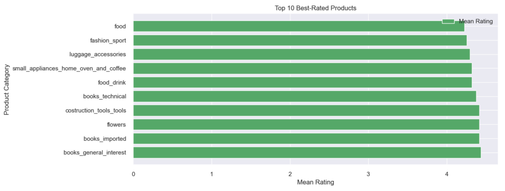
Zhou, R., Wang, X., Shi, Y., Zhang, R., Zhang, L., & Guo, H. (2019). Measuring e-service quality and its importance to customer satisfaction and loyalty: an empirical study in a telecom setting. *Electronic Commerce Research*, *19*(3), 477–499. https://doi.org/10.1007/s10660-018-9301-3

Παραρτήματα

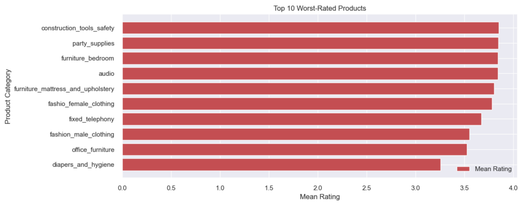
Appendix A

product_category_name_english	mean	count
books_general_interest	4.438.503	561
books_imported	4.419.355	62
flowers	4.419.355	31
costruction_tools_tools	4.415.842	101
books_technical	4.375.465	269
food_drink	4.324.138	290
small_appliances_home_oven_and_coffee	4.320.513	78
luggage_accessories	4.295.945	1159
fashion_sport	4.258.065	31
food	4.228.963	511

Appendix Figure 1 Οι καλύτερες κατηγορίες προϊόντων βάσει αξιολογήσεων



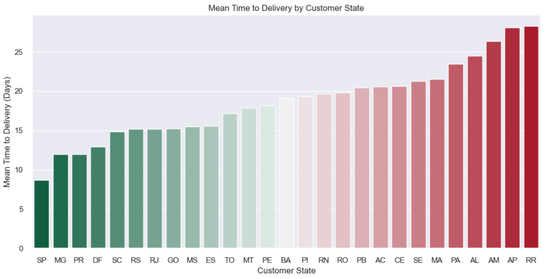
Appendix Figure 2 Οι καλύτερες κατηγορίες προϊόντων βάσει βαθμολογίας



Appendix Figure 3 Οι 10 καλύτερες κατηγορίες προϊόντων βάσει βαθμολογίας

City Name	lat	lng
Acre	-9,07000324	-68,66997929
Alagoas	-9,48000405	-35,83996769
Amapá	-0,03959837	-51,17998743
Amazonas	-3,28958087	-60,6199797
Bahia	-16,2800024	-39,0299797
Ceará	-2,89999225	-40,85002364
Distrito Federal	-15,7833402	-47,91605229
Espírito Santo	-20,8500077	-41,12998071
Goiás	-17,7300431	-49,10998458
Maranhão	-5,80999551	-46,14998438
Mato Grosso	-15,650015	-56,14002059
Mato Grosso do Sul	-22,5300085	-55,7299681
Minas Gerais	-18,7800049	-42,95002466
Pará	-1,19001911	-47,17999903
Paraíba	-7,01958576	-37,29000838
Paraná	-24,089965	-54,2699797
Pernambuco	-8,11001015	-35,02004358
Piaui	-4,82003009	-42,18001998
Rio de Janeiro	-22,9111	-43,2056
Rio Grande do Norte	-5,65000527	-37,80000309
Rio Grande do Sul	-30,8800415	-55,53000615
Rondônia	-11,6400272	-61,20999536
Roraima	1,81623151	-61,12767481
Santa Catarina	-27,2300317	-52,03001306
São Paulo	-23,55	-46,6333
Sergipe	-11,2696106	-37,45002446
Tocantins	-6,3195768	-47,41998438

Appendix Figure 4 Γεωγραφικό πλάτος και μήκος πόλεων Βραζιλίας



Appendix Figure 5 Μέσος χρόνος παράδοσης παραγγελίας ανά περιοχή πελάτη

Appendix Figure 6 Μέση βαθμολογία ηλεκτρονικού καταστήματος ανά έτος