ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

UNIVERSITY OF PIRAEUS

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ «ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ & ΑΝΑΛΥΤΙΚΗ»

Μέθοδοι Τμηματοποίησης Πελατών με Χρήση Αλγορίθμων Μηχανικής Μάθησης

Βικτωρία Κοντογεώργου

ΑΜ:1716

Επιβλέπων Καθηγητής: Μιχαήλ Φιλιππάκης

Φεβρουάριος 2020

Περίληψη

Η σύγχρονη εποχή είναι ιδιαίτερα ανταγωνιστική για τις επιχειρήσεις. Το περιβάλλον μέσα στο οποίο κινούνται είναι μεταβαλλόμενο και απαιτεί σωστή λήψη αποφάσεων κάθε στιγμή και δεδομένου των συνθηκών. Σήμερα, η διαδικασία της λήψης αποφάσεων είναι αποτέλεσμα πολύπλοκων διαδικασιών μέσα σε μία εταιρεία. Είναι σημαντικό όμως ότι η σύγχρονη επιχείρηση έχει στην διάθεσή της πληθώρα δεδομένων που με την κατάλληλη επεξεργασία μπορούν ν' αποκτήσουν αξία για την ίδια. Τα δεδομένα αυτά έχουν διαφορετικές πηγές, προέρχονται από διαφορετικά πληροφοριακά συστήματα και καλείται να συνδυαστούν ώστε να φέρουν κάποιο αποτέλεσμα. Αν μια εταιρεία καταφέρει να αντλήσει πληροφορίες μέσα από τα δεδομένα, αυτό είναι μεγάλο κέρδος διότι αυτόματα διαμορφώνει της στρατηγική της. Η παρούσα εργασία κινείται ακριβώς γύρω από αυτόν τον άξονα. Η Τμηματοποίηση Πελατών και οι μέθοδοι υλοποίησής της έχουν σκοπό να παρέχουν το κέρδος γνώσης στο τμήμα Μάρκετινγκ. Η ανίχνευση ομάδων πελατών μέσα από το σύνολο δεδομένων παρέχει την δυνατότητα στην επιχείρηση να διαμορφώσει και να εφαρμόσει πιο στοχευμένες καμπάνιες διαφήμισης και πιο προσωποποιημένες στρατηγικές για προσέλκυση, απόκτηση και διατήρηση πελατών. Οι μέθοδοι που χρησιμοποιούνται για την υλοποίηση της Τμηματοποίησης θα πρέπει να είναι όσο γίνεται πιο αντιπροσωπευτικές της πραγματικής κατάστασης και έχει σημασία η απόδοσή τους να ενισχύεται και να εμπλουτίζεται σε συνδυασμό με την πρόοδο της επιστήμης. Η παρούσα εργασία αποτελεί μία τέτοιου είδους μελέτη και αποδεικνύει ότι στο κομμάτι της τμηματοποίησης πελατών έχει αποτελεσματικότητα δίνοντας ταυτόχρονα το έδαφος σε περαιτέρω έρευνα και εξέλιξη.

Λέξεις κλειδιά: Τμηματοποίηση Πελατών, Διαχείριση Ομάδων Πελατών, Αγοραστική Συμπεριφορά, RFM Μοντέλο, Χρηματική Αξία Πελάτη, Αλγόριθμοι Μηχανικής Μάθησης, Κατηγοριοποίηση

.

Abstract

Viktoria Kontogeorgou

Customer Segmentation Methods with Machine Learning Algorithms

University of Piraeus

Department of Digital Systems

Msc. Big Data & Analytics

Thesis 2020

Today, increased competition between organizations has led them to seek a better understanding of their business. In this way they will be able to estimate and make right decisions according to dynamic changes. Better understanding of a business is a procedure where strong and analytical skills are essential. Innovative ways of storing and analysing data are required. The Data Science has a major development on this field. Purpose of the thesis is to apply Machine Learning Methods on Implementation of Customer Segmentation. Customer satisfaction is one of the most important organizational goals. Since all customers do not represent the same profitability to an organization, understanding and identifying the valuable customers is considered crucial. Thus, understanding customers' behavioural variables and categorizing customers based on these characteristics could provide better insight that will help business owners and industries to adopt appropriate marketing strategies such as up-selling and cross-selling. The current thesis presents a research on customer segmentation through Machine Learning Algorithms based on classic RFM Model. They are performed experiments on how a classic model behaves in a Machine Learning Environment. The results ensure and enhance the RFM model, providing the ability for future research.

Keywords: Customer Segmentation, Customer Segments Management, Customer Behaviour, RFM Model, Customer Lifetime Value, Machine Learning Algorithms, Classification.

Ευχαριστίες

Με την ολοκλήρωση της μεταπτυχιακής διπλωματικής μου εργασίας, νιώθω την ανάγκη να εκφράσω τις θερμές μου ευχαριστίες σε όλους όσους συνέβαλλαν στην εκπόνησή της.

Ευχαριστώ θερμά τον επιβλέπων καθηγητή μου, κύριο Μιχαήλ Φιλιππάκη, αρχικά για την εμπιστοσύνη που μου έδειξε, με την ανάθεση της διπλωματικής και στην πορεία για όλη την καθοδήγηση και επιμονή του κατά την διάρκεια της συνεργασίας μας.

Τέλος, θα ήθελα εκφράσω την ευγνωμοσύνη μου και να ευχαριστήσω την οικογένειά μου και ιδιαίτερα τον αδερφό μου Μιλτιάδη, για όλη τη στήριξη ηθική και οικονομική, την συμπαράσταση και την κατανόηση, καθ' όλη τη διάρκεια της διπλωματικής μου εργασίας αλλά και των μεταπτυχιακών μου σπουδών.

Βικτωρία Κοντογεώργου

Αθήνα, Φεβρουάριος 2020

Περιεχόμενα

Περίληψη.....iii
Abstract.....v
Ευχαριστίες.....vii
Λίστα Εικόνων – Πινάκων – Αποτελέσματα Κώδικα & Weka.....xi
Κεφάλαιο 1 Εισαγωγή.....1
1.1Αντικείμενο της Διπλωματικής.....1
1.2Σκοπός της Διπλωματικής.....2
1.3Δομή της Διπλωματικής.....2
Κεφάλαιο 2 Τμηματοποίηση Πελατών – Θεωρητικό Υπόβαθρο.....4
2.1Εισαγωγή.....5
2.1.1Marketing και Μεγάλα Δεδομένα - Data Driven Marketing.....5
2.2Τμηματοποίηση Πελατών - Customer Segmentation.....6
2.2.1Ορισμός – Χρησιμότητα.....6
2.2.2Βασικοί Τύποι Τμηματοποίησης Πελατών – Types of Customer Segmentation.....7
2.3RFM Μοντέλο Κατηγοριοποίησης.....9
2.3.1Ορισμός.....9
2.3.2Θεωρία.....9
2.3.3Ανάλυση Μοντέλου.....10
2.3.4Χρησιμότητα.....12
2.3.5Εφαρμογή - Εκτέλεση.....13
2.3.6Διαχείριση Αποτελεσμάτων.....14
2.3.7Αξιοποίηση – Πλεονεκτήματα Χρήσης.....17
2.4Αξία Ενός Πελάτη – Customer Lifetime Value.....20
2.5Κατηγοριοποίηση στην Μηχανική Μάθηση.....21
2.5.1Εισαγωγή.....21
2.5.2Επαγωγικοί Αλγόριθμοι και Μοντέλα.....22
2.5.3Στάδια Κατηγοριοποίησης.....23
2.5.4Κριτήρια Αξιολόγησης Μεθόδων Κατηγοριοποίησης.....24

Κεφάλαιο 3 Τμηματοποίηση Πελατών – Πρακτικό Μέρος 26
3.1Εισαγωγή 27
3.1.1Εργαλεία Υλοποίησης Εργασίας 27
3.1.2Σύντομη Αναφορά στην Χρησιμότητα της Προεπεξεργασίας Δεδομένων 30
3.2Προεπεξεργασία Δεδομένων – Data Preprocessing 32
3.2.1Περιγραφή -Συνόλου Δεδομένων 32
3.2.2Προεπεξεργασία Δεδομένων 33
3.3Εφαρμογή RFM Μοντέλου για Τμηματοποίηση Πελατών 37
3.4Τμηματοποίηση Πελατών με βάση την τιμή Customer Lifetime Value 42
3.5Πρόβλεψη Κατηγορίας και Ενίσχυση RFM Μοντέλου με χρήση WEKA 44
Κεφάλαιο 4 Συμπεράσματα - Προτάσεις 47
4.1Συμπεράσματα 48
4.1.1Τμηματοποίηση Πελατών βάσει της Αγοραστικής Συμπεριφοράς με εφαρμογή του RFM Μοντέλου και της Μηχανικής Μάθησης 48
4.1.2Τμηματοποίηση Πελατών βάσει της Customer Lifetime Value και πρόβλεψή της με το RFM Μοντέλο μέσω της Μηχανικής Μάθησης 49
4.2Συνολική Αξιολόγηση - Προτάσεις 50
Κεφάλαιο 5 Βιβλιογραφία 51

Λίστα Εικόνων – Πινάκων – Αποτελεσμάτων Κώδικα & Weka

Εικόνες

Εικόνα 2.1: Τύποι Τμηματοποίησης Πελατών βάσει της Κατηγορίας Δεδομένων.....7
Εικόνα 2.2: Γραφική Αναπαράσταση "Αρχής του Pareto" στην Τμηματοποίηση Πελατών...10
Εικόνα 2.3: Βασικές Μετρικές που ορίζουν το R,F,M.....11
Εικόνα 2.4: Γραφικές Περιοχές των Ομάδων Πελατών.....15

Πίνακες

Πίνακας 3.1: Χαρακτηριστικά – Γνωρίσματα του Συνόλου Δεδομένων 32
Πίνακας 3.2: Αριθμητική Εκτίμηση του Συνόλου Δεδομένων μετά την Προεπεξεργασία 36
Πίνακας 3.3: Ποσοστό Επιτυχής CLV_Κατηγοριοποίησης ανά Αλγόριθμο με χρήση των R, F, M σε διαφορετική έκφραση 45
Πίνακας 3.4: Ποσοστό Επιτυχής CLV_Κατηγοριοποίησης ανά συνδυασμό των actual R, F, M 46

Αποτελέσματα Κώδικα & Weka

Code Output 1: Παρουσίαση Πληροφοριών του Συνόλου Δεδομένων – Εντολή - .info().....33
Code Output 2: Περιγραφή του Συνόλου Δεδομένων - Εντολή .describe().....34
Code Output 3: Εύρεση των 10 Περιοχών με τους περισσότερους Πελάτες. 35
Code Output 4: Ετικέτα Ομάδας ανά Πελάτη 39
Code Output 5: Γράφημα Κατάταξης Ομάδων Πελατών Συγκεκρ.Περιοχής με Ποσοστά 39
Code Output 6: Κατανομή των Πελατών ανά Recency, Frequency. 40
Code Output 7: Οι μέσες τιμές των μετρικών R, F, M ανά Ομάδα Πελατών 40
Code Output 8: Γράφημα Κατάταξης Ομάδων Πελατών Όλων των Περιοχών με Ποσοστά .41
Code Output 9: Οι μέσες τιμές των μετρικών R, F, M ανά Ομάδα Πελατών 41
Code Output 6: Βασικές τιμές για τον τελικό υπολογισμό της τιμής Customer Value.....42
Code Output 7: Κατηγοριοποίηση Πελατών βάσει Customer Lifetime Value 43

Weka Output 1: Confusion Matrix J48 - Customer Segment Classification με βάση τις τιμές R, F, M 44
Weka Output 2: Summary J48 - Customer Segment Classification με βάση τα Δημογραφικά Χαρακτηριστικά
Weka Output 3: Summary J48 - CVL Segment Classification με βάση τα Δημογραφικά Χαρακτηριστικά 45
Weka Output 4: Classification CVL Κατηγορίας βάσει της Ομάδα Πελατών 46

Κεφάλαιο 1 Εισαγωγή

1.1 Αντικείμενο της Διπλωματικής

Η σημερινή εποχή είναι πολύ απαιτητική για όλες τις επιχειρήσεις και ιδιαίτερα του λιανικού εμπορίου. Οι επιχειρήσεις αυτού του κλάδου δραστηριοποιούνται συνήθως σε παγκόσμια κλίμακα και χρειάζεται να προβάλλουν την ποικιλία των προϊόντων τους στους καταναλωτές με ελκυστικό τρόπο ώστε να υπάρχει το επιθυμητό αποτέλεσμα. Το γεγονός όμως ότι η αγορά είναι μεγάλη και προσφέρει πολλές επιλογές του ίδιου προϊόντος καθιστά δύσκολο αυτόν τον σκοπό. Πιο συγκεκριμένα, οι πελάτες αλλάζουν εύκολα επιλογές στις μάρκες των προϊόντων που αγοράζουν, καθιστώντας δύσκολη την δημιουργία μίας σταθερής βάσης αξιόπιστων πελατών. Ο ανταγωνισμός είναι έντονος και χρειάζεται πλάνο για το πώς μία εταιρεία μπορεί να παραμείνει και να έχει μία σταθερή και ανοδική πορεία.

Η ραγδαία εξέλιξη της τεχνολογίας παίζει βασικό ρόλο στην εύρεση τρόπων αντιμετώπισης του ανταγωνισμού. Η δύναμη της πληροφορίας που λαμβάνεται για και από τον πελάτη συμβάλλει σημαντικά ώστε να βρεθούν στοχευμένοι τρόποι προσέλκυσης και διατήρησης των πελατών. Συγκεκριμένα, μία εταιρεία για να μπορέσει να τραβήξει και να διατηρήσει το ενδιαφέρον των καταναλωτών της βασίζεται στην προβολή, την διαφήμιση και γενικότερα στο Μάρκετινγκ. Για την επιτυχία της διαφήμισης όμως, δεν αρκεί μόνο η εμπειρία και οι προσωπικές ικανότητες του διαφημιστή. Η άνοδος της Επιστήμης των Δεδομένων ξεπερνά τα όρια της Στατιστικής καθώς παρέχει την δυνατότητα ν΄ ανιχνευθούν ομοιότητες και μοτίβα στα δεδομένα, ιδιαίτερα κερδοφόρα για τον τομέα του Μάρκετινγκ.

Η ανίχνευση ομάδων πελατών αποτελεί από τις σημαντικότερες πληροφορίες που προκύπτουν από την σωστή μελέτη των δεδομένων . Η γνώση είναι καθοριστική για τον σχεδιασμό των καμπάνιων διαφήμισης και προσέγγισης των πελατών. Η διαχείριση της κάθε ομάδας με τον ανάλογο τρόπο δημιουργεί την αίσθηση της προσωποποιημένης προσέγγισης και προσέλκυσης προς τον καταναλωτή. Επομένως, η διερεύνηση της αγοραστικής συμπεριφοράς αποτελεί το πρώτο βήμα πάνω στο οποίο πρέπει να στηριχτεί το Μάρκετινγ.

Στην παρούσα εργασία η Τμηματοποίηση πελατών θα εστιαστεί στην αγοραστική συμπεριφορά αλλά και ,σε ξεχωριστό κομμάτι, στο χρηματικό αντίκτυπο που έχεις ένας πελάτης για μία εταιρεία σε βάθος χρόνου. Η εφαρμογή μεθόδων και η επαλήθευση και ενίσχυσή τους μέσω της Μηχανικής Μάθησης είναι το κύριο αντικείμενο μελέτης.

1.2 Σκοπός της Διπλωματικής

Η παρούσα μεταπτυχιακή εργασία έχει σκοπό την μελέτη μεθόδων Τμηματοποίησης Πελατών σε συνδυασμό με τις τεχνολογίες Μηχανικής Μάθησης. Συγκεκριμένα θα παρουσιαστεί πώς κλασσικά μοντέλα που αξιοποιούνταν μέχρι σήμερα, συμπεριφέρονται όταν εφαρμόζονται στο πεδίο της Μηχανικής Μάθησης.

Αρχικά τονίζεται η χρησιμότητα της Τμηματοποίησης πελατών και αναλύονται τα οφέλη της ενώ στην συνέχεια γίνεται η περιγραφή των εννοιών και των μεθόδων που θα εφαρμοστούν. Η εργασία εστιάζει ιδιαίτερα στο κλασσικό μοντέλο RFM και την έννοια Customer Lifetime Value ενώ στο πεδίο της Μηχανικής Μάθησης τα πειράματα αφορούν τον επιβλεπόμενο τρόπο μάθησης, την Κατηγοριοποίηση.

Νοητά η μελέτη της παρούσας εργασίας μπορεί να χωριστεί σε δύο φάσεις. Στην πρώτη φάση εφαρμόζεται ένας κλασσικός τρόπος τμηματοποίησης πελατών όπου σε κάθε στιγμιότυπο του συνόλου δεδομένων αντιστοιχείται μία κατηγορία. Στην δεύτερη φάση επιχειρείται η πρόβλεψη της κατηγορίας μέσω της Μηχανικής Μάθησης με χρήση του ελεύθερου λογισμικού Weka.

Η επαλήθευση των προβλέψεων των κλασσικών μεθόδων μέσω της Μηχανικής Μάθησης παρουσιάζεται να έχει επιτυχία, ενισχύοντας με αυτόν τον τρόπο το αρχικό μοντέλο. Σαν γενικό συμπέρασμα προκύπτει ότι η Μηχανική Μάθηση ανιχνεύοντας συσχετίσεις και ομοιότητες σε συνδυασμό με την πληροφορία των κλασσικών μοντέλων, τ' αποτελέσματα εμπλουτίζονται και δίνονται νέες ερμηνείες στο βαθμό συνεισφοράς του κάθε γνωρίσματος ως προς την πρόβλεψη.

1.3 Δομή της Διπλωματικής

Το **Κεφάλαιο 2** εισάγει την έννοια της Τμηματοποίησης των Πελατών και τους τρόπους αξιοποίησης αυτής από το τμήμα Μάρκετινγκ. Γίνεται αναλυτική περιγραφή της κάθε ομάδας πελατών και των τρόπων καλύτερης διαχείρισης της αγοραστικής σχέσης με τους καταναλωτές. Στην συνέχεια παρουσιάζεται το θεωρητικό υπόβαθρο του RFM Μοντέλου και της τιμής Customer Lifetime Value ενώ αναλύονται τα οφέλη που προκύπτουν από την εφαρμογή τους και η συμβολή τους στην ανέλιξη της επιχείρησης. Τέλος, παρουσιάζεται λεπτομερώς η μέθοδος της Μηχανικής Μάθησης, Κατηγοριοποίηση, τα στάδια εφαρμογής της καθώς και τα κριτήρια αξιοποίησης των αλγόριθμων μηχανικής μάθησης για κατηγοριοποίηση.

Το **Κεφάλαιο 3** αποτελεί την πρακτική εφαρμογή της θεωρίας που προηγήθηκε. Συγκεκριμένα παρουσιάζεται η διαδικασία απόδοσης ετικέτας κατηγορίας σε κάθε πελάτη με βάση το RFM Μοντέλο και ως ξεχωριστή ομαδοποίηση με βάση την τιμή που υποδηλώνει την Χρηματική Αξία του Πελάτη. Εν μέρει γίνεται αναφορά σε εντολές του κώδικα Python για καλύτερη κατανόηση αυτού του σταδίου ενώ στο δεύτερο στάδιο εφαρμόζονται αλγόριθμοι

Μηχανικής Μάθησης με σκοπό την πρόβλεψη της κατηγορίας αλλά και της ανίχνευσης των γνωρισμάτων που συμβάλλουν σε αυτήν. Η εκτέλεση αυτής της διαδικασίας γίνεται στο λογισμικό Weka και η παρουσίαση πλαισιώνεται από τις εξαγωγές που προκύπτουν και από συγκριτικούς πίνακες.

Το **Κεφάλαιο 4** αποτελεί έκθεση των συμπερασμάτων από την πρακτική εφαρμογή της Τμηματοποίησης Πελατών. Πιο αναλυτικά, γίνεται αξιολόγηση της ακρίβειας κατηγοριοποίησης με βάση το RFM Μοντέλο και της κατηγοριοποίησης με βάση την τιμή Customer Lifetime Value. Παρουσιάζεται η συμβολή των τιμών R, F, M και στις δυο κατηγοριοποιήσεις και τέλος αξιολογείται συνολικά η μελέτη της παρούσας εργασίας ενώ ακολουθούν κάποιες προστάσεις για μελλοντική έρευνα.

Κεφάλαιο 2 Τμηματοποίηση Πελατών – Θεωρητικό Υπόβαθρο

"Market segmentation 's a natural result of the vast differences among people."

– Donald Norman

2.1Εισαγωγή 5
2.2Τμηματοποίηση Πελατών - Customer Segmentation 6
2.3RFM Μοντέλο Κατηγοριοποίησης 8
2.4Αξία Ενός Πελάτη – Customer Lifetime Value 19
2.5Κατηγοριοποίηση στην Μηχανική Μάθηση 20

2.1 Εισαγωγή

Οι επιχειρήσεις καθορίζουν την στρατηγική πορεία και εξέλιξή τους βάσει του Marketing και των τεχνικών που αναπτύσσει ο συγκεκριμένος τομέας. Πιο ειδικά, η επιστήμη του Marketing και της Επικοινωνίας αναφέρεται στις επιχειρηματικές δραστηριότητες με σκοπό την προαγωγή της αγοράς ή της πώλησης ενός προϊόντος ή υπηρεσίας. Το κύρια εργαλεία είναι η έρευνα αγοράς και η διαφήμιση ενώ ο τελικός στόχος η ικανοποίηση και η προσέλκυση των πελατών.

Η ραγδαία εξέλιξη της Επιστήμης των Δεδομένων και των τεχνολογιών που την απαρτίζουν έχει συμβάλλει σημαντικά στην ανάπτυξη και αποτελεσματικότητα του Ψηφιακού Μάρκετινγκ οδηγώντας έτσι στην καλύτερη λήψη αποφάσεων των εταιρειών. Αυτό με την σειρά του έχει δημιουργήσει μία νέα έννοια, γνωστή με τον ξένο όρο Data Driven Marketing.

2.1.1 Marketing και Μεγάλα Δεδομένα - Data Driven Marketing

Τα Μεγάλα Δεδομένα ή αλλιώς Big Data είναι ο μεγάλος και ολοένα αυξανόμενος όγκος δεδομένων συμπεριλαμβάνοντας τις δυνατότητές που προκύπτουν από την αποθήκευση και ανάλυσή τους. Η συμμετοχή όλων και περισσότερων εταιρειών στο ηλεκτρονικό εμπόριο έχει ενισχύσει την παραγωγή δεδομένων και έχει καταστήσει την ανάγκη εύρεσης νέων μεθόδων επεξεργασίας τους. Τα χαρακτηριστικά που προσδιορίζουν τα Μεγάλα Δεδομένα είναι τα εξής:

- Όγκος
- Ταχύτητα
- Ποικιλία
- Μεταβλητότητα
- Ακρίβεια
- Αξία

Όπως είναι λογικό, η συλλογή των μεγάλων δεδομένων δεν μπορεί να μείνει ανεκμετάλλευτη οδηγώντας στην παρακάτω κατηγοριοποίηση των αξιοποιήσιμων ,από το τμήμα Μάρκετινγκ, δεδομένων:

1. **Πελατειακού τύπου - Customer**: Η πιο δημοφιλής και χρήσιμη κατηγορία μεγάλων δεδομένων περιλαμβάνει στοιχεία πελατών όπως ηλικία, φύλο, ιστορικό αγορών και αναζήτησης. Πηγές συγκέντρωσης είναι τα κοινωνικά μέσα δικτύωσης, γκάλοπ πελατών, ιστοσελίδες, σημεία πώλησης, καμπάνιες μάρκετινγκ κ.α.
2. **Λειτουργικού τύπου - Operational** : Τα δεδομένα αυτά χρησιμοποιούνται σε μετρικές απόδοσης την επιχειρηματικής διαδικασίας. Λαμβάνονται από τμήματα όπως των προμηθειών και του ανθρώπινου δυναμικού με σκοπό την βελτίωση της απόδοσης και την μείωση του κόστους.

3. **Οικονομικού τύπου - Financial**: Η κατηγορία αυτή βοηθάει στην εκτίμηση της επιχειρηματικής οικονομικής δραστηριότητας. Εδώ περιλαμβάνονται δεδομένα που αφορούν πωλήσεις, έσοδα, κέρδη και γενικότερα την οικονομική κατάσταση της εταιρείας.

Η συλλογή και επεξεργασία της παραπάνω 1ης κατηγορίας μεγάλων δεδομένων που αφορούν τους πελάτες/καταναλωτές αποτελεί αναπόσπαστο κομμάτι των Μάρκετινγκ ερευνών και αναλύσεων. Η τμηματοποίηση πελατών (Customer Segmentation) αποτελεί μία τέτοια μέθοδος ανάλυσης που στόχο έχει την εξαγωγή χρήσιμων πληροφοριών για την επιχειρηματική δράση μιας εταιρείας.

2.2 Τμηματοποίηση Πελατών - Customer Segmentation

2.2.1 Ορισμός – Χρησιμότητα

Η τμηματοποίηση πελατών αποτελεί ένα πολύ βασικό εργαλείο για τις επιχειρήσεις. Ορίζεται ως ο διαχωρισμός των πελατών σε κατηγορίες βάσει των κοινών χαρακτηριστικών τους και με σκοπό την καλύτερη διαχείρισή τους μέσω πιο στοχευμένου Μάρκετινγκ.

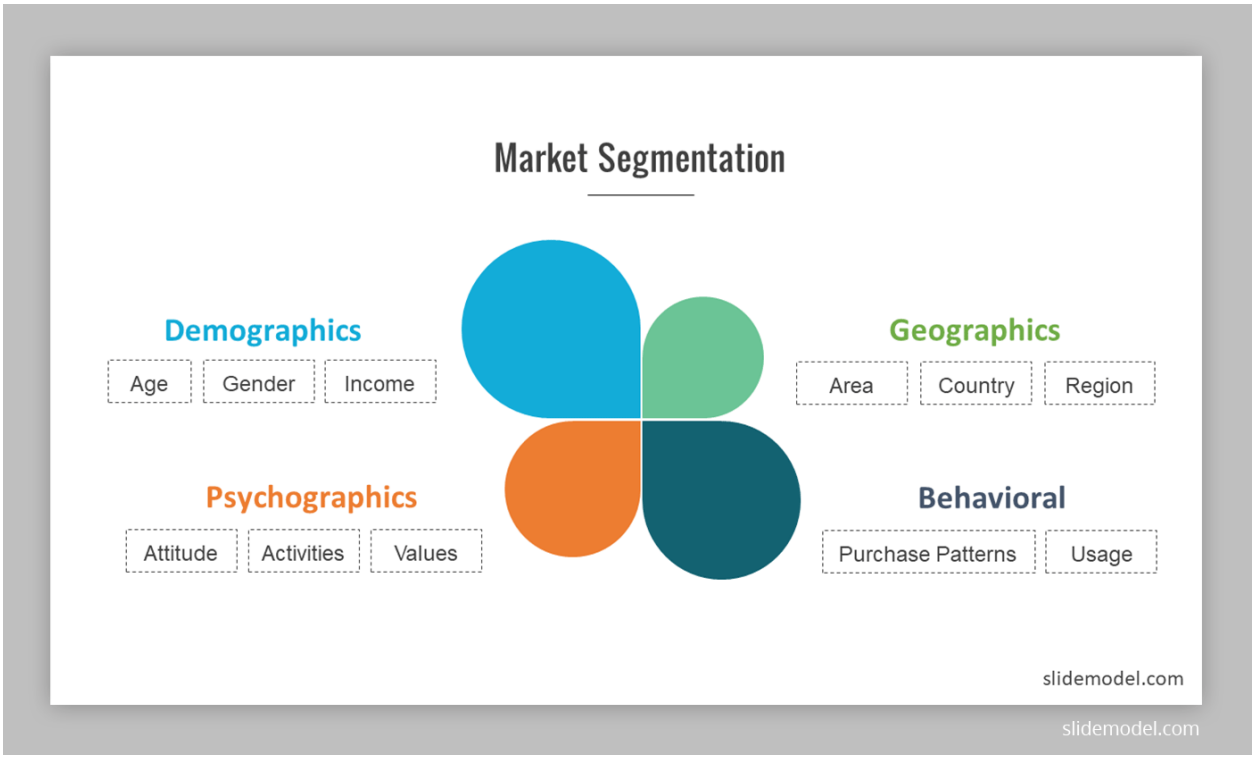
Συγκεκριμένα, οι πελάτες μιας επιχείρησης δεν έχουν όλοι την ίδια καταναλωτική δράση αλλά αποδεικνύεται πώς ένα μικρό ποσοστό αυτών συμβάλλει στην πλειοψηφία των κερδών. Σ’ αυτήν την περίπτωση, η επιχείρηση ενδιαφέρεται να κρατήσει ευχαριστημένη και ικανοποιημένη την συγκεκριμένη κατηγορία πελατών και να χτίσει μία αξιόπιστη σχέση μεταξύ τους. Σε άλλη περίπτωση, νέα προϊόντα και υπηρεσίες μπορούν να δημιουργηθούν για την προσέλκυση άλλων κατηγοριών πελατών με πιο ενδοιαστική αγοραστική συμπεριφορά.

Η τμηματοποίηση των καταναλωτών είναι ενδεικτική των προτιμήσεων καθώς και των αναγκών που ωθούν στην αγορά ενός προϊόντος, στοιχεία που συνεισφέρουν στις διαφημιστικές και προωθητικές καμπάνιες προς τους συγκεκριμένους πελάτες. Πιο ειδικά, η πρόβλεψη μελλοντικής αγοράς παρόμοιων προϊόντων βελτιωμένων εκδόσεων (up-selling) ή συμπληρωματικών προϊόντων προς παλαιότερες αγορές (cross-selling) μπορεί να εντοπιστεί πολύ πιο εύκολα και καθολικά και να ενισχυθεί αντίστοιχα. Ο εντοπισμός των αγοραστικών κινήτρων των καταναλωτών και πιο ειδικά αν το κίνητρο είναι άμεσα συνδεδεμένο με την αγοραστική αξία ή όχι έχει σαν αποτέλεσμα την καλύτερη διαμόρφωση της τιμολογιακής πολιτικής της εταιρείας.

Γενικότερα, η κατηγοριοποίηση πελατών είναι καίρια και αναπόσπαστη διαδικασία για την ανάπτυξη του προϊόντος μιας επιχείρησης (product development) καθώς βοηθάει να χτιστεί μία πιο ‘’προσωπική’’ σχέση με τον πελάτη έχοντας ως τελικό αποδέκτη το ίδιο το προϊόν ή υπηρεσία.

2.2.2 *Βασικοί Τύποι Τμηματοποίησης Πελατών – Types of Customer Segmentation*

Εικόνα 2.1: Τύποι Τμηματοποίησης Πελατών βάσει της Κατηγορίας Δεδομένων



Πηγή: https://slidemodel.com/market-segmentation-the-right-way/

Η τμηματοποίηση των πελατών γίνεται σε 4 βασικές κατηγορίες ή συνδυασμό αυτών με βάση τα **Δημογραφικά**, **Γεωγραφικά**, **Ψυχογραφικά** και **Συμπεριφορικά** χαρακτηριστικά τους όπως φαίνεται στην εικόνα. Παρακάτω θ΄ αναλύσουμε εκτενώς οι παραπάνω κατηγορίες.

Δημογραφικά χαρακτηριστικά ορίζονται τα πληθυσμιακά χαρακτηριστικά και απαντούν στο **«Ποιος».** Αποτελούν αντικειμενικά στοιχεία διαχωρισμού σε κατηγορίες**:**

- Φύλο
- Ηλικία
- Εισόδημα
- Επάγγελμα
- Μορφωτικό Επίπεδο
- Οικογενειακή Κατάσταση
- Αριθμός εξαρτώμενων τέκνων

Τα **Γεωγραφικά Χαρακτηριστικά** απαντούν στο **«Πού»** υποδεικνύοντας την γεωγραφική διασπορά των πελατών

- Πόλη ή Επαρχία
- Εντοπιότητα (Εγχώριος ή Διεθνής) Ονόματα Πόλεων για αξιολόγηση της Δυναμικής της Αγοράς
- Περιοχές και Διαμερίσματα
- Ταχυδρομικοί Κωδικοί

Η τμηματοποίηση με βάση τα **Ψυχογραφικά Χαρακτηριστικά** συχνά αναφέρεται ως lifestyle καθώς είναι ενδεικτική του τρόπου ζωής των καταναλωτών. Πολλές φορές υποδεικνύουν αν το αγοραστικό κίνητρο εξαρτάται από την τιμή ενός προϊόντος ή όχι:

- Χόμπι και Ενδιαφέροντα
- Δραστηριότητες
- Αξίες
- Χαρακτηριστικά Προσωπικότητας
- Καταναλωτικές συνήθειες

Τα **Συμπεριφορικά Χαρακτηριστικά** έχουν πολλά κοινά σημεία με τα ψυχογραφικά αλλά απαιτούν εις βάθος ανάλυση δεδομένων καθώς εστιάζουν κυρίως στην καταναλωτική συμπεριφορά:

- Μοτίβα καταναλωτικής συμπεριφοράς
- Αγοραστικός τρόπος σκέψης
- Στάση απέναντι στο brand και τρόπος χρήσης

Οι παραπάνω κατηγορίες είναι πολύ βασικές και σημαντικές αλλά για πιο αποτελεσματική τμηματοποίηση, ο συνδυασμός τους είναι απαραίτητος. Μ' αυτόν τον τρόπο το Μάρκετινγκ είναι πολύ πιο αποτελεσματικό καθώς λαμβάνει συνδυασμούς χαρακτηριστικών και ενώ φαινομενικά είναι εστιασμένο στον εκάστοτε πελάτη (customer tailor made), ταυτόχρονα είναι καθολικό καθώς απευθύνεται σε μεγαλύτερη μερίδα καταναλωτών.

Στην παρούσα εργασία θα εστιάσουμε στον 4^{ο} βασικό τύπο κατηγοριοποίησης με βάση τα Συμπεριφορικά χαρακτηριστικά των πελατών. Συγκεκριμένα θ' αξιολογήσουμε την στάση του καταναλωτή απέναντι στην εκάστοτε επιχείρηση και τον τρόπο αγοραστικής δράσης – χρήσης (customer's attitude & usage). Η υλοποίηση θα γίνει με δύο προσεγγίσεις, αρχικά με χρήση του RFM μοντέλου και στην συνέχεια με χρήση της τιμής Customer Lifetime Value.

2.3 RFM Μοντέλο Κατηγοριοποίησης

2.3.1 Ορισμός

Το RFM μοντέλο εφαρμόστηκε το 1994 από τον οικονομολόγο και καθηγητή **Arthur Middleton Hughes** ως μέτρο εκτίμησης της ικανοποίησης των πελατών για τις αγορές που είχαν πραγματοποιήσει (customer value). Στα χρόνια που ακολούθησαν η χρήση του

από τις επιχειρήσεις λιανικού εμπορίου συνέβαλαν στο να καθιερωθεί και να είναι ιδιαίτερα δημοφιλές.

Πιο ειδικά η RFM ανάλυση αποτελεί ένα ισχυρό εργαλείο κατηγοριοποίησης των πελατών με βάση την αγοραστική τους συμπεριφορά. Αυτό για μία επιχείρηση συνεπάγεται καλύτερο σχεδιασμό των προωθητικών ενεργειών της. Οι προωθητικές ενέργειες είναι πλέον στοχευμένες ν΄ ανταποκρίνονται στο προφίλ του εκάστοτε πελάτη αυξάνοντας έτσι την απόδοσή τους.

2.3.2 Θεωρία

Η RFM κατηγοριοποίηση και γενικότερα η ιδέα της κατηγοριοποίησης των πελατών έχει την βάση της ανάπτυξής της στην «Αρχή του Pareto» ή διαφορετικά τον «Κανόνα 80/20».

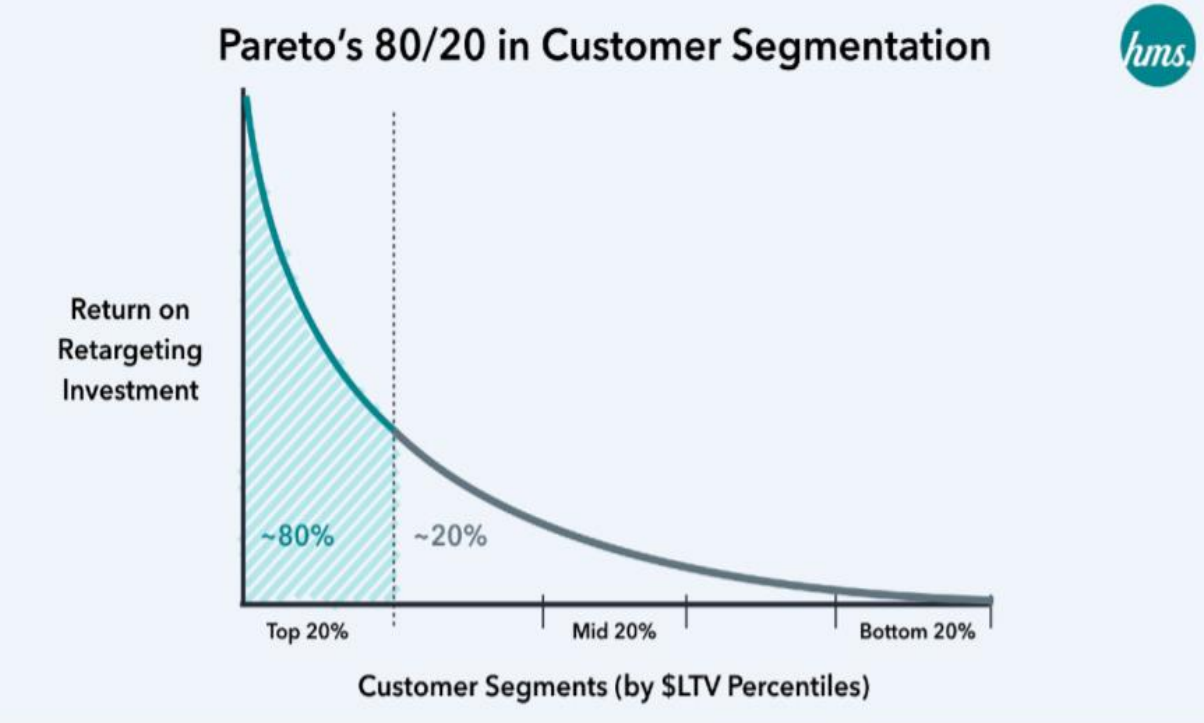
Ο **Vilfredo Federico Damaso Pareto (1848-1923)** ήταν Ιταλός μηχανικός, κοινωνιολόγος και οικονομολόγος. Ως ακαδημαϊκός καθηγητής απέκτησε φήμη χάρη στις καινοτόμες αντιλήψεις του στα οικονομικά και στην οικονομετρία, ενώ άφησε πίσω του και σημαντικό κοινωνιολογικό έργο[1].

Ο Pareto ήταν ο πρώτος που το 1897 παρατήρησε ότι το 80% των εσόδων στην Ιταλία προέρχεται από το 20% του πληθυσμού όπως επίσης και ότι το 20% του πληθυσμού κατέχει το 80% της γης. Ο Pareto συνεχίζοντας τις έρευνες παρατήρησε ότι η κατανομή ίσχυε όχι μόνο στην οικονομία και την κατανομή της γης αλλά και σε άλλους τομείς της ανθρώπινης δραστηριότητας. Το 1930 στις Ηνωμένες Πολιτείες, ο **Δρ. Joseph M. Juran** ήταν ο πρώτος που διατύπωσε μια καθολική αρχή την οποία απεκάλεσε κανόνα «των σημαντικών ολίγων και των ασήμαντων πολλών» (law of the vital few and the trivial many) ωστόσο η ονομασία «Αρχή του Pareto» επικράτησε.

Πιο κοντά στο σήμερα, πολλές εταιρείες παρατήρησαν ανάλογα μοτίβα. Η Microsoft διαπίστωσε πως το 80% του κώδικα μπορεί να γραφεί στο 20% του χρόνου καθώς και ότι το 20% επίλυσης των bugs ενός κώδικα λύνει το 80% των αστοχιών του. Αντίστοιχη αναλογία υπήρχε και στον αθλητικό τομέα με την παρατήρηση ότι χάρη στο περίπου 15% των επαγγελματιών παικτών του baseball επιτυγχάνεται το 85% των νικών την ομάδας.

Στον τομέα του εμπορίου και των επιχειρήσεων τώρα, μεταφράζοντας τον «Κανόνα 80/20» πρακτικά, αυτό σημαίνει ότι το 20% των πελατών αποφέρουν το 80% των συνολικών πωλήσεων καθώς και ότι το 80% των συνολικών κερδών προέρχονται από το 20% των πελατών, χωρίς τα παραπάνω 20% να συμπίπτουν. Εστιάζοντας λοιπόν στο 20% των καλύτερων πελατών και διαχειρίζοντας ανάλογα αυτήν την ιδιότητα, αυξάνεται η επιτυχία του σκοπού με ταυτόχρονη επιδίωξη της μικρότερης δυνατής δαπάνης.

Εικόνα 2.2: Γραφική Αναπαράσταση της "Αρχής του Pareto" στην Τμηματοποίηση Πελατών



Πηγή: https://how-many-steps-inc.webflow.io/rfm-segmentation-overview

Το RFM μοντέλο αξιοποιεί αυτήν την αρχή, ανιχνεύοντας μεταξύ των άλλων την ομάδα εκείνου του 20% των πελατών που αγοράζουν συχνά και ακριβά. Επιδιώκοντας μία ορθή προσέγγιση των «καλύτερων» πελατών, επιτυγχάνεται το 80% των επιθυμητών αποτελεσμάτων της καμπάνιας. Στην εφαρμογή του μοντέλου ένα 20 % θα συγκεντρώνει το υψηλότερο RFM score όπως θ' αναλυθεί παρακάτω έναντι του υπόλοιπου 80%. Αυτό δεν σημαίνει ότι δεν θα δοθεί σημασία και ανάλογη προσέγγιση στο 80% αλλά πάντα υπάρχουν συντελεστές που καθορίζουν τον βαθμό της ενασχόλησης μιας επιχείρησης με μία ομάδα και αυτό είναι μία γενική αρχή που ισχύει και ακολουθείται.

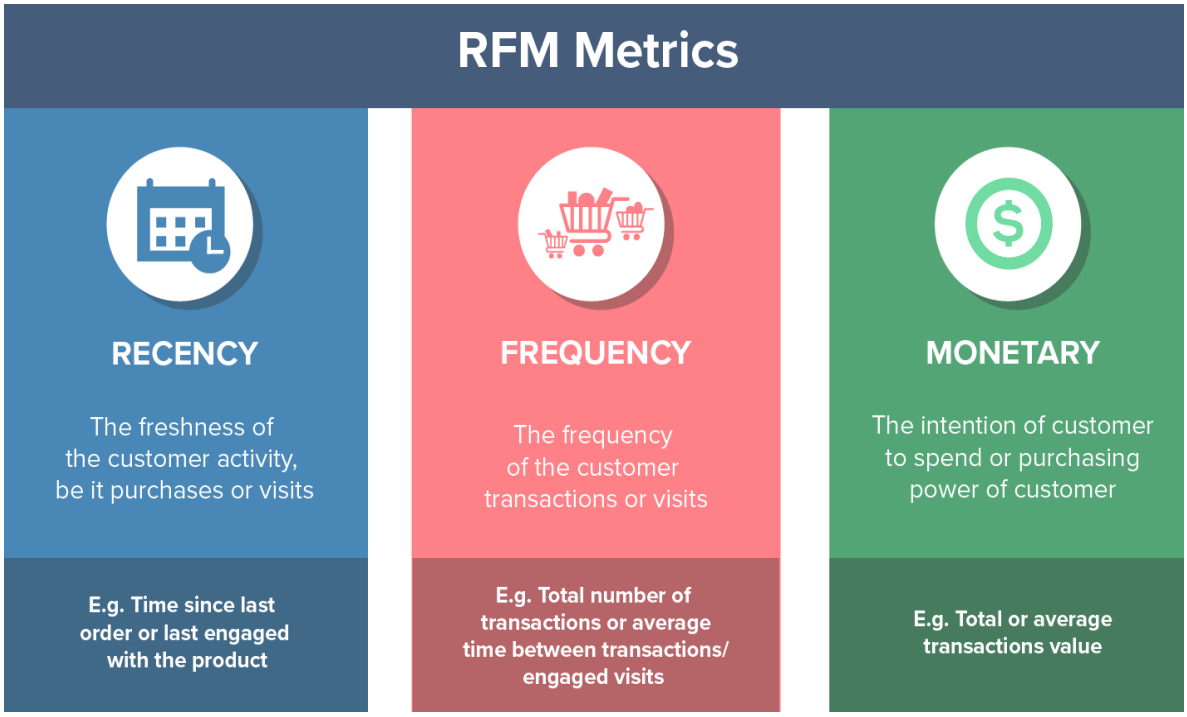
2.3.3 Ανάλυση Μοντέλου

Το RFM μοντέλο όπως υποδεικνύει και η ονομασία του βασίζεται στις τιμές των Recency, Frequency και Monetary. Οι μετρικές αυτές αποτελούν κλειδιά της καταναλωτικής συμπεριφοράς και ορίζονται παρακάτω εκτενώς.

- **Recency**: Ορίζεται ο χρόνος από την τελευταία καταγεγραμμένη δραστηριότητα του πελάτη/καταναλωτή. Ως δραστηριότητα μπορεί να θεωρηθεί είτε η επίσκεψη στο κατάστημα είτε η συναλλαγή από το κατάστημα. Η επιλογή είναι ανάλογη κάθε φορά με το είδος και τον σκοπό της ανάλυσης. Και στις δύο περιπτώσεις όσο πιο πρόσφατη είναι η δραστηριότητα ενός πελάτη τόσο πιο πιθανό ν' ανταποκριθεί σε μία νέα «επικοινωνία» της εταιρείας (Διαφήμιση, Ενημερωτικά mail κ.α.)

- **Frequency:** Ορίζεται η συχνότητα δραστηριότητας – συναλλαγής του πελάτη με την επιχείρηση. Η συχνότητα αποτελεί μετρική δέσμευσης του πελάτη απέναντι στην εταιρεία και τον καθιστά πιο πιστό καταναλωτή έναντι εκείνων που δεν έχουν συχνές επισκέψεις ή αγορές.
- **Monetary**: Ορίζεται το ποσό συναλλαγής του πελάτη με την επιχείρηση μέσα σ' ένα συγκεκριμένο χρονικό διάστημα. Μ' αυτόν τον τρόπο μπορεί ν' ανιχνευθεί η κατηγορία των πελατών που ξοδεύουν μεγάλα ποσά μέσα σε μια εταιρεία («Big Spenders») και μ ' αυτόν τον τρόπο να διαχειριστούν ανάλογα σε σχέση με άλλους πελάτες που ξοδεύουν λιγότερα. Η Monetary τιμή διαιρεμένη από την Frequency δίνει σαν αποτέλεσμα μία μέση τιμή συναλλαγής ανά επίσκεψη, παράγοντας πολύ σημαντικός για δευτερεύουσες αναλύσεις. Εικόνα 2.3: Οι Μετρικές που ορίζουν το RFM Μοντέλο.

Εικόνα 2.3: Βασικές Μετρικές που ορίζουν το R,F,M



Πηγή: https://clevertap.com/blog/rfm-analysis/

Οι παράγοντες της RFM ανάλυσης αντικατοπτρίζουν στην πράξη τις εξής βασικές διαπιστώσεις:

- Όσο πιο πρόσφατη μια αγορά , τόσο μεγαλύτερη ανταπόκριση υπάρχει από τον πελάτη σε νέα promotion της εταιρείας
- Όσο πιο συχνά ένας πελάτης επισκέπτεται και αγοράζει από μία επιχείρηση, τόσο ποιο ικανοποιημένος δείχνει να είναι και αντίστοιχα ισχυροποιείται η δέσμευση μεταξύ τους.

- Όσο μεγαλύτερο το ποσό που ξοδεύει ένας καταναλωτής αυτόματα τον κατατάσσει σε μία κατηγορία μόνη του έναντι αυτών που είναι πιο συντηρητικοί στο κομμάτι αυτό.

2.3.4 Χρησιμότητα

Η εφαρμογή του RFM μοντέλου από μία εταιρία, δίνει μία ποιοτική εικόνα των πελατών της σε σχέση με την αγοραστική τους συμπεριφορά. Πιο συγκεκριμένα μία εταιρεία μπορεί ν ‘ ανιχνεύσει:

- Τους καλύτερους πελάτες της.
- Ποιοι πελάτες είναι πιθανότερο ν’ ανταποκριθούν σ’ ενημερωτικές καμπάνιες της.
- Ποιοι έχουν δυναμική να εξελιχθούν σε πιστούς καταναλωτές.
- Ποιοι πελάτες ακριβοπληρώνουν.
- Ποιοι είναι πιθανό να εγκαταλείψουν την αγοραστική σχέση (customer churn).
- Ποιοι έχουν δυνατότητα να διατηρηθούν.

Η παραπάνω λίστα είναι ενδεικτική των ερμηνειών που δίνονται με βάση τις κατηγορίες των πελατών καθώς διαφορετικοί συνδυασμοί των τιμών R, F , M οδηγούν σε διαφορετικές εκτιμήσεις της αγοραστικής συμπεριφοράς.

2.3.5 Εφαρμογή - Εκτέλεση

Το RFM μοντέλο αποτελεί ένα λειτουργικό μοντέλο με πολλές δυνατότητες στον τρόπο χρήσης του. Για την εφαρμογή του, απαιτούνται 3 βασικά βήματα όπως αυτά ακολουθούν παρακάτω:

Βήμα 1ο

Για τον κάθε μοναδικό πελάτη που ορίζει το ID, αποδίδονται οι τιμές R , F , M, οι οποίες αντικατοπτρίζουν τις μέρες από την τελευταία επίσκεψη, τον αριθμό επισκέψεων και το ποσό που διέθεσε ο συγκεκριμένος πελάτης μέσα στο ορισμένο χρονικό διάστημα. Το χρονικό διάστημα και συγκεκριμένα ο χρόνος έναρξης, η διάρκεια και πολλές φορές η μονάδα μέτρησης (ώρα, μέρες, μήνες) επιλέγονται από τον αναλυτή ανάλογα με το σκοπό της ανάλυσης.

Βήμα 2ο

Οι τιμές R, F, M όπως προκύπτουν αυτούσιες δεν εξυπηρετούν την κατηγοριοποίηση καθώς είναι εξίσου μοναδικές με τον πελάτη. Για τον λόγο αυτό η κάθε τιμή χρειάζεται να χωριστεί

σε βαθμίδα ενδεικτική της σειράς κατάταξή της στο σύνολο. Ο συνδυασμός των τιμών RMF που προκύπτουν από τις βαθμίδες ονομάζεται και RFM score. Συνηθίζεται, όπως θα εφαρμοστεί και στην παρούσα εργασία, οι βαθμίδες να είναι 4 και συγκεκριμένα η βαθμίδα 1 να είναι και η επιθυμητή. Ο πελάτης με RFM score 111 θεωρείται ο καλύτερος πελάτης μιας εταιρείας.

Βήμα 3ο

Το 3ο βήμα είναι ο διαχωρισμός των ομάδων πελατών και η απόδοση χαρακτηριστικών σε αυτές βάσει των RFM score που εντοπίζονται στο σύνολο. Αυτή η διαδικασία είναι πολύ σημαντική καθώς αποτελεί την αρχή της αξιολόγησης και ερμηνείας των δεδομένων. Οι διάφοροι τύποι επικοινωνίας-προσέγγισης που θα επιλεχθούν για κάθε ομάδα αρχίζει να διαφαίνεται σε αυτό το στάδιο. Είναι ιδιαίτερα βοηθητικό να δοθούν ονόματα στην κάθε ομάδα πελατών και ενδεικτικά κάποιες από τις πιο βασικές είναι οι εξής:

- **Καλύτεροι Πελάτες (Champions)**: Οι πελάτες με πρόσφατες, συχνές και σίγουρα ακριβές αγορές. Συγκεντρώνουν RFM score 111.
- **Νέοι Πελάτες (High-spending New Customers):** Νέοι πελάτες θεωρούνται οι πελάτες με μοναδική αλλά πολύ πρόσφατη και ακριβή αγορά. Αναγνωρίζονται από τα score 141 και 142.
- **Σταθεροί - Ενεργοί πελάτες (Lowest-Spending Active Loyal Customers):** Η συγκεκριμένη ομάδα αποτελείται από πιστούς πελάτες με πρόσφατες, συχνές αγορές αλλά όχι ιδιαίτερα ακριβές. Συγκεντρώνουν RFM Score 113 ή 114,
- **Χαμένοι Πελάτες (Churned Best Customers) :** Οι πελάτες αυτής της ομάδας συνήθιζαν ν' αγοράζουν συχνά και αρκετά ακριβά ωστόσο έχει παρέλθει μεγάλο χρονικό διάστημα από την τελευταία αγορά. Συνήθως έχουν score 411, 412, 421, 422.

Οι παραπάνω ομάδες προκύπτουν από τους διαφορετικούς συνδυασμούς των βαθμίδων των μετρικών R, F , M. Για περισσότερη ακρίβεια πρόκειται για διατάξεις χωρίς απαραίτητα διαφορετικά στοιχεία και συγκεκριμένα εδώ ο μέγιστος αριθμός διατάξεων είναι $4^3 = 64$. Καθώς 64 ομάδες δεν αποτελούν ουσιαστική κατηγοριοποίηση, ο αναλυτής ανιχνεύει ποιες από τις ομάδες έχουν παρόμοια συμπεριφορά και μπορούν να προσεγγιστούν με κοινή πολιτική διαφήμισης.

2.3.6 Διαχείριση Αποτελεσμάτων

Η τμηματοποίηση των πελατών όπως αναλύθηκε στην προηγούμενη ενότητα, προκύπτει με βάση τα RFM score που υποδεικνύουν παρόμοιες αγοραστικές συμπεριφορές. Ωστόσο το ίδιο RFM score έχει διαφορετική ερμηνεία για μια εταιρία απ' ότι για κάποια άλλη. Ένα χαρακτηριστικό παράδειγμα εξαιρώντας τους πελάτες με RFM score 111, για τους οποίους δεν χωρά αμφισβήτηση ότι πρόκειται για τους καλύτερους, αφορά τους πελάτες

που συγκεκτρώνουν score 131, 231 και παρόμοιες διατάξεις που δείχουν μικρή συχνότητα. Μία εταιρεία με ηλεκτρικά είδη δεν μπορεί ν' αναμένει ότι κάποιος πελάτης θ' αγοράζει άσπρες συσκευές σε μηναία βάση. Σε αυτήν την περίπτωση η εταιρεία χρειάζεται να δώσει βαρύτητα στις τιμές Recency και Monetary και επομένως η ομάδα των πελατών με αυτά τα score δεν αντικατοπτρίζει απαραίτητα μη πιστούς πελάτες Αντιθέτως για μία εταιρεία ρουχισμού ή καλλυντικών η οποία βασίζεται στην συχνότητα αγορών/επισκέψεων, η συγκεκριμένα ομάδα πελατών υποδεικνύει ότι η αγοραστική σχέση με την εταιρεία είναι αδύναμη. Οι εταιρείες ρουχισμού, καλλυντικών καθώς και οι εταιρείες καταναλωτών αγαθών χρειάζεται να δίνουν ιδιαίτερη βαρύτητα στις Recency και Frequency τιμές έναντι της Monetary τιμής.

Παρατηρείται πώς ενώ στόχος κάθε εταιρείας είναι η αύξηση των κερδών μέσω της πώλησης ακριβών προϊόντων, δεν δίνεται πάντα αντίστοιχα το ίδιο βάρος στην Monetary τιμή ενός πελάτη. Αυτό εξηγείται διότι τις περισσότερες φορές, η συγκεκριμένη τιμή είναι άμεσα εξαρτημένη από τις άλλες δύο. Ένας ενεργός πελάτης με συχνές αγορές, αθροιστικά είναι κερδοφόρος για μία εταιρεία. Επομένως η συλλογή όλων των πιθανών συνδυασμών των R, F, M θα μπορούσε ν' αρκεστεί στους συνδυασμούς των R και F (εδώ 4^2=16). Γενικότερα η αξιολόγηση και "μετάφραση" των RFM scores είναι σε άμεση σχέση με την φύση της εταιρείας και για ορθότερη ανάλυση πρέπει να λαμβάνεται υπόψιν.

Σε συνέχεια και αφού έχουν εντοπιστεί οι επιθυμητές ομάδες, σειρά έχει ο σχεδιασμός των τακτικών που θ' ακολουθήσει μία εταιρεία και συγκεκριμένα το τμήμα του Marketing απέναντι στους πελάτες. Για κάθε ομάδα πελατών, υπάρχει συγκεκριμένος σκοπός όπως Προσέγγιση, Διατήρηση, Επαναπροσέγγιση και αντίστοιχα πληθώρα τακτικών επίτευξης του επικείμενου σκοπού.

Στην εικόνα που ακολουθεί παρουσιάζεται ο διαχωρισμός των πελατών σε ομάδες σε συνάρτηση των τιμών R, F και M. Όπως είναι λογικό πάνω δεξιά παρατηρούνται οι καλύτεροι πελάτες ενώ κάτω αριστερά οι πελάτες που έχουν διακόψει, όπως όλα δείχνουν, την αγοραστική τους σχέση με την εταιρεία Για κάθε ομάδα από τις παρακάτω θα παρουσιαστούν ενδεικτικά κάποιοι τρόποι διαχείρησής της από το τμήμα Μάρκετινγκ.

Εικόνα 2.4: Γραφικές Περιοχές των Ομάδων Πελατών

Πηγή: https://how-many-steps-inc.webflow.io/rfm-segmentation-overview–

1. **Champions - Καλύτεροι Πελάτες**

   Πρόκειται για τους καλύτερους πελάτες οπότε μία έξυπνη κίνηση είναι η ανταμοιβή τους (δώρο, ειδική προσφορά κ.α) με σκοπό να διατηρηθεί η αγοραστική σχέση στον ίδιο βαθμό. Επίσης μία έξυπνη κίνηση της εταιρείας είναι η συλλογή αξιολογήσεων μέσω ερωτηματολογίων. Με αυτόν τον τρόπο θα μπορούσε να εκτιμηθεί η ανταπόκρισή των πελατών αυτών σε νέα προϊόντα /υπηρεσίας που σκοπεύει να λανσάρει η εταιρεία.

2. **Loyal Customs – Πιστοί Καταναλωτές**

   Στην συγκεκριμένη ομάδα πελατών θα μπορούσε να εφαρμοστεί η τακτική του Up-selling. Πρακτικά αυτό σημαίνει ότι θα προτείνονται μέσω newsletters βελτιωμένα και ακριβότερα προϊόντα ίδιας κατηγορίας ή αναβαθμισμένες εκδόσεις προϊόντων παλαιότερων αγορών. Αυτή η κίνηση στοχεύει στην Monetary τιμή ενώ δώρα και κουπόνια προσφορών ενισχύουν την σχέση δέσμευσης του πελάτη με την εταιρεία διατηρώντας τις τιμές R,F στον ίδιο ικανοποιητικό βαθμό.

3. **Promising – Υποσχόμενοι Πελάτες**

   Εδώ θα μπορούσε, η πρόταση για συνδρομή ή πρόγραμμα ανταμοιβής πόντων να λειτουργήσει θετικά αυξάνοντας την συχνότητα αγορών και επισκέψεων. Επίσης οι προτάσεις αγορών αποτελούν εξίσου έναν ελκυστικό τρόπο.

4. **New Customers – Νέοι Πελάτες**

   Δώρα και εκπτώσεις στις πρώτες αγορές. Εφαρμογή post-sale υπηρεσιών. Πρακτικά αυτό μεταφράζεται ως εξυπηρέτηση των πελατών σε σχέση με τον τρόπο χρήσης/λειτουργίας

των προϊόντων που ήδη έχουν αποκτήσει. Η επιδίωξη προσωποποιημένης σχέσης με τον πελάτη βοηθάει να χτιστεί μία σχέση δέσμευσης με την εταιρεία.

5. **Abandoned Checkouts – Πελάτες που δεν ολοκληρώνουν την αγορά:**

   Εφαρμογή pre-sale υπηρεσιών όπως επικοινωνία με σκοπό την αναγνώριση των αναγκών και προτιμήσεων του πελάτη. Επιδίωξη ανάπτυξης προσωποποιημένης σχέσης όπως και στους νέους πελάτες.

6. **Warm Leads – Ένθερμοι Ακόλουθοι**

   Πρόκειται για πελάτες που στο παρελθόν έχουν δείξει ενδιαφέρον για τα προϊόντα ή τις υπηρεσίες της επιχείρησης και πολλές φορές την «ακολουθούν» στα κοινωνικά μέσα. Σε αυτήν την περίπτωση μία μέθοδος επικοινωνίας που θα κατατοπίζει και θα λύνει τυχόν απορίες των πελατών (proactive support) μπορεί να έχει θετικό αποτέλεσμα για την επίτευξη αγοράς. Παράθυρα επικοινωνίας αντιπροσώπων της επιχείρησης στην ιστοσελίδα ή κάποιο κομμάτι με συχνές ερωταπαντήσεις (FAQ) είναι από τους πιο συχνούς τρόπους τέτοιου τύπου επικοινωνίας.

7. **Cold Leads – Μη ενδιαφερόμενοι Ακόλουθοι**

   Σε αντίθεση με την προηγούμενη ομάδα πελατών, οι συγκεκριμένοι πελάτες έχουν δείξει απροθυμία σε παρελθοντικές προσεγγίσεις της εταιρείας. Ωστόσο μετά από ένα εύλογο διάστημα η εταιρεία θα μπορούσε να επαναπροσεγγίσει τον πελάτη σε προσωπικό επίπεδο και να μάθει τυχόν ανάγκες ή προτιμήσεις του.

8. **Need Attention- Πελάτες που θέλουν προσοχή**

   Εδώ είναι οι πελάτες που έχουν ένα μέσο RFM score. Με σκοπό να μετακυλήσουν στην κατηγορία των υποσχόμενων πελατών, διάφορες προσφορές με συγκεκριμένο χρόνο αξιοποίησης ή προτάσεις βασιζόμενες στις προτιμήσεις και ανάγκες τους θα μπορούσαν να λειτουργήσουν θετικά.

9. **Shouldn’t Lose - Πελάτες που πρέπει να παραμείνουν στο αγοραστικό κοινό**

   Πρόκειται για Champion πελάτες του παρελθόντος που όμως έχουν καιρό να επισκεφτούν το κατάστημα (φυσικό/ηλεκτρονικό) ή να κάνουν κάποια συναλλαγή. Σε αυτήν την περίπτωση οι συγκεκριμένοι πελάτες πρέπει οπωσδήποτε να παραμείνουν και ν’ αναβαθμιστούν ξανά στο καλύτερο αγοραστικό κοινό. Σημαντικό είναι ν’ ανακτηθεί επικοινωνία, οπότε τηλεφωνική επικοινωνία με τον ίδιο τον πελάτη, ειδικές προσφορές, προσφορά δώρων στην επόμενη αγορά είναι κάποιοι από τους τρόπους επανάκτησης (win-back).

10. **Sleepers – Ανενεργοί Πελάτες**

    Με βάση την εικόνα, αυτοί οι πελάτες διέθεταν ένα σημαντικό ποσό στις αγορές τους αλλά έχει περάσει μεγάλο χρονικό διάστημα επανεμφάνισής τους. Προσωποποιημένα e-mails, ειδικές εκπτώσεις αλλά και έρευνα για ανίχνευση των προτιμήσεών τους είναι ενδεικτικά του ενδιαφέροντος της εταιρείας προς τους πελάτες και μπορούν να έχουν θετική επίδραση.

11. Lost – Χαμένοι Πελάτες

Αυτή η κατηγορία αφορά τους πελάτες που εξαρχής δεν είχαν ιδιαίτερα στενή αγοραστική σχέση με την επιχείρηση. Το κίνητρο για προσπάθεια επαναπροσέγγισής τους είναι χαμηλό. Ενημερωτικά mails υπενθύμισης του brand ή/ και εποχιακών εκπτώσεων και Black Friday προσφορών θα μπορούσαν ίσως να οδηγήσουν σε μία δεύτερη ευκαιρία.

2.3.7 Αξιοποίηση – Πλεονεκτήματα Χρήσης

Η αξιοποίηση των αποτελεσμάτων της RFM ανάλυσης από μία επιχείρηση και το τμήμα Marketing, όπως παρουσιάζεται ακολούθως έχει πολυάριθμα πλεονεκτήματα για την λειτουργικότητά της.

I. Αποδοτικότερο E-mail Marketing

Ομαδοποιώντας τους πελάτες με βάση την αγοραστική τους συμπεριφορά μπορεί να διασφαλιστεί ότι τα κατάλληλα e-mails απευθύνονται στους κατάλληλους πελάτες. Το RFM μοντέλο συμβάλλει ώστε να εφαρμοστεί η βασική αρχή του online Marketing; το σωστό Προϊόν, για τον σωστό Πελάτη, με την σωστή Προώθηση (3 P's of Marketing).

II. Υψηλότερη πραγματική αξία του πελάτη μακροπρόθεσμα (Customer Lifetime Value)

Όπως είναι λογικό, τα πάντα σε μία επιχείρηση είναι άμεσα συνδεδεμένα. Στην περίπτωση αυτή, η πιο προσωποποιημένη διαφήμιση ,παραπάνω, είναι το κλειδί για την αύξηση παραμονής των πελατών και αντίστοιχα την αποφυγή «εγκατάλειψής» τους καθώς και την αύξηση των πωλήσεων με χρήση της τακτικής up-selling και cross- selling.Αυτό συνεπάγεται αύξηση των πωλήσεων στους ήδη υπάρχοντες πελάτες και ταυτόχρονα αύξηση της μακροπρόθεσμης αξίας τους για την επιχείρηση. Έχει παρατηρηθεί ότι οι πελάτες που επανακτώνται με πιο στοχευμένη προσέγγιση, αγοράζουν πιο συχνά και ακριβά σε σχέση με νεοαποκτηθέντες πελάτες.

III. Μεγαλύτερη απήχηση των νέων προϊόντων κυκλοφορίας

Εστιάζοντας στους καλύτερους και πιστούς καταναλωτές και αντλώντας πληροφορίες από τις συγκεκριμένες ομάδες, είναι σημαντικό ο σχεδιασμός και η ανάπτυξη των νέων προϊόντων να διαμορφώνεται με βάση τις προτιμήσεις και ανάγκες τους. Με αυτόν τον τρόπο, η κυκλοφορία αυτών είναι πιθανότερο να έχει την επιθυμητή ανταπόκριση και προπώληση και να πετύχει τον τελικό σκοπό της, το κέρδος.

IV. Ενίσχυση της αξιοπιστίας και της δέσμευσης με τον πελάτη (Loyalty & Engagement)

Η ανίχνευση των υποσχόμενων πελατών δίνει την ευκαιρία στην επιχείρηση να τους προσεγγίσει με τέτοιο τρόπο που να συγκαταλέγονται πλέον στους πιστούς πελάτες. Η προσωποποιημένη διαφήμιση και ένδειξη εκτίμησης της προτίμησης τους είναι πιθανό να οδηγήσουν στην επόμενη αγορά. Η προσφορά για συνδρομή με προνόμια είναι επίσης πολύ ενθαρρυντική.

V. Περιορισμός απώλειας πελατών και εγκαταλελειμμένων καλαθιών (Customer churn & Abandoned checkouts)

Για την αντιμετώπιση του συγκεκριμένου φαινομένου, οι επιχειρήσεις πρέπει να διαχειριστούν έξυπνα την ομάδα όχι τόσο των ήδη χαμένων πελατών όσο την ομάδα των ανενεργών για μεγάλο χρονικό διάστημα πελατών (sleepers) ώστε να τους επανακτήσουν. Σ' αυτές τις περιπτώσεις η επιχείρηση πρέπει να προχωρήσει σε μία προσφορά που είναι δύσκολο ο πελάτης ν' αρνηθεί και να προσπεράσει. Πολύ σημαντικό ακόμα είναι η διερεύνηση των αίτιων που οδήγησαν τον πελάτη να είναι ανενεργός. Αντίστοιχη διερεύνηση χρειάζεται και το φαινόμενο του «εγκαταλελειμμένου καλαθιού» (abandoned cart). Αν η επιχείρηση μπορέσει ν' αποσπάσει μέσω έρευνας τους λόγους μη ολοκλήρωσης της αγοράς είναι χρήσιμο για την αντιμετώπιση και αποφυγή αντίστοιχων μελλοντικών συμπεριφορών. Πολλές φορές ακόμα και αν η επανάκτηση των πελατών είναι αδύνατη, η πληροφορία που λαμβάνεται από αυτήν την έρευνα αυτής της ενέργειας, αποτελεί κέρδος για την εταιρεία.

VI. Μείωση κόστους Marketing

Οι διαφημίσεις στην γνωστή πλατφόρμα του Facebook αλλά και του Google έχουν κόστος όπως και η αποστολή Emails και SMS.Καθώς η διαφήμιση είναι απαραίτητο μέσο προσέγγισης των πελατών, σημασία έχει η αύξηση της αποτελεσματικότητάς της. Για να γίνει αυτό, απαραίτητο είναι το περιεχόμενο της διαφήμισης να είναι εξατομικευμένο βάσει της αγοραστικής συμπεριφοράς του καταναλωτή (targeting strategy). Επομένως ακόμα και αν ο συνολικός αριθμός των εξερχόμενων μηνυμάτων είναι ο ίδιος, η απόδοση αυτών θα είναι μεγαλύτερη. Αν μία εταιρεία μπορεί να μειώσει τον αριθμό των διαφημίσεων αυτό θα μπορούσε να είναι στην περίπτωση των χαμένων ή ανενεργών πελατών. Αν το κόστος για την επανάκτησή τους υπερβαίνει του κέρδους από αυτήν τότε μπορεί να μειωθεί η συχνότητα της διαφήμισης. Με βάση την «Αρχή του Pareto» αν μια εταιρεία επενδύσει 100 Ε σε διαφημιστική καμπάνια επαναπροσέγγισης μέσω της Facebook πλατφόρμας, έχει παρατηρηθεί ότι μόλις το ποσό των 20 Ε μόνο θα συνεισφέρει στο 80% της επιστροφής επί της επένδυσης (ROI). Εφαρμόζοντας την αποτελεσματικότητα κατά Pareto και με χρήση της RFM ανάλυσης αν μια εταιρεία μπορεί ν' ανιχνεύσει την ομάδα που θ΄ ανταποκριθεί στον μεγαλύτερο βαθμό, τότε μπορεί να επενδύσει μόνο στην συγκεκριμένη ομάδα το ποσό των 20 Ε ή ακόμα και όλο το αρχικό ποσό των 100 Ε και να μεγιστοποιήσει και άλλο την επιστροφή επί της επένδυσης.

VII. Εις βάθος κατανόηση της επιχείρησης (Business Understanding)

Πολλές εταιρίες ειδικά μικρές ή μικρομεσαίες εκλείπουν βαθύτερης γνώσης του αγοραστικού τους κοινού. Η RFM ανάλυση αποτελεί ένα εύχρηστο εργαλείο και η εφαρμογή του μπορεί να οδηγήσει για παράδειγμα στην ανακάλυψη πελατών που μία εταιρεία αγνοούσε μέχρι χθες ότι μπορούσε να εφαρμόσει την τακτική της συμπληρωματικής πώλησης. Ακόμα η επικοινωνία με άλλη ομάδα πελατών όπως για παράδειγμα των warm leaders θα μπορούσε να έχει θετικό αντίκτυπο για βελτιώσεις αναφορικά με την λειτουργικότητα της ιστοσελίδας κ.α. Μία επιχείρηση επαφίεται στους πελάτες της, η καλύτερη κατανόηση τούς θα πρέπει ν' αποτελεί προτεραιότητα γι' αυτήν και το RFM μοντέλο συμβάλλει σημαντικά σε αυτό.

2.4 Αξία Ενός Πελάτη – Customer Lifetime Value

Η αξία ενός πελάτη ή σύμφωνα με τον ξένο όρο Customer Lifetime Value (CLV) είναι αναμφίβολα, μία από τις σημαντικότερες μετρικές αξιολόγησης για κάθε αναπτυσσόμενη ή και ανεπτυγμένη επιχείρηση καθώς είναι ενδεικτική της αγοραστικής εμπειρίας και ικανοποίησης ενός πελάτη. Πρόκειται, ουσιαστικά, για μία τιμή που υποδεικνύει τα συνολικά έσοδα που μπορεί ν' αναμένει μία εταιρεία από έναν πελάτη σε βάθος χρόνου. Πιο επίσημα, η τιμής της Customer Lifetime Value αντανακλά την πραγματική αξία που έχει ο κάθε πελάτης για μία εταιρεία κατά την διάρκεια της αγοραστικής σχέσης και συγκριτικά με το κόστος της απόκτησής του (Cost-per-Acquisition) αλλά και διατήρησής του. Συγκεκριμένα, λαμβάνοντας υπόψιν την αξία των εσόδων που φέρνει ένας πελάτης σε μία επιχείρηση το CLV συσχετίζει τη μεταβλητή αυτή με την προβλεπόμενη διάρκεια ζωής του πελάτη και το αντικατοπτρίζει σε μία τιμή.

Ο υπολογισμό της τιμής CLV μιας επιχείρησής δεν είναι ιδιαίτερα πολύπλοκος. Αρχικά χρειάζεται να εντοπιστούν τα έσοδα του πελάτη προς την επιχείρηση και να οριστεί η μέση τιμής τους ή αλλιώς η μέση τιμή αγοράς. Στην συνέχεια η τιμή που προκύπτει θα πολλαπλασιαστεί με το μέσο ποσοστό συχνότητας αγοράς για να καθοριστεί η Αξία του Πελάτη. Για την Lifetime Value αρκεί η Αξία του Πελάτη να πολλαπλασιαστεί με την μέση διάρκεια «ζωής» του πελάτη.

Η χρησιμότητα της τιμής Customer Lifetime Value είναι πολυδιάστατη. Χάρη σε αυτήν προλαμβάνονται πιθανές, σημαντικές ζημίες κατά το λανσάρισμα των marketing campaign. Αφετέρου, η σχέση με τους πελάτες εμβαθύνεται, με κύριο γνώμονα την απόλυτη ικανοποίησή τους και κατά επέκταση την αύξηση της κερδοφορίας. Είναι λογικό ότι όσο πιο συχνά ένας πελάτης συνεχίζει να αγοράζει από μία εταιρεία, τόσο αυξάνεται η CLV. Επομένως προκύπτει ότι ένας αφοσιωμένος πελάτης, που ίσως να έχει αποκτηθεί δυσκολότερα, να αξίζει πολύ περισσότερο από δέκα νέους, οι οποίοι δεν σκοπεύουν να προβούν σε επαναλαμβανόμενες αγορές μελλοντικά.

Η γνώση της τιμής CVL από τις επιχειρήσεις συμβάλλει ώστε ν΄ αναπτύσσουν εύστοχες και αποτελεσματικές στρατηγικές για την απόκτηση νέων πελατών και τη διατήρηση των υφιστάμενων. Η εύρεση της τιμής CLV, διεξάγει σαφή συμπεράσματα για την επιχείρηση, βελτιώνοντας αντίστοιχα το CPA (κόστος απόκτησης ανά πελάτη). Ακόμη, ιδανικά, παίζει σημαντικό ρόλο στην υιοθέτηση μίας περισσότερο πελατοκεντρικής φιλοσοφίας, στην οποία οι ανάγκες και η επίλυση προβλημάτων των πελατών κατέχουν πρωταγωνιστικό ρόλο. Επιπρόσθετα, η μεταβλητή της τιμής Customer Lifetime Value, μπορεί να σταθεί ως στρατηγικός σύμβουλός σε περιόδους που υπάρχει ανάγκη από άμεση ρευστότητα (cash flow), αποτρέποντάς από καταστροφικά λάθη (π.χ. υπερβολικό ύψος εκπτώσεων) και οριοθετώντας με σύνεση τις δαπάνες προωθητικών εκστρατειών.

Τέλος, η ύπαρξη CLV στα πλαίσια της γενικότερης στρατηγικής διευρύνει τους επιχειρηματικούς ορίζοντες μιας εταιρείας. Ειδικότερα, κινητοποιεί την καινοτομία και τη δημιουργικότητά, ωθώντας στην αναζήτηση σύγχρονων πηγών προστιθέμενης αξίας, οι οποίες συμβάλλουν στην εκτόξευση των πωλήσεών σας και κατά συνέπεια, της συνολικής τιμής customer lifetime value.

2.5 Κατηγοριοποίηση στην Μηχανική Μάθηση

2.5.1 Εισαγωγή

Η κατηγοριοποίηση (Classification) είναι μια από τις βασικότερες μεθόδους στην Εξόρυξη Δεδομένων, με εφαρμογές σε πολλούς τομείς όπως ενδεικτικά τον οικονομικό, ιατρικό και εμπορικό. Η κατηγοριοποίηση ανήκει στους επιβλεπόμενους τρόπους μάθησης. Στόχος της επιβλεπόμενης μάθησης είναι η ανίχνευση συσχέτισης ανάμεσα σ’ ένα γνώρισμα - στόχο και σε ένα σύνολο γνωρισμάτων. Το γνώρισμα – στόχος λειτουργεί ως εξαρτημένη μεταβλητή, ενώ τα υπόλοιπα γνωρίσματα ως ανεξάρτητες μεταβλητές. Η επιβλεπόμενη μάθηση επιτυγχάνει την δημιουργία ενός μοντέλου ο οποίος είναι ικανός να προβλέπει τις τιμές της εξαρτημένης μεταβλητής χρησιμοποιώντας τις ανεξάρτητες μεταβλητές. Το μοντέλο μπορεί να έχει διάφορες μορφές, όπως ένα σύνολο κανόνων ή μια εξίσωση ή το πλέγμα των νευρώνων και συνδέσεων ενός Νευρωνικού Δικτύου.

Στην επιβλεπόμενη μάθηση εκτός από την Κατηγοριοποίηση (Classification) ανήκει ακόμα, η Παλινδρόμηση (Regression). Η κατηγοριοποίηση και η παλινδρόμηση έχουν πολλές ομοιότητες. Στόχος των δύο μεθόδων είναι η πρόβλεψη των τιμών ενός γνωρίσματος, με χρήση άλλων γνωρισμάτων ενώ και στις δύο περιπτώσεις χρησιμοποιείται ένα σύνολο δεδομένων εκπαίδευσης, κατά την επεξεργασία του οποίου κατασκευάζεται το μοντέλο. Η διαφορά ανάμεσα στην κατηγοριοποίηση και στην παλινδρόμηση έχει άμεση σχέση με τον τύπο της εξαρτημένης μεταβλητής. Στόχος της παλινδρόμησης είναι η πρόβλεψη μιας εξαρτημένης μεταβλητής, η οποία περιέχει συνεχόμενες (αριθμητικές) τιμές. Αντιθέτως, κατηγοριοποίηση είναι η πρόβλεψη διακριτών ονομαστικών τιμών. Οι τιμές αυτές είναι συγκεκριμένες

και εξαρχής ορισμένες ενώ ορίζουν την κλάση (κατηγορία) στην οποία ανήκει κάθε αντικείμενο. Για τον λόγο αυτό, η εξαρτημένη μεταβλητή σε προβλήματα κατηγοριοποίησης καλείται και γνώρισμα κλάσης.

Η ένταξη αντικειμένων σε ομάδες αποτελεί εφαρμογή και για άλλη μέθοδο Εξόρυξης Δεδομένων, την Συσταδοποίηση (Clustering). Η Συσταδοποίηση ομαδοποιεί τα αντικείμενα με βάση τα κοινά χαρακτηριστικά τους. Το βασικό σημείο διαφοροποίησης είναι ότι οι συστάδες καθώς και ο αριθμός τους δεν είναι γνωστές και δεν έχουν οριστεί μέχρι και την εκτέλεση. Επίσης, τα δεδομένα δεν περιέχουν κάποιο πεδίο που να καθορίζει την ομάδα – συστάδα. Για τον λόγο αυτό η Συσταδοποίηση θεωρείται μη-επιβλεπόμενη μάθηση.

Με αναφορά πάλι στο κεντρικό θέμα της παρούσας παραγράφου, είναι χρήσιμο να οριστεί ο επαγωγικός αλγόριθμος που συναντάται στην Κατηγοριοποίηση. Ένας Επαγωγικός Αλγόριθμος είναι μια οντότητα, η οποία επεξεργάζεται ένα σύνολο δεδομένων και κατασκευάζει ένα μοντέλο. Το μοντέλο είναι μια τυποποίηση, που περιγράφει τη γενίκευση της σχέσης ανάμεσα σε μια εξαρτημένη μεταβλητή και σε ένα σύνολο ανεξάρτητων μεταβλητών. Πιο συγκεκριμένα, το μοντέλο δέχεται ως είσοδο τις τιμές των ανεξάρτητων μεταβλητών και παράγει ως έξοδο μια τιμή για την εξαρτημένη μεταβλητή. Το μοντέλο, που προκύπτει, εφαρμόζεται για να χρησιμοποιηθεί για την πρόβλεψη της κλάσης νέων παρατηρήσεων.

2.5.2 Επαγωγικοί Αλγόριθμοι και Μοντέλα

Στους επαγωγικούς αλγορίθμους ή μεθόδους κατηγοριοποίησης ανήκουν τα Δένδρα Αποφάσεων, τα Μπεϋζιανά Δίκτυα, τα Νευρωνικά Δίκτυα τύπου Multilayer Perceptron κα. Τα μοντέλα που προκύπτουν από κάθε μέθοδος στηρίζονται σε διαφορετική θεωρία και λογική και έχουν μεγάλες διαφορές.

Στην παρούσα εργασία θα γίνει χρήση των αλγορίθμων του Δένδρου Αποφάσεων και του Μπεϋζιανού. ‘Ενα μοντέλο Δένδρου Αποφάσεων είναι μια δενδρική δομή, όπου κάθε κόμβος είναι ένας έλεγχος για κάποιο γνώρισμα, κάθε κλάδος είναι ένα αποτέλεσμα του ελέγχου και κάθε φύλο είναι μια απόφαση κατηγοριοποίησης .Ένα μοντέλο που βασίζεται στην Μπεϋζιανή Θεωρία, περιγράφεται με έναν κατευθυνόμενο ακυκλικό γράφο και μια κατανομή πιθανοτήτων συσχέτισης μεταξύ των μεταβλητών. Στον γράφο κάθε κόμβος αντιστοιχεί σε μια μεταβλητή.

Όπως αναφέρθηκε και προηγουμένως, ένας επαγωγικός αλγόριθμος κατόπιν της επεξεργασίας ενός συνόλου δεδομένων παράγει ένα μοντέλο. Αν με I συμβολίζεται ένα επαγωγικός αλγόριθμος και με D ένα σύνολο δεδομένων, τότε το I(D) συμβολίζει το μοντέλο που προκύπτει από την επεξεργασία του συνόλου δεδομένων D από τον I. Αν ο ίδιος αλγόριθμος επεξεργαστεί ένα διαφορετικό σύνολο δεδομένων D’, τότε το αποτέλεσμα θα είναι ένα διαφορετικό μοντέλο I(D’). Εάν το D’ περιέχει διαφορετικά πεδία από το D, είναι προφανές ότι τα δύο μοντέλα θα είναι διαφορετικά. Η διαφορά υπάρχει ακόμα και αν τα δύο σύνολα δε-

δομένων περιέχουν τα ίδια πεδία, με διαφορετικές παρατηρήσεις. Για παράδειγμα στα Δένδρα Αποφάσεων, μικρές διαφορές στα σύνολα δεδομένων έχουν σαν αποτέλεσμα τη δημιουργία σημαντικά διαφορετικών μοντέλων.

Μια μέθοδος κατηγοριοποίησης μπορεί να είναι αιτιοκρατική (deterministic) ή στοχαστική (stochastic). Οι στοχαστικές μέθοδοι καλούνται και πιθανολογικές (probabilistic). Μια αιτιοκρατική μέθοδος δημιουργεί μοντέλα, όπου σε κάθε κλάση αντιστοιχείται μία παρατήρηση ενώ στα στοχαστικά μοντέλα για κάθε κλάση υπολογίζεται η πιθανότητα να ανήκει η παρατήρηση σε αυτήν. Παράδειγμα αιτιοκρατικής μεθόδου είναι τα Δένδρα αποφάσεων τύπου C4.5, ενώ παράδειγμα στοχαστικής μεθόδου είναι τα Μπεϋζιανά Δίκτυα

2.5.3 Στάδια Κατηγοριοποίησης

Η κατηγοριοποίηση χωρίζεται σε τρία στάδια, στο στάδιο της εκπαίδευσης του μοντέλου, στο στάδιο της επαλήθευσης του μοντέλου και στο στάδιο της χρήσης του μοντέλου. Αναλυτικότερα, οι εργασίες που λαμβάνουν χώρα σε κάθε στάδιο είναι οι ακόλουθες:

- **Εκπαίδευση Μοντέλου**

Στο στάδιο αυτό, πραγματοποιείται η ανάλυση του συνόλου δεδομένων προκειμένου να βρεθούν οι σχέσεις μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών. Το αποτέλεσμα αυτής της ανάλυσης είναι η κατασκευή ενός μοντέλου. Το σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση του μοντέλου, ονομάζεται σύνολο εκπαίδευσης (training data set). Η επιλογή του συνόλου εκπαίδευσης είναι καθοριστικής σημασίας, καθώς το μοντέλο που προκύπτει αποτυπώνει τις σχέσεις που υπάρχουν στο σύνολο εκπαίδευσης. Επομένως μεροληπτικά σύνολα εκπαίδευσης παράγουν μεροληπτικά μοντέλα, γεγονός που δεν είναι επιθυμητό.

- **Επαλήθευση Μοντέλου**

Στο στάδιο αυτό ελέγχεται η ακρίβεια πρόβλεψης του μοντέλου, η ικανότητα του δηλαδή να προβλέπει σωστά την κλάση των παρατηρήσεων. Το μοντέλο εκπαιδεύεται με παρατηρήσεις, όπου η κλάση είναι γνωστή και δίνει την αντίστοιχη πρόβλεψη. Στην συνέχεια γίνεται σύγκριση ανάμεσα στην πρόβλεψη του μοντέλου και την πραγματική τιμή της κλάσης. Αν το μοντέλο έχει ορθή πρόβλεψη για την κλάση ενός ικανοποιητικού ποσοστού παρατηρήσεων, τότε θεωρείται επιτυχημένο και μπορεί να χρησιμοποιηθεί για τη διατύπωση προβλέψεων. Η διαδικασία δοκιμής του μοντέλου καλείται επαλήθευση (validation) και το σύνολο δεδομένων που χρησιμοποιείται για τη δοκιμή καλείται σύνολο επαλήθευσης (validation set ή test set). Σκοπός ενός μοντέλου είναι να αξιοποιηθεί καθολικά για τη διατύπωση προβλέψεων στην πραγματική ζωή και όχι για την ανάλυση ενός συγκεκριμένου συνόλου δεδομένων. Για τον λόγο αυτό, το μοντέλο πρέπει να αποδείξει την επιτυχή πρόβλεψη

της κλάσης άγνωστων παρατηρήσεων πέρα απ' αυτών που χρησιμοποιήθηκαν για την εκπαίδευση του.

- **Χρήση Μοντέλου**

Στο στάδιο αυτό και αφού το μοντέλο, έχει εκπαιδευτεί και επαληθευτεί, χρησιμοποιείται για τη διατύπωση προβλέψεων. Πιο συγκεκριμένα, μία νέα παρατήρηση, της οποίας η κλάση είναι άγνωστη, εισάγεται στο μοντέλο, το οποίο χρησιμοποιώντας τις τιμές των ανεξάρτητων μεταβλητών υπολογίζει την τιμή της κλάσης.

2.5.4 Κριτήρια Αξιολόγησης Μεθόδων Κατηγοριοποίησης

Για την εφαρμογή της Κατηγοριοποίησης, υπάρχουν πολλοί και διαφορετικοί αλγόριθμοι. Πιο συγκεκριμένα, τα Νευρωνικά Δίκτυα, έχουν μεγάλη ικανότητα στην πρόβλεψη της κλάσης άγνωστων παρατηρήσεων. Θα μπορούσε η μέθοδος αυτή να θεωρηθεί καταλληλότερη από κάποια άλλη όμως η ακρίβεια δεν είναι το μοναδικό κριτήριο αξιολόγησης των μεθόδων κατηγοριοποίησης. Παρακάτω ακολουθούν αναλυτικά τα κριτήρια με τα οποία αξιολογείται η κάθε μέθοδος κατηγοριοποίησης.

- **Ακρίβεια Πρόβλεψης (Accuracy)**

Πρόκειται για την ικανότητα της επιλεγμένης μεθόδου - μοντέλου να προβλέπει σωστά την κλάση άγνωστων παρατηρήσεων. Προφανώς είναι από τα σημαντικότερα κριτήρια και μεγάλο μέρος της εστιάζει στην εύρεση μεθόδων υψηλών επιδόσεων.

- **Ταχύτητα (Speed)**

Η ταχύτητα σχετίζεται με την πολυπλοκότητα της μεθόδου και το υπολογιστικό κόστος που αυτή συνεπάγεται. Η εκτέλεση πολύπλοκων αλγορίθμων συνεπάγονται συνήθως καθυστερήσεις καθώς απαιτούν μεγαλύτερο αριθμό εκτέλεσης υπολογισμών. Η καθυστέρηση μπορεί να εστιάζεται στη διαδικασία εκπαίδευσης και κατασκευή του μοντέλου αλλά και κατά την διάρκεια της χρήσης του.

Ορισμένες μέθοδοι, όπως τα Δένδρα Αποφάσεων, διαθέτουν γρήγορους αλγορίθμους με μικρό χρόνο κατασκευής των μοντέλων. Αντίθετα, μέθοδοι όπως τα Νευρωνικά Δίκτυα, απαιτούν πολύ περισσότερο χρόνο για την εκπαίδευση των μοντέλων όμως ο χρόνος χρήσης τους είναι πολύ μικρός. Τέλος, υπάρχουν μέθοδοι, όπως οι k-Πλησιέστεροι Γείτονες, όπου δεν χρειάζεται εκπαίδευσε ενός μοντέλου, ωστόσο ο χρόνος για την κατηγοριοποίηση νέων παρατηρήσεων είναι μεγάλος.

- **Ερμηνευσιμότητα (Interpretability)**

  Πρόκειται για το σημαντικό χαρακτηριστικό της μεθόδου να παράγει μοντέλα, τα οποία είναι κατανοητά από την ανθρώπινη λογική. Για παράδειγμα, στα Δένδρα Αποφάσεων ο τρόπος λήψης της απόφασης κατηγοριοποίησης είναι εύκολα κατανοητός και το μοντέλο μπορεί εύκολα μετατρέπεται σ΄ένα σύνολο κανόνων της μορφής EAN-TOTE. Αντίθετα, μοντέλα άλλων μεθόδων, όπως τα Νευρωνικά Δίκτυα και οι Μηχανές Διανυσμάτων Υποστήριξης, λειτουργούν έχουν πιο πολύπλοκο τρόπο λειτουργίας για την πρόβλεψη της κλάσης. Η ερμηνευσιμότητα είναι μια σημαντική ιδιότητα των μεθόδων κατηγοριοποίησης καθώς έτσι οι χρήστες των μοντέλων, μπορούν να επέμβουν με παραμέτρους και να φέρουν τ ‘ αποτελέσματα πιο κοντά στο επιθυμητό. Επίσης, στο μοντέλο μπορεί να καταγραφούν νέες και άγνωστες σχέσεις μεταξύ των δεδομένων. Αν το μοντέλο είναι ερμηνεύσιμο η ανίχνευση των νέων σχέσεων κατά την εφαρμογή της μεθόδου κατηγοριοποίησης θα χρησιμοποιηθεί ως εργαλείο ανάλυσης, ικανό να προσφέρει καινοτόμα γνώση.

- **Επεκτασιμότητα (Scalability)**

  Η επεκτασιμότητα αναφέρεται στην ικανότητα των μεθόδων να χειρίζονται εύκολα σύνολα μεγάλου όγκου δεδομένων. Η Μηχανική Μάθηση και η Στατιστική προσφέρουν μεθόδους κατηγοριοποίησης, ωστόσο η εφαρμογή αυτών σε σύνολα μεγάλων δεδομένων δεν είναι πάντα εύκολα εκτελέσιμη. Σε πολλές περιπτώσεις η υπολογιστική πολυπλοκότητα των μεθόδων ορίζεται από το πλήθος των παρατηρήσεων, ενώ οι περισσότεροι μέθοδοι απαιτούν την εγκατάσταση του συνόλου εκπαίδευσης στην κύρια μνήμη του υπολογιστή. Τα ζητήματα αυτά θέτουν αυτόματα περιορισμούς στην δυνατότητα εφαρμογής των μεθόδων. Ωστόσο η σημερινή εποχή επιτάσσει την αξιοποίηση των Μεγάλων Δεδομένων και είναι απαραίτητο η ανάλυση και οι μέθοδοι να προσαρμοστούν σε αυτό το χαρακτηριστικό. Όπως χαρακτηριστικά επισημαίνουν οι Fayyad, Piatetsky-Shapiro and Smyth (1996), η πρόκληση για την κοινότητα των ερευνητών Εξόρυξης Δεδομένων είναι η κατασκευή μεθόδων που διευκολύνουν τη χρήση αλγορίθμων εξόρυξης δεδομένων σε βάσεις δεδομένων του πραγματικού κόσμου.

- **Ανθεκτικότητα (Robustness)**

  Η Ανθεκτικότητα ορίζεται η ικανότητα των μεθόδων να πραγματοποιούν ορθές προβλέψεις ακόμα και όταν τα δεδομένα χαρακτηρίζονται από προβλήματα, όπως ο θόρυβος και οι χαμένες τιμές. Κατά την εφαρμογή Μεθόδων Μηχανικής Μάθησης είναι πολύ συνηθισμένο να συναντώνται θορυβώδη δεδομένα. Αυτό συμβαίνει διότι πολλές φορές τα δεδομένα προέρχονται από διαφορετικά πληροφοριακά συστήματα και η συνένωσή τους μπορεί να δημιουργήσει κενά σε κάποια πεδία. Επιπλέον, λάθη του ίδιου του υπαλλήλου που κάνει την καταγραφή – εισαγωγή ή ακόμα και σφάλματα του ίδιου του συστήματος είναι πολύ συχνά. Επομένως κρίνεται αναγκαίο εκτός της καλής προεπεξεργασίας των δεδομένων, οι μέθοδοι που χρησιμοποιούνται να μην μεροληπτούν απέναντι στις προβληματικές τιμές που τυγχάνει να υπάρχουν.

Κεφάλαιο 3 Τμηματοποίηση Πελατών – Πρακτικό Μέρος

"Above all else, show the data."

–Edward R. Tufte

3.1Εισαγωγή 26
3.2Προεπεξεργασία Δεδομένων – Data Preprocessing 31
3.3Εφαρμογή RFM Μοντέλου για Τμηματοποίηση Πελατών.....36
3.4Τμηματοποίηση Πελατών με βάση την τιμή Customer Lifetime Value 41
3.5Πρόβλεψη Κατηγορίας και Ενίσχυση RFM Μοντέλου με χρήση WEKA 43

3.1 Εισαγωγή

3.1.1 Εργαλεία Υλοποίησης Εργασίας

Το πρακτικό μέρος της εργασίας υλοποιήθηκε σε γλώσσα Python στο περιβάλλον εργασίας του λογισμικού Jupyter. Το κομμάτιτης Μηχανικής Μάθησης εφαρμόστηκε στο ελεύθερο λογισμικό Weka με κύρια χρήση των αλγορίθμων του Decision Tree και Naïve Bayes.

Python

Η Python επινοήθηκε και αναπτύχθηκε από τον Ολλανδό Guido van Rossum, το 1990. Διαθέτει ένα μεγάλο αριθμό βιβλιοθηκών που διευκολύνουν πολλές συνηθισμένες εργασίες και χαρακτηρίζεται για την ταχύτητα εκμάθησης της. Αναπτύσσεται ως open source software (ανοιχτό λογισμικό) και η διαχείρισή της γίνεται από τον μη κερδοσκοπικό οργανισμό Python Software Foundation. Ακριβώς λόγω του ανοικτού της κώδικα, η Python λειτουργεί σε πολλές πλατφόρμες και όλα προγράμματά σας μπορούν να δουλέψουν σε οποιαδήποτε από αυτές τις πλατφόρμες χωρίς να είναι απαραίτητες κάποιες αλλαγές. Η Python μπορεί να χρησιμοποιηθεί παντού, από το Linux και τα Windows μέχρι το Macintosh και το PlayStation. Η δομή της Pythonστηρίζεται στον αντικειμενοστραφή προγραμματισμό, δηλαδή έχει ως κεντρική έννοια την κλάση η οποία συνδέεται με τα αντικείμενα.

Οι βιβλιοθήκες της Python στην Επιστήμη Δεδομένων

Η μεγάλη δύναμη της Python είναι το πλήθος των ελεύθερων third-party βιβλιοθηκών. Της αρχές του 2019, το επίσημο Python repository περιείχε γύρω της 165 χιλιάδες projects. Σχεδόν για κάθε περίπτωση, υπάρχει διαθέσιμη μια υψηλής ποιότητας βιβλιοθήκη. Μεταξύ αυτών υπάρχουν της εξαιρετικές βιβλιοθήκες για την επιστήμη των δεδομένων, καλύπτοντας κάθε στάδιο της ανάλυσης των δεδομένων.

Pandas

Μια πολύ χρήσιμη και διαδεδομένη βιβλιοθήκη είναι η Pandas, που παρέχει υψηλής απόδοσης δομές δεδομένων - data structures για τον χειρισμό, τον καθαρισμό και την προετοιμασία των δεδομένων με σκοπό ν' αποτελέσουν input για την διαδικασία της μηχανικής μάθησης. Η παροχή "καθαρών" δεδομένων με τη σωστή μορφή είναι ζωτικής σημασίας για την πραγματοποίηση σωστών προβλέψεων στο machine learning, και σίγουρα είναι ένα απαραίτητο πρώτο βήμα. Η βιβλιοθήκη Pandas, χάρη στην κατάλληλη μορφοποίηση δίνει στα

τα δεδομένα μορφή ελαχιστοποιώντας τον κίνδυνο να συμπεριληφθούν προβληματικά ή ελλιπή δεδομένα στις αναλύσεις.

Scikit-learn

Αφού τα δεδομένα έχουν προεπεξεργαστεί σωστά από την βιβλιοθήκη Pandas, το επόμενο βήμα της μηχανικής μάθησης - machine learning, υλοποιείται με την βιβλιοθήκη scikit-learn. Ένα από τα πράγματα που κάνει η συγκεκριμένη βιβλιοθήκη, είναι ότι παρέχει έναν μεγάλο αριθμό αλγορίθμων μηχανικής μάθησης κάτω από ένα κοινό API. Αυτό σημαίνει ότι πολλές τεχνικές μπορούν να χρησιμοποιηθούν με την ίδια σύνταξη, ώστε να υπάρχει εύκολη εναλλαγή μεταξύ διαφορετικών τεχνικών ανάλογα με τον τύπο των δεδομένων που αναλύονται και το πρόβλημα που επιλύεται. Η βιβλιοθήκη Scikit-learn παρέχει επίσης εργαλεία για την επικύρωση των αποτελεσμάτων και τη διασφάλιση της βέλτιστης επιλογής των μοντέλων.

Jupyter

Το Jupyter Notebook είναι μια διαδικτυακή εφαρμογή, ελεύθερου λογισμικού και ανοιχτού κώδικα που επιτρέπει την δημιουργία και διαμοιρασμό εγγράφων που περιέχουν κώδικα, εξισώσεις, γραφήματα, απεικονίσεις και επεξηγηματικό κείμενο. Το Jupyter notebook προσφέρει δυνατότητες παρουσίασης μορφοποιημένου κείμενου, μαθηματικών εξισώσεων και συμβόλων, και διαδραστικά στοιχεία με εκτελέσιμο κώδικα υπολογιστή, μέσα από το γνώριμο περιβάλλον του web browser.

Weka

Το Weka (Waikato Environment for Knowledge Analysis) μια σουίτα λογισμικού για μηχανική μάθηση και Εξόρυξη Δεδομένων που αναπτύχθηκε απο το Πανεπιστήμιο του Waikato της Ν. Ζηλανδία. Είναι γραμμένο σε Java και ανήκει στην κατηγορία του λεγόμενου ελεύθερου λογισμικού (freeware) επιτρέποντας στους χρήστες να χρησιμοποιούν, αλλά και να τροποποιούν ελεύθερα το λογισμικό του. Είναι ένα από τα πιο διαδεδομένα λογισμικά Εξόρυξης Δεδομένων και έχει χρησιμοποιηθεί σε μεγάλο αριθμό επιστημονικών εργασιών ενώ η δημοφιλία του οφείλεται στα ειδικά χαρακτηριστικά του και στις δυνατότητες που προσφέρει. Συγκεριμένα το WEKA παρέχει ποικιλία μεθόδων για κατηγοριοποίηση, παλινδρόμηση, ανάλυση συστάδων, κανόνες συσχέτισης, προεπεξεργασία των δεδομένων καθώς και εργαλεία οπτικοποίησης.

Decision Tree J48

Ο Decision Tree J48 είναι ένας εποπτευόμενος αλγόριθμος μάθησης που βασίζεται στον αλγόριθμο C4.5. Ξεκινώντας από τα δεδομένα εκπαίδευσης, δημιουργείται ένα μοντέλο πρόβλεψης το οποίο χαρτογραφείται σε μια δομή δέντρου. Ο στόχος είναι να επιτευχθεί άριστη ταξινόμηση με ελάχιστο αριθμό αποφάσεων, αν και δεν είναι πάντα εφικτό λόγω θο-

ρύβου ή ασυνέπειας στα δεδομένα. Κάθε επίπεδο χωρίζει τα δεδομένα σύμφωνα με διαφορετικά χαρακτηριστικά, οι κόμβοι που δεν έχουν φύλλα αντιπροσωπεύονται από τα χαρακτηριστικά, ενώ οι κόμβοι των φύλλων αντιπροσωπεύουν την προβλεπόμενη μεταβλητή.
Τα βασικά στοιχεία του αλγορίθμου είναι:

- Ξεκινάει ελέγχοντας όλο το σύνολο δεδομένων εκπαίδευσης.

- Βρίσκει τα χαρακτηριστικά που δίνουν το μεγαλύτερο πληροφοριακό κέρδος με βάση την εντροπία ώστε να προχωρήσει στον καλύτερο διαχωρισμό των κόμβων.

- Το κλάδεμα είναι μια πολύ σημαντική τεχνική που πρέπει να χρησιμοποιηθεί στη δημιουργία δέντρων εξαιτίας των ακραίων τιμών. Είναι πιθανό να υπάρχουν δεδομένα που μπορεί να περιέχουν ελάχιστα υποσύνολα περιπτώσεων που δεν είναι καλά καθορισμένες. Για να τα ταξινομήσετε σωστά, μπορεί να χρησιμοποιηθεί κλάδεμα με τους δύο τύπους όπως παρακάτω:

    - Κλάδεμα μετά τη δημιουργία του δέντρου.
    - Online κλάδεμα (που εκτελείται κατά τη δημιουργία του δέντρου)

Naïve Bayes

Στη Μηχανική Μάθηση, οι Naïve Bayes κατηγοριοποιητές είναι μια οικογένεια πιθανοτικών ταξινομητών που βασίζονται στην εφαρμογή του θεωρήματος Bayes με ισχυρή ανεξαρτησία μεταξύ των χαρακτηριστικών.

Ο Naïve Bayes παραμένει από την δεκαετία του 1960 μια δημοφιλής μέθοδος για την κατηγοριοποίηση κειμένων καθώς και εκτίμηση εγγράφων σχετικά με το είδος τους όπως ανεπιθύμητο, νόμιμο, αθλητικό, πολιτικό, κλπ. Αυτό βασίζεται στην συχνότητα κάποιων λέξεων που ορίζονται ως χαρακτηριστικά. Οι ταξινομητές Naïve Bayes λειτουργούν κλιμακωτά απαιτώντας μια σειρά γραμμικών παραμέτρων στον αριθμό των μεταβλητών (χαρακτηριστικά) κατά την εκπαίδευσή τους.

Ο Naïve Bayes με την κατάλληλη προεπεξεργασία, αποτελεί έναν πολύ ανταγωνιστικό αλγόριθμο με πολλές εφαρμογές ακόμα και στην αυτόματη ιατρική διάγνωση .

3.1.2 Σύντομη Αναφορά στην Χρησιμότητα της Προεπεξεργασίας Δεδομένων

Καθαρισμός Δεδομένων – Data Cleansing

Η προεπεξεργασία Δεδομένων αποτελεί ένα απαραίτητο στάδιο στην διαδικασία της Εξόρυξης Δεδομένων. Η αναγκαιότητα είναι εμφανής καθώς είναι πολύ συνηθισμένο ένα σύνολο δεδομένων να περιέχει διπλοκαταχωρημένες εγγραφές, εγγραφές με αντικρουόμενο περιεχόμενο, τιμές που παραβιάζουν λογικούς κανόνες (πχ αρνητικές ηλικίες),εσφαλμένες τιμές, χρήση συνωνύμων τιμών σε κάποιο πεδίο (πχ η «Θεσσαλονίκη» μπορεί να καταχωρείται ολογράφως, ή ως «Θεσ/νικη») κλπ. Κάτι τέτοιο παρατηρείται συνήθως κατά την συγχώνευση δεδομένων από διαφορετικές πηγές, Συμπερασματικά, τα σύνολα δεδομένων τείνουν να μην περιέχουν άρτια πληροφορία είτε λόγω χαμένων τιμών είτε λανθασμένων.

Οι χαμένες τιμές (missing values) σχετικά με τον πελάτη μπορεί να οφείλονται στο γεγονός ότι δεν είναι διαθέσιμες τη στιγμή δημιουργίας της καρτέλας του και καταχωρούνται μόνο τα διαθέσιμα στοιχεία. Διαγραφή δεδομένων μπορεί να γίνει από ανθρώπινο χειριστικό λάθος ή και από αστοχία του εξοπλισμού. Δεν είναι σπάνιες οι περιπτώσεις, που ορισμένες πληροφορίες θεωρήθηκαν «μη σημαντικές » και δεν καταγράφηκαν από τους χειριστές των συστημάτων ή περιπτώσεις κακής συνεννόησης μεταξύ των χειριστών και των προϊσταμένων τους.

Οι λανθασμένες τιμές συχνά ορίζονται ως «θόρυβος». Τα δεδομένα, μπορεί να είναι «θορυβώδη» λόγω σφαλμάτων των χειριστών τη στιγμή της καταχώρησης των δεδομένων ή της χρήσης του λογισμικού, αλλά και εξαιτίας τεχνικών προβλημάτων, όπως είναι η κακή λειτουργία συσκευών που καταγράφουν δεδομένα (πχ συσκευών ανάγνωσης ετικετών RFID), ή/και τα πιθανά προβλήματα στη μετάδοση των δεδομένων μέσω ενός δικτύου.

Ωστόσο, υπάρχουν και άλλες περιπτώσεις όπου η προεπεξεργασία αποδεικνύεται το βασικότερο πρώτο βήμα. Η ύπαρξη δεδομένων με ακραίες τιμές είναι μία τέτοια περίπτωση. Αυτά τα δεδομένα-εξαιρέσεις δεν προσφέρουν στην ανάλυση χρήσιμη πληροφορία, καθώς περιγράφουν σπάνιες και μεμονωμένες περιπτώσεις χωρίς κάποια κανονικότητα. Αντιθέτως, μπορούν ν’ αποπροσανατολίσουν τους αλγορίθμους εξόρυξης και να τους οδηγήσουν σε εσφαλμένα ή μεροληπτικά συμπεράσματα. Ωστόσο, είναι σημαντικό ν’ αναφερθεί ότι η αξία τους είναι πολύ μεγάλη σε περιπτώσεις όπως του εντοπισμού κάποιας απάτης. Σ’ αυτές τις εργασίες εξόρυξης δεδομένων, το ενδιαφέρον επικεντρώνεται αποκλειστικά στα δεδομένα που αποκλίνουν από το «κανονικό» ή το συνηθισμένο. Σε κάθε περίπτωση όμως, τα δεδομένα με ακραίες τιμές πρέπει να τυγχάνουν ειδικό χειρισμό και αξιολόγηση από τον αναλυτή.

Μετασχηματισμός Δεδομένων – Data Transformation

Ο καθαρισμός των δεδομένων δεν είναι το μοναδικό αντικείμενο της προεπεξεργασίας. Η προσαρμογή των δεδομένων στις απαιτήσεις των μεθόδων επεξεργασίας είναι εξίσου σημαντική.

Στην περίπτωση των αλγορίθμων που δεν μπορούν να χειριστούν συνεχείς τιμές (αριθμούς), αλλά μόνο ονομαστικές, αυτό θα πρέπει να είναι εφικτό. Εάν το σύνολο δεδομένων περιλαμβάνει αριθμητικά πεδία, τότε ο αναλυτής πρέπει να εντοπίσει τα δεδομένα αυτά και να μετατρέψει τις αριθμητικές τιμές σε ονομαστικές. Η διαδικασία αυτή ονομάζεται Διακριτοποίηση (Discretization).

Σε άλλη περίπτωση, υπάρχουν μέθοδοι που αντιμετωπίζουν προβλήματα εάν ορισμένες μεταβλητές περιέχουν πολύ μεγάλες τιμές, ενώ άλλες μεταβλητές περιέχουν μικρές τιμές. Κατά την χρήση του αλγορίθμου k-Πλησιέστεροι Γείτονες, η ύπαρξη τέτοιων συνθηκών μπορεί να λειτουργήσει αποπροσανατολιστικά καθώς οι μεταβλητές με τις μεγάλες τιμές θα καθορίσουν το αποτέλεσμα, ενώ η επιρροή των μεταβλητών με τις μικρές τιμές θα είναι ανύπαρκτη. Επίσης, χρήσιμο χαρακτηριστικό για τα Νευρωνικά Δίκτυα είναι ό,τι λειτουργούν καλύτερα όταν οι τιμές που επεξεργάζονται κυμαίνονται στην περιοχή [0..1]. Σε αυτές τις περιπτώσεις λοιπόν, πρέπει να γίνει αναγωγή των αριθμητικών τιμών σε άλλες αριθμητικές τιμές, που να κυμαίνονται εντός των ορίων της επιθυμητής περιοχής. Η διαδικασία αυτή ονομάζεται Κανονικοποίηση (Normalization). Τέλος, τα δεδομένα μπορούν να τροποποιηθούν έτσι ώστε να εκφράζουν γενικεύσεις, πχ συγκεντρωτικές πωλήσεις ανά περιοχή. Εργασίες μετατροπής των δεδομένων όπως η κανονικοποίηση και η γενίκευση ονομάζονται μετασχηματισμός των δεδομένων (data transformation) και αποτελούν μια ακόμα μορφή προεπεξεργασίας των δεδομένων.

Ένα άλλο σύνολο εργασιών, που εκτελούνται στα πλαίσια της προεπεξεργασίας, αφορούν τη μείωση των δεδομένων. Η ικανότητα χειρισμού μεγάλου όγκου δεδομένων είναι ένα από τα βασικά χαρακτηριστικά της Εξόρυξης Δεδομένων. Όμως δεδομένα μεγάλου όγκου μπορούν να προκαλέσουν προβλήματα στις μεθόδους επεξεργασίας και μεγάλες καθυστερήσεις στη διεξαγωγή των αναλύσεων. Για τον λόγο αυτό εφαρμόζονται διαδικασίες για την μείωση του όγκου των δεδομένων. Η μείωση του όγκου δεν είναι μια τετριμμένη εργασία, καθώς τα αποτελέσματα της ανάλυσης των μειωμένων δεδομένων πρέπει να είναι τα ίδια ή περίπου τα ίδια με τα αποτελέσματα της ανάλυσης του συνόλου των δεδομένων. Μίας μέθοδος επίτευξης αυτού, είναι η επιλογή και διατήρηση των σημαντικών χαρακτηριστικών (feature selection). Τα σύνολα δεδομένων περιλαμβάνουν πολλά χαρακτηριστικά (στήλες) που για κάποια εργασία εξόρυξης, δεν είναι απαραίτητα χρήσιμα όλα. Πληροφορίες που δεν σχετίζονται με το αντικείμενο της ανάλυσης μπορούν να παραλειφθούν. Επίσης, στις περιπτώσεις όπου υπάρχουν χαρακτηριστικά που είναι διαφορετικές εκδοχές της ίδιας πληροφορίας χρειάζεται διερεύνηση ποια μεταβλητή είναι η πλέον κατάλληλη για την αναλυτική εργασία που διεξάγεται. Η διαδικασία κατάληξης σ’ εκείνο το υποσυνόλο των χαρακτηριστικών που είναι το πλέον κατάλληλο για μια συγκεκριμένη εργασία εξόρυξης γνώσης είναι πολύ σημαντική και γνωστή ως επιλογή χαρακτηριστικών.

3.2 Προεπεξεργασία Δεδομένων – Data Preprocessing

3.2.1 Περιγραφή -Συνόλου Δεδομένων

Για την παρούσα εργασία, το σύνολο δεδομένων που χρησιμοποιήθηκε κατά την υλοποίηση είναι ένα συνθετικό σύνολο δεδομένων από προσωπική πηγή. Η επιλογή αυτή προκύπτει σεβόμενη την Προστασία Δεδομένων (GDPR) και επιδιώκοντας την αποφυγή εργασίας σε όμοιο σύνολο δεδομένων από τα ευρέως ελεύθερα στο Διαδίκτυο.

Το συγκεκριμένο αρχείο είναι περίπου 1 MB και περιέχει δεδομένα αξιοποιήσιμα με σκοπό τον Μάρκετινγκ. Πιο αναλυτικά, πρόκειται για δεδομένα πελατών και των αγορών τους σε κάποιο κατάστημα ηλεκτρονικών ειδών κατά το έτος 2001. Περιλαμβάνει περίπου 350,000 συναλλαγές και πάνω από 59,000 πελάτες.

Τα δεδομένα πελατών είναι δημογραφικά (φύλο, ηλικία, οικογενειακή κατάσταση κ.α) και γεωγραφικά (περιοχή, ταχυδρομικός κώδικας κ.α) ενώ τα δεδομένα των αγορών αφορούν στο είδος της αγοράς, ποσό, ημερομηνία κ.α. Για πληρέστερη εικόνα ακολουθεί ο αναλυτικός πίνακας των στηλών δεδομένων.

Πίνακας 3.1: Χαρακτηριστικά – Γνωρίσματα του Συνόλου Δεδομένων

Data columns (total 23 columns)	
CardID	Αριθμός Κάρτας Πελάτη Πόντων
PostalCode	Ταχυδρομικός Κώδικας Περιοχής Διαμονής
Region	Τοποθεσία Διαμονής
CardStartDate	Ημερομηνία Απόκτησης Κάρτα Πελάτη
Gender	Φύλο
DateOfBirth	Ημερομηνίας Γέννησης
MaritalStatus	Οικογενειακή Κατάσταση
HasChildren	Ύπαρξη Τέκνων
NumChildren	Αριθμός Τέκνων
YoungestChild	Ηλικία Νεότερου Τέκνου
Population	Πληθυσμός Περιοχής Διαμονής
HouseHolds	Αριθμός Νοικοκυριών Περιοχής Διαμονής
ItemCode	Κωδικός Προϊόντος
TransactionID	Αριθμός Συναλλαγής
TimeStamp	Χρονικό Αποτύπωμα Συναλλαγής
PaymentMethod	Τρόπος Πληρωμής
ItemNumber	Αριθμός Τεμαχίων
Amount	Τιμή Προϊόντος
ItemDescription	Περιγραφή Προϊόντος
CategoryDescription	Κατηγορία Προϊόντος
SubCategoryDescription	Υποκατηγορία Προϊόντος
BrandDescription	Μάρκα Προϊόντος
Date	Ημερομηνία Αγοράς

3.2.2 Προεπεξεργασία Δεδομένων

Μετά την απαραίτητη εισαγωγή των βιβλιοθηκών όπως παρουσιάστηκαν στην εισαγωγή του κεφαλαίου, η διαδικασία συνεχίζεται με την προεπεξεργασία των δεδομένων ανά στάδιο.

Όπως αναφέρθηκε το αρχείο περιέχει 351,233 στιγμιότυπα αγορών από 59,477 πελάτες. Παρακάτω παρουσιάζεται πώς η σύνθεση και το μέγεθος του αρχείου μπορεί να μετασχηματιστεί για να εξυπηρετήσει τον σκοπό της παρούσας εργασίας.

Αρχική Εκτίμηση του Συνόλου Δεδομένων

Πρωταρχικό βήμα της όλης διαδικασίας είναι η εκτίμηση του συνόλου δεδομένων προκειμένου να φανούν οι ανάγκες του για προεπεξεργασία. Γι' αυτόν τον λόγο γίνεται χρήση των εντολών .info() και .describe().

Με την εντολή .info() δίνεται ο τύπος και ο αριθμός των στιγμιότυπων ανά στήλη χαρακτηριστικών (attributes). Αντίστοιχα με την εντολή .describe(), εξάγεται όλη η πληροφορία για τ' αριθμητικά δεδομένα που αφορά τον μέσο, την τυπική απόκλιση, το min-max κάθε χαρακτηριστικού και τα εκατοστημόρια.

Code Output 1: Παρουσίαση Πληροφοριών του Συνόλου Δεδομένων – Εντολή - .info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 351233 entries, 0 to 351232
Data columns (total 23 columns):
CardID                    351233 non-null object
PostalCode                351233 non-null object
Region                    351233 non-null object
CardStartDate             351233 non-null object
Gender                    351233 non-null object
DateOfBirth               351233 non-null object
MaritalStatus             233099 non-null object
HasChildren               351233 non-null object
NumChildren               351233 non-null int64
YoungestChild             351233 non-null int64
Population                351233 non-null int64
HouseHolds                351233 non-null int64
ItemCode                  351233 non-null object
TransactionID             351233 non-null object
TimeStamp                 350241 non-null object
PaymentMethod             351233 non-null object
ItemNumber                351233 non-null int64
Amount                    351233 non-null float64
ItemDescription           351233 non-null object
CategoryDescription       351233 non-null object
SubCategoryDescription    351233 non-null object
BrandDescription          351233 non-null object
Date                      350241 non-null datetime64[ns]
dtypes: datetime64[ns](1), float64(1), int64(5), object(16)
memory usage: 61.6+ MB
```

Code Output 2: Περιγραφή του Συνόλου Δεδομένων - Εντολή .describe()

	NumChildren	YoungestChild	Population	HouseHolds	ItemNumber	Amount
count	351233.000000	351233.000000	351233.000000	351233.000000	351233.000000	351233.000000
mean	1.153277	11.143591	631.820259	268.850635	1.071548	65.605783
std	1.435603	10.320331	154.545312	65.769422	0.268653	118.865541
min	-1.000000	-5.000000	161.000000	69.000000	1.000000	1.490000
25%	0.000000	2.000000	504.000000	214.000000	1.000000	4.990000
50%	1.000000	9.000000	630.000000	271.000000	1.000000	14.990000
75%	2.000000	17.000000	740.000000	314.000000	1.000000	79.000000
max	7.000000	73.000000	1175.000000	480.000000	5.000000	2249.000000

Διαδικασία Καθαρισμού των Δεδομένων

Ο καθαρισμός κρίνεται κυρίως από τις χαμένες και τις λανθασμένες τιμές. Βάσει των παραπάνω αποτελεσμάτων, παρατηρείται ότι στα πεδία Marital Status, TimeStamp και Date τα στιγμιότυπα είναι λιγότερα από το γενικότερο σύνολο. Επίσης στα δεύτερα αποτελέσματα, παρουσιάζεται αρνητικός αριθμός στον αριθμό των τέκνων καθώς και στην ηλικία του νεότερου παιδιού γεγονός το οποίο αποτελεί λογικά λανθασμένη τιμή.

Ένας γενικός κανόνας διαχείρισης είναι η παράλειψη των στιγμιότυπων με την λανθασμένη ή ελλιπή πληροφορία. Ωστόσο επειδή ταυτόχρονα είναι σημαντικό να μην παραλείπονται στοιχεία που μπορούν να συμβάλλουν στ΄ αποτελέσματα κρίνεται σκόπιμο να γίνει μία πρώτη εκτίμηση της αναγκαιότητας των στηλών χαρακτηριστικών με την μη άρτια πληροφορία. Αν για παράδειγμα στη συνέχεια αποκλειστούν οι συγκεκριμένες στήλες ως περιττές, δεν υπάρχει λόγος να χαθεί όλο το στιγμιότυπο.

Συγκεκριμένα εδώ, από τις στήλες με τα θορυβώδη δεδομένα, για τον σκοπό της εργασίας κρίνονται απαραίτητες οι στήλες Date και εν δυνάμει Timestamp. Επομένως, αυτόματα δόθηκε εντολή να παραλειφθούν όλα τα στιγμιότυπα με χαμένες τιμές σε αυτά τα πεδία, μειώνοντας έτσι τον αριθμό των στιγμιότυπων από 351.233 σε 350.241.

Διαδικασία Επιλογής Κύριων Χαρακτηριστικών του συνόλου Δεδομένων

Όπως παρουσιάστηκε στην εισαγωγή, η μείωση του συνολικού όγκου δεδομένων είναι ιδιαίτερα βοηθητική για τους αλγόριθμους που θα εφαρμοστούν καθώς διευκολύνει και επιταχύνει τον χρόνο εκτέλεσης.

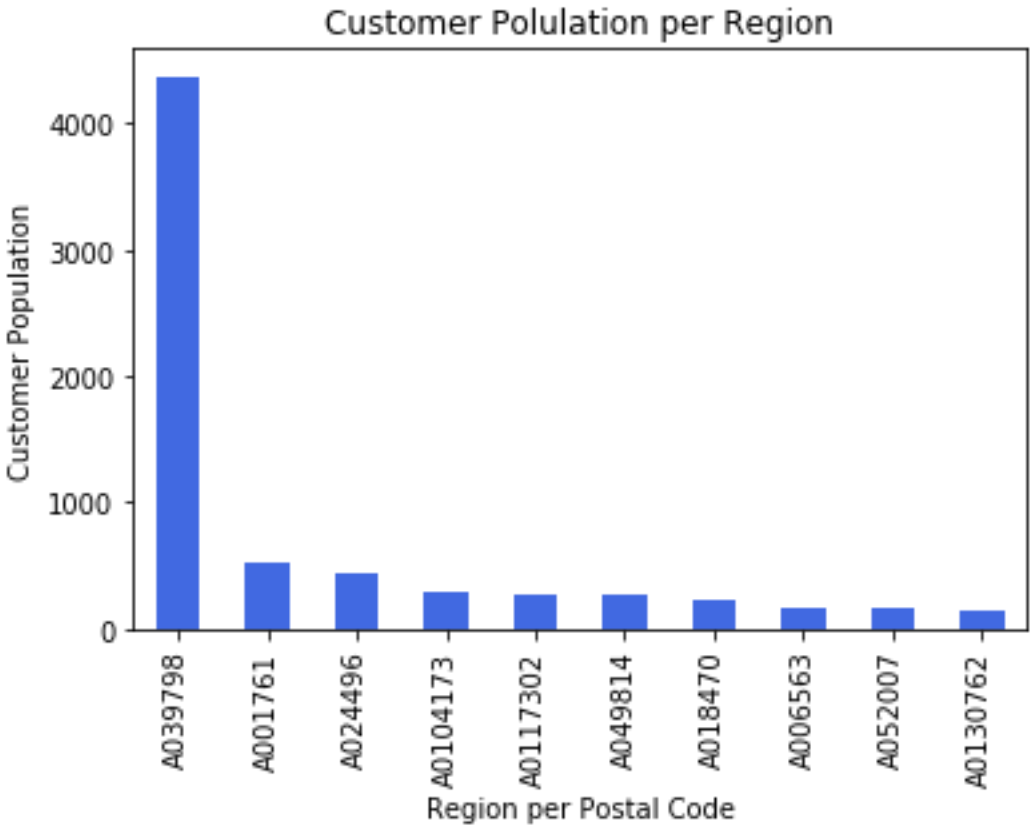
Στο παρών σύνολο δεδομένων, εκτιμήθηκε ότι θα ήταν αποδοτικότερο η Τμηματοποίηση πελατών να εστιαστεί στην περιοχή – συνοικία με το μεγαλύτερο πλήθος καταναλωτών. Σε αυτήν την περίπτωση η αγοραστική συμπεριφορά είναι πιο εύκολο να συγκριθεί και να ερμηνευθεί καθώς υπάρχει μία κοινή παράμετρος, της περιοχής κατοικίας. Η περιοχή κατοικίας πολλές φορές είναι ενδεικτική της οικονομικής δυνατότητας, επομένως δεν μπορεί

ν’ αναμένεται η ίδια συχνότητα και τα ίδια ποσά αγορών από πελάτες μίας περιοχής αποτελούμενη από χαμηλότερα οικονομικά στρώματα απ’ ότι περιοχές με υψηλότερα.

Για την εύρεση της περιοχής με την υψηλότερη αγοραστική κίνηση, τα δεδομένα περιορίστηκαν στις στήλες Card ID και Postal Code. Στην συνέχεια με την εντολή .drop_duplicates() παρέμειναν στις στήλες δεδομένων τα μοναδικά ζεύγη χαρακτηριστικών. Με αυτόν τον τρόπο τα filtered_data περιέχουν όλους τους πελάτες του καταστήματος ηλεκτρονικών ειδών και των περιοχών που αυτοί διαμένουν βάσει του Ταχυδρομικού Κώδικα.

Ο συνολικός αριθμός των περιοχών είναι 32.339. Ανάμεσα σε αυτές θα επιλεγεί η περιοχή με την υψηλότερη αγοραστική κίνηση με την βοήθεια του παρακάτω γραφήματος.

Code Output 3: Εύρεση των 10 Περιοχών με τους περισσότερους Πελάτες.



Το παραπάνω γράφημα περιλαμβάνει τις 10 πρώτες περιοχές με το μεγαλύτερο αριθμητικά αγοραστικό κοινό του καταστήματος.

Η πληροφορία αυτή θ’ αξιοποιηθεί για την εκ -νέου μείωση των χαρακτηριστικών του συνόλου δεδομένων. Το νέο σύνολο δεδομένων που ορίστηκε περιλαμβάνει τις κινήσεις αγορών των πελατών που δηλώνουν ως κατοικία την παραπάνω πρώτη σε σειρά. Μετά τον επαναπροσδιορισμό ο αριθμός των στιγμιότυπων του συνόλου έχει φτάσει πλέον τις 28.207 συναλλαγές και οι πελάτες τους 4.373.

Τέλος απομένει το σύνολο δεδομένων να περιοριστεί στ' απαραίτητα χαρακτηριστικά που εξυπηρετούν τον σκοπό της εργασίας. Τα χαρακτηριστικά αυτά είναι τα εξής:

- **Card ID** – Το μοναδικό ID του πελάτη που καθορίζεται από την Κάρτα Πόντων του
- **Transaction ID** – Ο μοναδικός αριθμός της συναλλαγής
- **Item Number** - Ο αριθμός των τεμαχίων του προϊόντος που αγοράστηκε.
- **Amount** – Το ποσό που ξοδεύτηκε ανά συναλλαγή
- **Date** – Η ημερομηνία πραγματοποίησης της συναλλαγής.

Συνοψίζοντας τις διαδικασίες προεπεξεργασίας που προηγήθηκαν, θα ήταν χρήσιμο να εκτιμηθεί η συμβολή του στην διαμόρφωση ενός καταλληλότερου συνόλου δεδομένων.

Πίνακας 3.2: Αριθμητική Εκτίμηση του Συνόλου Δεδομένων μετά την Προεπεξεργασία

	Αριθμός Συναλλαγών	**Μοναδικός Αριθμός Πελατών**	**Αριθμός Στηλών Χαρακτηριστικών**
Αρχικό Σύνολο Δεδομένων	351.233	59.477	23
Σύνολο μετά την Διαδικασία Καθαρισμού	350.241	59.477	23
Σύνολο μετά την Διαδικασία Επιλογής Χαρακτηριστικών	28.207	4.373	5

Το τελικό σύνολο δεδομένων που θα χρησιμοποιηθεί στην παρούσα εργασία αφορά το 8% των συνολικών συναλλαγών από το 7,5% των συνολικών πελατών του καταστήματος.

3.3 Εφαρμογή RFM Μοντέλου για Τμηματοποίηση Πελατών

Το RFM μοντέλο όπως αναλύθηκε εκτενώς βασίζεται στις 3 μετρικές Recency, Frequency και Monetary. Το βασικό βήμα για την εφαρμογή του μοντέλου είναι να υπολογιστούν οι 3 αυτές τιμές για τον κάθε πελάτη ξεχωριστά.

Με βάση τα υπάρχοντα δεδομένα, υπάρχουν κάποια προαπαιτούμενα βήματα προκειμένου να εφαρμοστούν οι μετρικές εξισώσεις.

Αρχικά χρειάζεται να υπολογιστεί το συνολικό χρηματικό ποσό αγοράς ανά συναλλαγή, πολλαπλασιάζοντας την τιμή Item Number με την τιμή Amount.

```
Data['Total Amount']=Data['ItemNumber'] * Data['Amount']
```

Στην συνέχεια, χρειάζεται να εκτιμηθεί το χρονικό διάστημα που καλύπτουν τα δεδομένα και να οριστεί η ημερομηνία βάσει της οποίας θα γίνει η αξιολόγηση της προσωρινής αγοράς. Στο συγκεκριμένο σημείο χρειάζεται να διευκρινιστεί ότι καθώς τα δεδομένα καλύπτουν το έτος 2001, σαν ημερομηνία πραγματοποίησης της διερεύνησης θα οριστεί η 1η Ιανουαρίου 2002.

```
Data['Date'].min(), Data['Date'].max()
PRESENT = dt.datetime(2002,1,1)
```

Πλέον τα δεδομένα έχουν όλη την απαραίτητη πληροφορία για να εφαρμοστεί ο παρακάτω κώδικας που θα δώσει τις απαιτούμενες τιμές.

```
rfm= Data.groupby('CardID').agg({'Date': lambda date: (PRESENT - date.max()).days,
                                 'TransactionID': lambda num: len(num),
                                 'Total Amount': lambda price: price.sum()})
```

Συγκεκριμένα εδώ, ανά πελάτη υπολογίστηκε

- η ημερομηνία της τελευταίας αγοράς του από την ημερομηνία διερεύνησης που ορίστηκε (Recency)
- το άθροισμα των συναλλαγών του σε όλο το χρονικό διάστημα
- το συνολικό χρηματικό ποσό που διέθεσε στις αγορές του

Η πληροφορία που έχει προκύψει μέχρι στιγμής δεν είναι αξιοποιήσιμη καθώς ο συνδυασμός των παραπάνω τιμών κατά κύριο λόγο διαφέρει από πελάτη σε πελάτη ώστε να-διερευνηθούν ομοιότητες και να ανιχνευθούν συστάδες,

Σε αυτήν την περίπτωση, η λύση έρχεται από την κατάταξη των παραπάνω τιμών σε τεταρτημόρια όπου το 1ο τεταρτημόριο ορίστηκε να είναι ενδεικτικό όσων ψώνισαν πολύ πρόσφατα, οι αγορές τους είναι οι πιο συχνές και τα χρηματικά ποσά αγορών τα μεγαλύτερα ενώ το 4ο τεταρτημόριο είναι το πιο αδύναμο σε σχέση με τα παραπάνω.

```
rfm['R'] = pd.qcut(rfm['Recency'], 4, ['1','2','3','4'])
rfm['F'] = pd.qcut(rfm['Frequency'], 4, ['4','3','2','1'])
rfm['M'] = pd.qcut(rfm['Monetary'], 4, ['4','3','2','1'])
```

Σε αυτό το σημείο μπορεί ,πλέον να υπολογιστεί το RFM Score κάθε πελάτη με την ενοποίηση των τιμών που προέκυψαν.

```
rfm['RFM_Score'] = rfm.R.astype(str)+ rfm.F.astype(str) + rfm.M.astype(str)
```

Το RFM Score είναι το βασικότερο κριτήριο ανίχνευσης των ομάδων πελατών, γεγονός που αποτυπώνεται στον κώδικα που ακολουθεί.

```
segt_map = {


r'111': 'Champions',
r'[1-2][1-2][1-2]': 'High-spending Loyal Customers',
r'[1-2][1-2][3-4]': 'Low-spending Loyal Customers',
r'[1-2][3][1-4]': 'Potential Customers',
r'[1-2][4][1-4]': 'New Customers',
r'[3][1-2][1]': 'Should not Loose',
r'[3][3-4][1-4]': 'Need Attention',
r'[4][1][1]': 'Churned Best Customers',
r'[3-4][1-2][1-4]': 'Sleepers',
r'[4][3-4][1-4]': 'Lost',


}
rfm['Segment'] = rfm['R'].map(str) + rfm['F'].map(str)+rfm['M'].map(str)
rfm['Segment'] = rfm['Segment'].replace(segt_map, regex=True)
```

Εδώ είναι ένα κομμάτι που χρειάζεται αρκετή ανάλυση. Ανατρέχοντας στο θεωρητικό υπόβαθρο, είχε αναφερθεί ότι η ανίχνευση ομάδων – συστάδων μπορεί να λειτουργήσει μετέπειτα σαν κλάση – κατηγορία για την εφαρμογή άλλων μεθόδων.

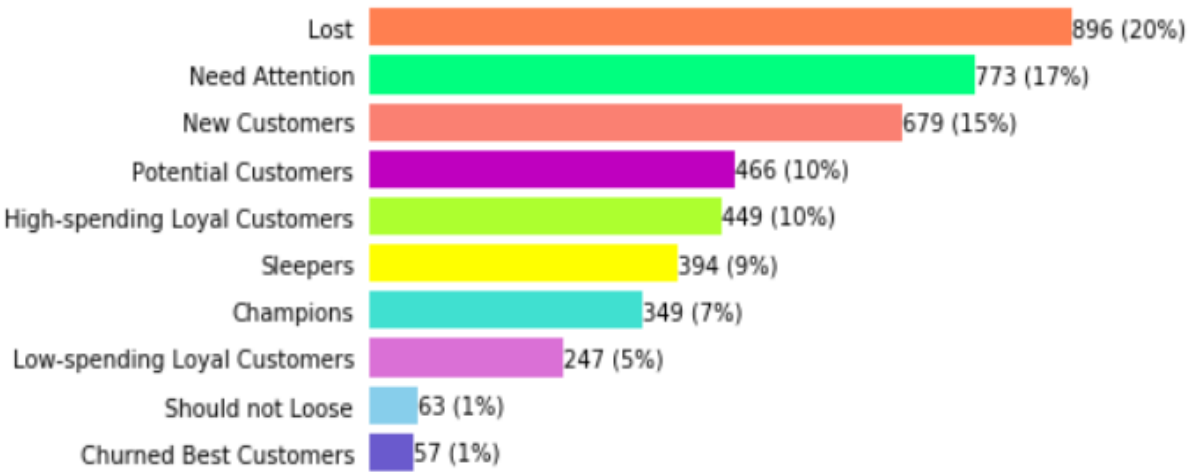
Το RFM καθώς είναι μία μέθοδος με συγκεκριμένο σκοπό, οι ομάδες που αφορούν ένα σύνολο δεδομένων πελατών είναι κατά κάποιο τρόπο ορισμένες από πριν και ο αναλυτής επιλέγει ποιες τον ενδιαφέρουν περισσότερο για διερεύνηση ανάλογα με τον σκοπό της έρευνας του. Επομένως ο παραπάνω κώδικας λειτουργεί σαν μια «κατηγοριοποίηση» αντιστοιχίζοντας κάθε συνδυασμό των τιμών R, F, M σε μία ομάδα. Οι ομάδες που επιλέχθηκαν εδώ είναι 9 και η λογική της αντιστοίχισης έχει αναλυθεί εκτενώς στο θεωρητικό υπόβαθρο.

Συγκεκριμένα για κάθε τιμή έχει δοθεί ένα εύρος τιμών που σε συνδυασμό με τις υπόλοιπες αντιστοιχίζουν κάθε πελάτη σε μία ομάδα. Τ' αποτελέσματα που προκύπτουν έχουν ως εξής.

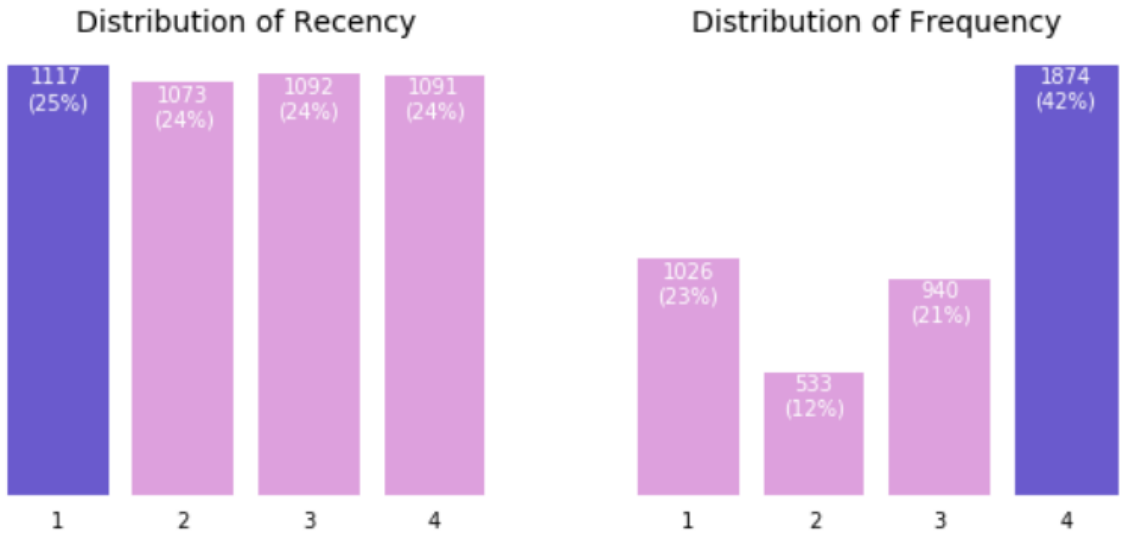
Code Output 4: Ετικέτα Ομάδας ανά Πελάτη

CardID	Recency	Frequency	Monetary	R	F	M	RFM_Score	Segment
C0100000444	3	22	847.23	1	1	1	111	Champions
C0100000789	3	2	478.00	1	4	2	142	New Customers
C0100001714	23	2	398.00	2	4	2	242	New Customers
C0100001916	115	4	356.99	3	2	2	322	Sleepers
C0100003293	35	4	341.99	2	2	2	222	High-spending Loyal Customers

Code Output 5: Γράφημα Κατάταξης Ομάδων Πελατών Συγκεκριμένης Περιοχής με Ποσοστά



Code Output 6: Κατανομή των Πελατών ανά Recency, Frequency.



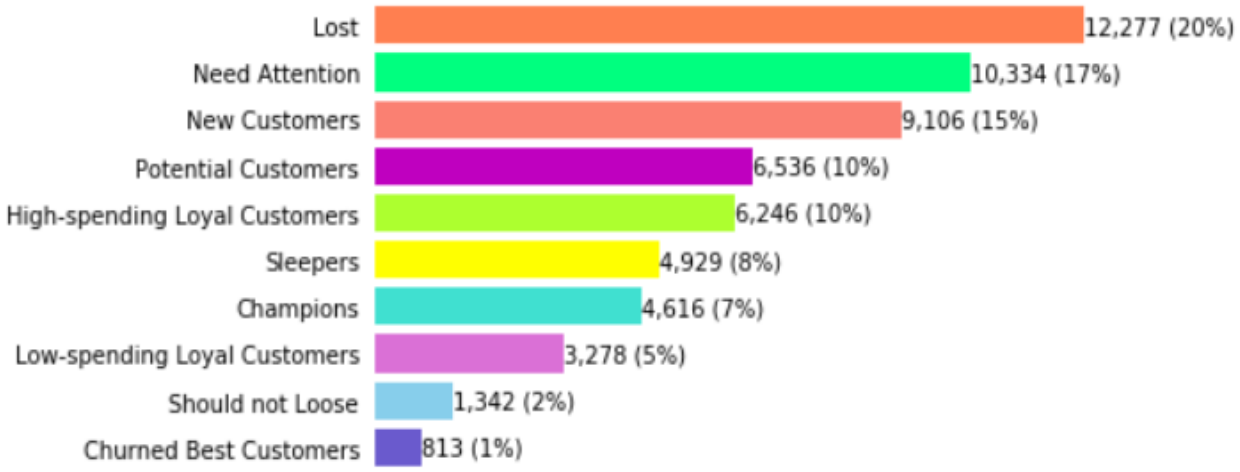
Στην συνέχεια ακολουθεί η εκτίμηση των παραπάνω μετρικών R, F, M σε σχέση με κάθε συστάδα.

Code Output 7: Οι μέσες τιμές των μετρικών R, F, M ανά Ομάδα Πελατών

	Recency mean	Frequency mean	Monetary mean	count
Segment				
Champions	9.7	27.3	1048.4	349
Churned Best Customers	200.6	17.5	912.0	57
High-spending Loyal Customers	28.2	15.1	694.4	449
Lost	231.7	1.6	274.3	896
Low-spending Loyal Customers	24.8	6.5	204.5	247
Need Attention	98.0	2.3	404.9	773
New Customers	25.0	1.7	392.2	679
Potential Customers	27.4	3.0	433.5	466
Should not Loose	92.4	9.4	945.5	63
Sleepers	125.3	7.4	340.2	394

Ακολουθεί η εφαρμογή του μοντέλου σε όλο το αρχείο ανεξαρτήτως περιοχής με τ΄ αντίστοιχα αποτελέσματα.

Code Output 8: Γράφημα Κατάταξης Ομάδων Πελατών Όλων των Περιοχών με Ποσοστά



Code Output 9: Οι μέσες τιμές των μετρικών R, F, M ανά Ομάδα Πελατών

	Recency mean	Frequency mean	Monetary mean	count
Segment				
Champions	10.1	26.1	974.4	4616
Churned Best Customers	200.1	15.7	708.8	813
High-spending Loyal Customers	30.0	11.8	653.9	6246
Lost	233.0	1.7	211.8	12277
Low-spending Loyal Customers	26.5	5.8	204.6	3278
Need Attention	101.1	2.3	336.0	10334
New Customers	27.2	1.7	293.9	9106
Potential Customers	28.1	3.0	369.4	6536
Should not Loose	95.7	8.5	840.2	1342
Sleepers	131.8	6.8	326.2	4929

Παρατηρείται ότι η αρχική περιοχή που επιλέχθηκε να μελετηθεί αποτελεί μία μικρογραφία των αποτελεσμάτων για όλες τις περιοχές ενώ αποτελεί μόλις το 10% των συνολικών πελατών ώστε να είναι μεροληπτικό το αποτέλεσμα. Θεωρείται αντιπροσωπευτικό δείγμα και όποια ενέργεια γίνει σε αυτό θα μπορούσε να γενικευθεί.

3.4 Τμηματοποίηση Πελατών με βάση την τιμή Customer Lifetime Value

Στο σημείο αυτό θα υπολογιστεί η χρηματική αξία του κάθε πελάτη (Customer Value). Το χρηματικό ποσό θα αποφέρει ένας πελάτης σε βάθος χρόνου της αγοραστικής του σχέσης με την εταιρεία (Customer Lifetime Value) είναι πολύ σημαντικό μέτρο εκτίμησης και εξαρτάται άμεσα από την παρούσα χρηματική αξία ενός πελάτη. Το γεγονός αυτό καθιστά την τιμή Customer Lifetime Value ένα χαρακτηριστικό βάσει του οποίου μπορεί να γίνει Τμηματοποίηση των πελατών και να δοθεί η αντίστοιχη σημασία στην κάθε κατηγορία.

Αρχικά γίνεται ο υπολογισμός των βασικών τιμών που ακολουθούν:

Code Output 6: Βασικές τιμές για τον τελικό υπολογισμό της τιμής Customer Value

```
purchase_frequency,repeat_rate,churn_rate
```

```
(5.888679657682801, 0.8300183264118903, 0.16998167358810967)
```

Αναλυτικότερα τι σημαίνει και πως υπολογίστηκε ο κάθε τύπος.

- Το μέσο χρηματικό ποσό που ξοδεύει ένας πελάτης ανά αγορά του:

```
rfm_total['avg_order_value' ]= (rfm_total['Monetary']/rfm_total['Frequency'])
```

- Η μέση αγοραστική συχνότητα που παρατηρείται στην διάρκεια του χρόνου που εξετάζεται:

```
purchase_frequency = sum(rfm_total['Frequency'])/rfm_total
```

- Η τιμή που εκφράζει την πιθανότητα επανεπισκεψιμότητας – αγοράς ενός πελάτη:

```
repeat_rate = rfm_total[rfm_total.Frequency > 1].shape[0]/rfm_total.shape[0]
```

- Η τιμή που εκφράζει την πιθανότητα μεταστροφής ενός πελάτη:

```
churn_rate = 1 - repeat_rate
```

- Το ποσοτικό κέρδος της εταιρίας ανά αγορά ενός πελάτη, δεδομένου ότι το ποσοστιαίο κέρδος έχει παρατηρηθεί και οριστεί από την εταιρεία ως 5%:

```
rfm_total['profit_margin']=(rfm_total['Monetary']*0.05)
```

- Υπολογισμός της χρηματικής αξίας ενός πελάτη βάσει της μέσης τιμής ανά αγορά του επί της αγοραστικής συχνότητας δια της πιθανότητας μεταστροφής του

```
rfm_total['CLV']=((rfm_total['avg_order_value']*purchase_frequency)/churn_rate
```

- Υπολογισμός της προβλεπόμενης χρηματικής αξίας που θ' αποφέρει ένας πελάτης στο σύνολο σε βάθος χρόνου.

```
rfm_total['cust_lifetime_value']=(rfm_total['CLV']*rfm_total['profit_margin'])
```

Στην συνέχεια γίνεται η ανάλογη αντιστοίχιση του εκάστοτε πελάτη στις κατηγορίες

- Most – Profitable
- High – Profitable
- Medium – Profitable
- Low – Profitable

Το συγκεκριμένο κομμάτι γίνεται με την μέθοδο των τεταρτημόριων όπως και με τις τιμές R, F, M.

Code Output 7: Κατηγοριοποίηση Πελατών βάσει Customer Lifetime Value

	CardID	RFM_Score	Segment	avg_order_value	CVL_Category
0	C0100000111	411	Churned Best Customers	147.16	Most-Profitable
6	C0100000199	131	Potential Customers	199.00	Most-Profitable
9	C0100000343	421	Sleepers	138.24	Most-Profitable
13	C0100000375	223	Low-spending Loyal Customers	56.00	Medium-Profitable
17	C0100000392	111	Champions	30.60	High-Profitable

Με μία πρώτη ματιά, παρατηρείται ότι η CVL Κατηγορία δεν έχει άμεση σύνδεση με την Κατηγορία Πελατών που ανήκει ένας πελάτης. Συγκεκριμένα θ' ανέμενε κάποιος ότι κάποιος πελάτης στην κατηγορία Champions θα άνηκε αυτόματα στην κατηγορία Most-Profitable, παρόλ΄ αυτά κάτι τέτοιο δεν ισχύει καθώς ο συγκεκριμένος πελάτης δείχνει να διαθέτει ένα σχετικά χαμηλό ποσό ανά αγορά του. Σε αυτό το σημείο η αναλυτικότερη εκτίμηση είναι ανάλογη της κατηγορίας που ανήκει η εταιρεία και ποιες είναι οι επιδιώξεις της.

3.5 Πρόβλεψη Κατηγορίας και Ενίσχυση RFM Μοντέλου με χρήση WEKA

Στο σημείο αυτό θα γίνει χρήση αλγορίθμων κατηγοριοποίησης με σκοπό την πρόβλεψη των κατηγοριών πελατών και CVL όπως αυτές παρουσιάστηκαν παραπάνω.

Η εκτέλεση της κατηγοριοποίησης με θα γίνει με χρήση του ελεύθερου λογισμικού Μηχανικής Μάθησης WEKA και συγκεκριμένα των αλγορίθμων J48 και Naïve Bayes.

Weka Output 1: Confusion Matrix J48 - Customer Segment Classification με βάση τις τιμές R, F, M

```
== Confusion Matrix ===

   a     b     c     d     e     f     g     h     i     j   <-- classified as
 749     0     0     0     0     0     0     0     0     0 |     a = Churned Best Customers
   0  5896     0     0     0     0     0     0     0     0 |     b = Potential Customers
   0     0  4435     0     0     0     0     0     0     0 |     c = Sleepers
   0     0     0  2933     0     0     0     0     0     0 |     d = Low-spending Loyal Customers
   0     0     0     0  4156     0     0     0     0     0 |     e = Champions
   0     0     0     0     0  9312     0     0     0     0 |     f = Need Attention
   0     0     0     0     0     0  5593     0     0     0 |     g = High-spending Loyal Customers
   0     0     0     0     0     0     0 11019     0     0 |     h = Lost
   0     0     0     0     0     0     0     0  8219     0 |     i = New Customers
   0     0     0     0     0     0     0     0     0  1217 |     j = Should not Loose
```

Ο διαχωρισμός στις ομάδες πελατών έγινε βάσει των τιμών R, F, M κάτι που διαπιστώνεται άμεσα με την χρήση WEKA όπου η χρήση των συγκεκριμένων τιμών μπορεί να προβλέψει την ομάδα πελατών με 100% επιτυχία.

Στην συνέχεια, εξετάζεται η πρόβλεψη της ομάδας πελατών με βάση τα δημογραφικά χαρακτηριστικά – attributes; Φύλο, Ηλικία, Οικογενειακή, Ύπαρξη Παιδιών.

Weka Output 2: Summary J48 - Customer Segment Classification με βάση τα Δημογραφικά Χαρακτηριστικά

```
Time taken to build model: 0.71 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       12030               20.2263 %
Incorrectly Classified Instances     47447               79.7737 %
Kappa statistic                         -0
Mean absolute error                      0.1728
Root mean squared error                  0.2947
Relative absolute error                100.0009 %
Root relative squared error            100.2695 %
Total Number of Instances            59477
```

Όπως είναι εμφανές δεν μπορεί ν ανιχνευθεί κάποια συσχέτιση μεταξύ των Δημογραφικών Χαρακτηριστικών και της κατηγορίας Πελατών. Το ποσοστό επιτυχίας είναι πολύ χαμηλό προκειμένου να επιδιωχθεί βελτιστοποίηση. Το γεγονός αυτό οδηγεί στο συμπέρασμα ότι τα δημογραφικά χαρακτηριστικά δεν είναι ικανά για την πρόβλεψη της κατηγορίας. Ίσως η πληροφορία για το εισόδημα να συνείσφερε αλλά όχι καθοριστικά. Αυτό ισχυροποιεί την δύναμη των R, F, M Δεικτών.

Ωστόσο το γεγονός ότι οι ομάδες των πελατών είναι 9 να θεωρείται ένας αρκετά μεγάλος αριθμός κατηγοριών ώστε ν' ανιχνευθεί κάποιο μοτίβο. Για το λόγο αυτό η πρόβλεψη της CVL Κατηγορίας ,που είναι συνολικά 4 να είναι πιο αποδοτική με βάση τα Δημογραφικά Χαρακτηριστικά.

Weka Output 3: Summary J48 - CVL Segment Classification με βάση τα Δημογραφικά Χαρακτηριστικά

```
=== Summary ===

Correctly Classified Instances       16062               27.0054 %
Incorrectly Classified Instances     43415               72.9946 %
Kappa statistic                          0.0267
Mean absolute error                      0.3743
Root mean squared error                  0.433
Relative absolute error                 99.8079 %
Root relative squared error             99.9945 %
Total Number of Instances            59477
```

Παρατηρείται πώς η κατηγοριοποίηση στις 4 Κατηγορίες που αντιπροσωπεύουν τη χρηματική αξία μιας αγοραστικής σχέσης σε βάθος χρόνου, έχει μικρό βαθμό εξάρτησης από τα δημογραφικά χαρακτηριστικά.

Αντίστοιχα θα γίνει το ίδιο πείραμα με βάσει τις τιμές R, F, M ως ένδειξη 1, 2, 3, 4 καθώς και των πραγματικών τιμών Recency, Frequency, Monetary που εκφράζουν τις μέρες που έχουν προηγηθεί από την τελευταία αγορά, τις φορές που ο πελάτης επισκέφτηκε και αγόρασε ένα προϊόν και το συνολικό ποσό που διέθεσε στο κατάστημα στον χρόνο που μελετάται.

Πίνακας 3.3: Ποσοστό Επιτυχής CLV_Κατηγοριοποίησης ανά Αλγόριθμο με χρήση των R, F, M σε διαφορετική έκφραση

	R, F, M Indicators	R, F, M Actual
Decision Tree	70%	99%
Naive Bayes	59%	51%

Είναι εμφανές ότι σε αυτήν την περίπτωση η πρόβλεψη της κατηγορίας αποδίδει καλύτερα με χρήση του Decision Tree (J48) και βάσει των πραγματικών τιμών R, F, M και όχι των δεικτών που αντιπροσωπεύουν τα εκατοστημόρια.

Πιο ειδικά, εκτιμάται ότι ο αριθμός αγορών ενός πελάτη το τελευταίο διάστημα και το χρηματικό ποσό που διέθεσε γι' αυτές είναι τα κύρια χαρακτηριστικά που συμβάλλουν στην εύρεση της κατηγορίας χρηματικής αξίας.

Πίνακας 3.4: Ποσοστό Επιτυχής CLV_Κατηγοριοποίησης ανά συνδυασμό των actual R, F, M

	R, F, M	M	R, M	F, M
J48	99%	84.9%	83.6%	99.6%

Όπως επιβεβαιώνεται και με τα παραπάνω αποτελέσματα, το πόσο πρόσφατα έχει αγοράσει ένας πελάτης από το συγκεκριμένο κατάστημα δε συμβάλλει ιδιαίτερα στην πρόβλεψη της Customer Lifetime Value Κατηγορίας. Παρατηρείται ότι το ποσό που έχει διαθέσει ο πελάτης είναι καθοριστικό και ο συνδυασμός με τον αριθμό των φορών αγορών από το κατάστημα ενισχύει σημαντικά την απόδοση του αλγόριθμου.

Τέλος θα ήταν χρήσιμο να εξεταστεί και πειραματικά η παρατήρηση που έγινε παραπάνω και αφορά στο κατά πόσον η πληροφορία της ομάδας πελατών συμβάλλει στην πρόβλεψη της CVL κατηγορίας.

Weka Output 4: Classification CVL Κατηγορίας βάσει της Ομάδα Πελατών

Correctly Classified Instances	22160	37.2581 %
Incorrectly Classified Instances	37317	62.7419 %
Kappa statistic	0.1635	
Mean absolute error	0.3507	
Root mean squared error	0.4188	
Relative absolute error	93.5291 %	
Root relative squared error	96.7129 %	
Total Number of Instances	59477	

Παρατηρείται ότι και ο αλγόριθμος J48 και ο Naïve Bayes συμπεριφέρονται κατά τον ίδιο τρόπο. Όπως είχε διαπιστωθεί δεν υπάρχει πλήρη αντιστοιχία ανάμεσα στους καλύτερους πελάτες και τους πιο κερδοφόρους.

Κεφάλαιο 4 Συμπεράσματα - Προτάσεις

"Data beats emotions."

–Larry Page

4.1Συμπεράσματα 48
4.2Συνολική Αξιολόγηση – Προτάσεις.....49

4.1 Συμπεράσματα

4.1.1 Τμηματοποίηση Πελατών βάσει της Αγοραστικής Συμπεριφοράς με εφαρμογή του RFM Μοντέλου και της Μηχανικής Μάθησης

Το RFM μοντέλο, όπως εφαρμόστηκε στην παρούσα εργασία, αποδείχθηκε πολύ ευέλικτο. Το γεγονός ότι παρέχει την δυνατότητα να οριστούν εξ' αρχής οι ομάδες πελατών που επιθυμεί ο αναλυτής να ανιχνευθούν, έχει το προτέρημα να προσαρμοστεί το μοντέλο στα μέτρα της εκάστοτε έρευνας. Συγκεκριμένα εδώ επιλέχθηκαν 9 κατηγορίες καλύπτοντας τη μεγαλύτερη γκάμα αγοραστικών συμπεριφορών.

Καθώς η τμηματοποίηση των πελατών αντικατοπτρίζει την αγοραστική συμπεριφορά των πελατών, οι δείκτες R, F, M αποτελούν το πιο αντιπροσωπευτικό χαρακτηριστικό χάρη στο οποίο μπορεί να προβλεφθεί η κατηγορία. Εδώ, η αξιοποίησή τους στο πεδίο της Μηχανικής Μάθησης με χρήση του WEKA ήταν καθοριστική.

Πιο αναλυτικά, η ανίχνευση της κάθε ομάδας πελατών που έγινε στο πρώτο μέρος αποτέλεσε ετικέτα για την μετέπειτα εφαρμογή αλγορίθμων κατηγοριοποίησης. Αποδείχθηκε ότι οι τιμές R, F, M αποτελούν τα κυριότερα χαρακτηριστικά για την πρόβλεψη της κατηγορίας σε αντίθεση με τα δημογραφικά χαρακτηριστικά, για την ώρα.

Το ουσιαστικό κέρδος από αυτή την εφαρμογή είναι το γεγονός ότι μπορεί να γενικευθεί στα περισσότερα σύνολα δεδομένων. Για παράδειγμα εάν η εκπαίδευση του αλγορίθμου πραγματοποιηθεί στο παρών σύνολο δεδομένων αλλά η επαλήθευση του γίνει σε κάποιο άλλο σύνολο δεδομένων που φέρει τους αντίστοιχους δείκτες του R, F, M η πρόβλεψη με Decision Tree έχει 100% επιτυχία. Το παραπάνω πείραμα επιβεβαιώθηκε, ανάμεσα στο σύνολο δεδομένων της παρούσας εργασίας και το συνόλο δεδομένων Online Retail από τον ιστότοπο UCI.

Η επιτυχία αυτή οφείλεται στο γεγονός ότι οι τιμές R,F,M είναι δείκτες και όχι οι πραγματικές τιμές που αφορούν το συγκεκριμένο σύνολο και πελάτη. Υπενθυμίζεται ότι οι τιμές που αφορούσαν στα χαρακτηριστικά Recency, Frequency και Monetary αντιστοιχήθηκαν σ΄ έναν αριθμό που αντικατοπτρίζει το τεταρτημόριο που ανήκουν. Επομένως ο αλγόριθμος έψαξε και αναγνώρισε μοτίβο ανάμεσα στους αριθμούς αυτούς και τις κατηγορίες αντίστοιχα. Οι πραγματικές τιμές είχαν εξίσου επιτυχία αλλά η εκπαίδευση βασιζόμενη σε αυτές δεν θα μπορούσε να εφαρμοστεί σε κάποιο άλλο σύνολο. Επιπρόσθετα αξίζει να σημειωθεί ότι 9 κατηγορίες είναι ένα μεγάλος αριθμός για κατηγοριοποίηση, παρόλα αυτά η αποτελεσματικότητα δεν επηρεάστηκε.

Τα παραπάνω αποτελέσματα ενισχύουν την απόδοση του RFM μοντέλου και το καθιστούν ένα πολύ δυνατό εργαλείο για Τμηματοποίηση πελατών. Τέλος η αξιοποίηση των

δεικτών R, F, M του μοντέλου συμβάλλει καθοριστικά στην πρόβλεψη της κατηγορίας μέσω της Μηχανικής Μάθησης.

4.1.2 Τμηματοποίηση Πελατών βάσει της Customer Lifetime Value και πρόβλεψή της με το RFM Μοντέλο μέσω της Μηχανικής Μάθησης

Στο μέρος που αφορούσε την χρηματική αξία ενός πελάτη σε βάθος χρόνου, αυτή υπολογίστηκε και διαχωρίστηκε σε 4 κατηγορίες ανάλογα με το πόσο κερδοφόρος θα μπορούσε ν' αποδειχθεί ένας πελάτης. Στην ουσία, όπως στις ομάδες πελατών, δημιουργήθηκαν με ανάλογο τρόπο οι ετικέτες – labels για την μετέπειτα εφαρμογή των αλγορίθμων κατηγοριοποίησης στο WEKA.

Τα πειράματα που εφαρμόστηκαν είχαν σκοπό να ανιχνεύσουν συσχετίσεις ανάμεσα στην Customer Lifetime Value και τα δημογραφικά χαρακτηριστικά του συνόλου δεδομένων, τις ομάδες πελατών της προηγηθείσας τμηματοποίησης και των τιμών R, F, M.

Παρατηρήθηκε ότι η κατηγοριοποίηση είχε σημαντική απόδοση βασιζόμενη στις τιμές με τις οποίες λειτουργεί και το RFM μοντέλο. Είναι σημαντικό ν' αναφερθεί ότι στην περίπτωση της Customer Lifetime Value Κατηγορίας δεν υπάρχει άμεση σύνδεση των τιμών στον υπολογισμό της, παρόλα αυτά οι τιμές Frequency και Monetary αξιοποιούμενες σε αλγόριθμο Μηχανικής Μάθησης υποδεικνύουν πως υπάρχει σχετικότητα.

Στο σημείο αυτό θα πρέπει να διαχωριστούν τ' αποτελέσματα αυτών των πειραμάτων με όσα προηγήθηκαν. Η χρήση των αλγορίθμων είχε την απαιτούμενη απόδοση βασιζόμενη στις πραγματικές τιμές Frequency και Monetary και όχι στις τιμές δείκτες 1,2,3,4. Το γεγονός αυτό διαφοροποιεί σημαντικά το συγκεκριμένο πείραμα καθώς η εκπαίδευση του αλγόριθμου μ' ένα αρχείο δεν μπορεί να λειτουργήσει μ' επιτυχία σ΄ ένα δεύτερο αρχείο. Θα πρέπει αντίστοιχα και στο δεύτερο αρχείο να υπάρξει training και testing. Αυτό συμβαίνει γιατί οι τιμές Frequency και Monetary είναι ορισμένες σχετικά με το ποια είναι η συχνότητα και το χρηματικό άθροισμα για τα δεδομένα μίας εταιρείας στο μελετούμενο χρονικό διάστημα.

Συνολικά κρίνεται ότι η χρήση του RFM μοντέλου έχει ουσιαστική εφαρμογή και σε προβλήματα κατηγοριοποίησης βάσει της Customer Lifetime Value. Είναι πολύ κερδοφόρο όταν ένα μοντέλο μπορεί ταυτόχρονα να λειτουργήσει σε δύο ειδών κατηγοριοποιήσεις και τ΄ αποτελέσματα αυτά να συνδυαστούν και ν' αξιοποιηθούν ανάλογα.

4.2 Συνολική Αξιολόγηση – Προτάσεις

Οι εφαρμογές και τα πειράματα που παρουσιάστηκαν και έγιναν στα πλαίσια της παρούσας μεταπτυχιακής εργασίας είχαν εξαιρετικό ενδιαφέρον. Οι συνιστώσες του RFM μοντέλου, κύριο εργαλείο για την τμηματοποίηση πελατών με βάση την αγοραστική συμπεριφορά, είχαν μία πολύ αποδοτική συμπεριφορά στο πεδίο της Μηχανικής Μάθησης. Είναι

σημαντικό, κλασσικά μοντέλα που χρησιμοποιούνταν στο παρελθόν να αναβαθμίζονται μέσα από την αξιοποίησή τους στην Μηχανική Μάθηση ενισχύοντας με αυτόν τον τρόπο τ΄ αποτελέσματα.

Η κατηγοριοποίηση έγινε σε δύο επίπεδα, προβλέποντας όχι μόνο την αγοραστική συμπεριφορά αλλά και την εκτίμηση της χρηματικής προσφοράς που θα έχει ένας πελάτης σε μία εταιρεία σε βάθος χρόνου. Όσο περισσότερες είναι οι πληροφορίες που εξάγει ο αναλυτής από τα δεδομένα του, τόσο πιο κερδοφόρο είναι αυτό για την έρευνα. Συγκεκριμένα τ' αποτελέσματα που προκύπτουν, όπως αυτά παρουσιάστηκαν είναι ικανά να μεταφέρουν γνώση στο τμήμα Μάρκετινγκ προς δύο κατευθύνσεις. Για παράδειγμα η ταυτόχρονη πληροφορία ότι ένας Champion Πελάτης ανήκει ταυτόχρονα στους Most – Profitable έναντι άλλου Champion Πελάτη με High Profitability, θα μπορούσε να οδηγήσει την εταιρεία σε άλλο τρόπο διαχείρισης του καθενός.

Η Μηχανική Μάθηση είναι ένας τομέας που δίνει απαντήσεις βρίσκοντας συσχετίσεις. Αυτή η ιδιότητα αξιοποιήθηκε ώστε να ενισχυθεί η αξία του RFM μοντέλου. Παρόλα αυτά συσχετίσεις της τμηματοποίησης πελατών βάσει άλλων χαρακτηριστικών του συνόλου δεδομένων διερευνήθηκαν. Τα δημογραφικά χαρακτηριστικά σύμφωνα με την κοινή λογική καθορίζουν ως έναν βαθμό την αγοραστική συμπεριφορά ενός καταναλωτή. Για να μπορεί ωστόσο αυτό να ανιχνευθεί με κάποιον αλγόριθμο, φαίνεται να χρειάζεται διαφορετική διαχείριση. Αρχικά θα ήταν πολύ βοηθητική η συγκέντρωση περισσότερων πληροφοριών από της πλευρά της εταιρείας και όσο αυτό είναι επιτρεπτό. Για παράδειγμα η γνώση του επαγγέλματος ενός πελάτη θα μπορούσε σε εταιρείες που παρέχουν ένα προϊόν σχετικό με το αντικείμενό του να συμβάλλει στην καλύτερη πρόβλεψη μιας εταιρίας. Ακόμα, ένα πεδίο που θα είχε ενδιαφέρον θα ήταν η διερεύνηση πώς τα δημογραφικά χαρακτηριστικά μπορούν να επηρεάσουν τους R, F, M δείκτες και κατά επέκταση την πρόβλεψη της ομάδας πελατών.

Σ΄ έναν ωραίο επίλογο, θα ήταν χρήσιμο να αναφερθεί πως το γεγονός ότι η παρούσα εργασία αποτελεί μία μελέτη σύνδεσης ενός κλασσικού μοντέλου με τις τεχνολογίες της σημερινής εποχής, ανοίγοντας νέα πεδία έρευνας είναι ιδιαίτερα ενθαρρυντικό και αποδεικνύει πως η επιστήμη αναπαράγει επιστήμη.

Κεφάλαιο 5 Βιβλιογραφία

Aziz, A. (χ.χ.). *Customer Segmentation based on Behavioural Data in E-marketplace.*

Ayçin, E. (χ.χ.). CUSTOMER SEGMENTATION BY USING RFM MODEL AND CLUSTERING METHODS: A CASE STUDY IN RETAIL INDUSTRY. *International Journal of Contemporary Economics and Administrative Sciences*.

Counselling, S. (χ.χ.). Data mining application in retailer segmentation based on LRFM variables: case study.

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

Cheng, C., & Chen, Y. (2009). Classifying the segmentation of customer value via RFM model and RS theory. *Expert Systems with Applications, 36*(3 PART 1), 4176-4184.

Counselling, S. (χ.χ.). Data mining application in retailer segmentation based on LRFM variables: case study.

Doğan, O., Ayçin, E., & Bulut, Z. (χ.χ.). *CUSTOMER SEGMENTATION BY USING RFM MODEL AND CLUSTERING METHODS: A CASE STUDY IN RETAIL INDUSTRY.*

Fonseca, J., & Cardoso, M. (χ.χ.). Retail Clients Latent Segments. *Lecture Notes in Computer Science*, 348-358.

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., . . . Sriram, S. (2006, 11). Modeling customer lifetime value. *Journal of Service Research, 9(2)*, 139-155.

Hosseini, M., & Shabani, M. (2015, 9 1). New approach to customer segmentation based on changes in customer value. *Journal of Marketing Analytics, 3*(3), 110-121.

iaeme, i. (χ.χ.). DEVELOPING RFM MODEL FOR CUSTOMER SEGMENTATION IN RETAIL INDUSTRY. *International journal of Marketing & Human Resource Management (IJMHRM)*.

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. *Procedia Computer Science. 3*, σσ. 57-63. Elsevier.

Monalisa, S., Nadya, P., & Novita, R. (2019). Analysis for customer lifetime value categorization with RFM model. *Procedia Computer Science. 161*, σσ. 834-840. Elsevier B.V.

(χ.χ.). *Retail Customer Segmentation RFM Segmentation to drive sales and loyalty Smart Customer Segmentation.*

Otto, P., Davies, G., Chater, N., & Stott, H. (2009, 1). From spending to understanding: Analyzing customers by their spending behavior. *Journal of Retailing and Consumer Services, 16*(1), 10-18.

Safari, F., Safari, N., & Montazer, G. (2016, 6 6). Customer lifetime value determination based on RFM model. *Marketing Intelligence and Planning, 34*(4), 446-461.

Sari, J., Nugroho, L., Ferdiana, R., & Santosa, P. (2016, 10 1). Review on customer segmentation technique on ecommerce. *Advanced Science Letters, 22*(10), 3018-3022.

Yoseph, F., & AlMalaily, M. (2019, 2 28). NEW MARKET SEGMENTATION METHODS USING ENHANCED (RFM), CLV, MODIFIED REGRESSION AND CLUSTERING METHODS. *International Journal of Computer Science and Information Technology, 11*(01), 43-60.

(χ.χ.). *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών «Πληροφορική και Διοίκηση» Τμήματα Πληροφορικής και Οικονομικών Επιστημών Τμηματοποίηση Των Πελατών Με Την Χρήση Εργαλείων Επιχειρηματικής Ευφυΐας.*

NEW MARKET SEGMENTATION METHODS USING ENHANCED (RFM), CLV, MODIFIED REGRESSION AND CLUSTERING METHODS | International Journal of Computer Science and Information Technology (IJCSIT) - Academia.edu. (χ.χ.). Ανάκτηση από https://www.academia.edu/38528740/NEW_MARKET_SEGMENTATION_METHODS_USING_ENHANCED_RFM_CLV_MODIFIED_REGRESSION_AND_CLUSTERING_METHODS?overview=true

A RFMV Model and Customer Segmentation Based on Variety of Products | Saeedeh Rajaee Harandi - Academia.edu. (χ.χ.). Ανάκτηση από https://www.academia.edu/38332240/A_RFMV_Model_and_Customer_Segmentation_Based_on_Variety_of_Products?auto=download

Developing a model for measuring customer's loyalty and value with RFM technique and clustering algorithms | Golriz Amooee - Academia.edu. (χ.χ.). Ανάκτηση από https://www.academia.edu/4454026/Developing_a_model_for_measuring_customer_s_loyalty_and_value_with_RFM_technique_and_clustering_algorithms?email_work_card=title

Customer Lifetime Value: The Path to Profitability - V. Kumar - Βιβλία Google. (χ.χ.). Ανάκτηση από https://books.google.gr/books?id=k1KHFxS3gosC&pg=PA5&lpg=PA5&dq=customer+lifetime+value+weka&source=bl&ots=rNglov_2cd&sig=ACfU3U2BPpUpDewkPVOixPadI8fot0CtGg&hl=el&sa=X&ved=2ahUKEwiIsqWgxeLnAhVQ3aQKHdpdBcQQ6AEwCXoECAUQAQ#v=onepage&q&f=false

Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study - ScienceDirect. (χ.χ.). Ανάκτηση από https://www.sciencedirect.com/science/article/pii/S1877050910003868

A review of the application of RFM model | guhan raj - Academia.edu. (χ.χ.). Ανάκτηση από https://www.academia.edu/37188440/A_review_of_the_application_of_RFM_model

Διαδικτυακές Πηγές

How we do Customer segmentation — Customer science & Analytics. (χ.χ.). Ανάκτηση από https://towardsdatascience.com/how-we-do-customer-segmentation-customer-science-analytics-e237f3db32bb

RFM Analysis for Customer Segmentation | CleverTap. (χ.χ.). Ανάκτηση από https://clevertap.com/blog/rfm-analysis/

RFM Customer Segmentation for Shopify | How Many Steps Inc. (χ.χ.). Ανάκτηση από https://how-many-steps-inc.webflow.io/rfm-segmentation-overview

RFM Segmentation | RFM Analysis, Model, Marketing & Software | Optimove. (χ.χ.). Ανάκτηση από https://www.optimove.com/resources/learning-center/rfm-segmentation

The Difference Between Cold, Warm and Qualified Leads. (χ.χ.). Ανάκτηση από https://www.successagency.com/growth/2014/06/26/the-difference-between-cold-warm-and-qualified-leads/

Ανάλυση RFM - Galaxy CRM | myGalaxy. (χ.χ.). Ανάκτηση από https://www.mygalaxy.gr/el/RFMAnalysis_GalaxyCRM_Ver2132

Ο κανόνας Pareto 80/20 - Πως πετυχαίνω περισσότερα κάνοντας λιγότερα | Blog. (χ.χ.). Ανάκτηση από https://braining.gr/blog/%CE%BF-

Customer Lifetime Value: Ποια είναι η αξία των πελατών σας;. (χ.χ.). Ανάκτηση από https://developgreece.com/customer-lifetime-value/

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import datetime as dt
import warnings
warnings.filterwarnings("ignore")
```

```
df = pd.read_excel("/content/Online Retail.xlsx")
df.head()
```

```
df.shape
```

```
(541909, 8)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype         
---  ------       --------------   -----         
 0   InvoiceNo    541909 non-null  object        
 1   StockCode    541909 non-null  object        
 2   Description  540455 non-null  object        
 3   Quantity     541909 non-null  int64         
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64       
 6   CustomerID   406829 non-null  float64       
 7   Country      541909 non-null  object        
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
```

```
df.describe()
```

	Quantity	InvoiceDate	UnitPrice	CustomerID
count	541909.000000	541909	541909.000000	406829.000000
mean	9.552250	2011-07-04 13:34:57.156386048	4.611114	15287.690570
min	-80995.000000	2010-12-01 08:26:00	-11062.060000	12346.000000
25%	1.000000	2011-03-28 11:34:00	1.250000	13953.000000
50%	3.000000	2011-07-19 17:17:00	2.080000	15152.000000
75%	10.000000	2011-10-19 11:37:00	4.130000	16791.000000
max	80995.000000	2011-12-09 12:50:00	38970.000000	18287.000000
std	218.081158	NaN	96.759853	1713.600303

```
df.isnull().sum()
```

```
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     135080
Country             0
dtype: int64
```

```
df['Country'].value_counts()
```

```
Country
United Kingdom          495478
Germany                   9495
France                    8557
EIRE                      8196
Spain                     2533
Netherlands               2371
Belgium                   2069
Switzerland               2002
Portugal                  1519
Australia                 1259
Norway                    1086
Italy                      803
Channel Islands            758
Finland                    695
Cyprus                     622
Sweden                     462
Unspecified                446
Austria                    401
Denmark                    389
Japan                      358
Poland                     341
Israel                     297
USA                        291
Hong Kong                  288
Singapore                  229
Iceland                    182
Canada                     151
Greece                     146
Malta                      127
United Arab Emirates        68
European Community          61
RSA                         58
Lebanon                     45
Lithuania                   35
Brazil                      32
Czech Republic              30
Bahrain                     19
Saudi Arabia                10
Name: count, dtype: int64
```

```
df['Country'].nunique()
```

```
38
```

```
df['InvoiceDate'].min(), df['InvoiceDate'].max()
```

```
(Timestamp('2010-12-01 08:26:00'), Timestamp('2011-12-09 12:50:00'))
```

Data Preprocessing

```
df.duplicated().sum()
```

```
5268
```

```
df = df.drop_duplicates()
df.duplicated().sum()
```

```
0
```

```
df.shape
```

```
(536641, 8)
```

```
df.isnull().sum()
```

```
InvoiceNo           0
StockCode           0
Description      1454
Quantity            0
InvoiceDate         0
UnitPrice           0
CustomerID     133626
Country             0
dtype: int64
```

```
df = df[df['CustomerID'].notnull()]
df.isnull().sum()
```

```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

```
df.shape
```

```
(401015, 8)
```

```
df = df[df['Quantity'] > 0]
df.shape
```

```
(392732, 8)
```

```
df = df[df['UnitPrice'] > 0]
df.shape
```

```
(392692, 8)
```

```
df['TotalPrice'] = df['Quantity'] * df['UnitPrice']
df.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalPrice
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34

```
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])
df['InvoiceDate'].dtype
```

```
dtype('<M8[ns]')
```

```
df['InvoiceDate'].min(), df['InvoiceDate'].max()
```

```
(Timestamp('2010-12-01 08:26:00'), Timestamp('2011-12-09 12:50:00'))
```

```
df['InvoiceDate'].dt.year.value_counts()
```

```
InvoiceDate
2011    379913
2010     12779
Name: count, dtype: int64
```

```
df['InvoiceDate'].dt.month.value_counts()
```

```
InvoiceDate
11    62157
12    41483
10    35078
9     32482
8     27182
7     26824
6     25361
5     28320
4     22641
3     27957
2     24032
1     21675
Name: count, dtype: int64
```

```
df['InvoiceDate'].dt.day.value_counts()
```

```
InvoiceDate
6     16500
13    15045
14    14844
20    14836
18    14666
12    14660
17    14475
5     14466
4     14190
11    13943
19    13943
3     13851
10    13491
7     13490
2     13113
9     12890
16    12890
15    12705
21    12358
8     11731
23    10854
22    10854
24     8354
27     7675
28     7675
26     7675
25     7675
29     6911
30     6911
31     4344
1      4344
Name: count, dtype: int64
```

```
df['InvoiceDate'].dt.hour.value_counts()
```

```
InvoiceDate
12    59424
13    49484
14    43890
15    41976
11    41909
10    33719
16    29126
9     23226
17    19813
18     8124
19     4400
20     2000
8      1900
7       101
Name: count, dtype: int64
```

```
df['InvoiceDate'].dt.day_name().value_counts()
```

```
InvoiceDate
Thursday     79674
Wednesday    77865
Tuesday      75204
Monday       69240
Sunday       49749
Friday       40960
Name: count, dtype: int64
```

```
df['InvoiceDate'].dt.day_name().value_counts().plot(kind='bar', color='skyblue')
plt.title("Customer Registrations by Day of Week")
plt.xlabel("Day of Week")
plt.ylabel("Number of Registrations")
plt.show()
```



```
df['InvoiceDate'].dt.month.value_counts().plot(kind='bar', color='skyblue')
plt.title("Customer Registrations by Month")
plt.xlabel("Month")
plt.ylabel("Number of Registrations")
plt.show()
```



```
df['InvoiceDate'].dt.hour.value_counts().plot(kind='bar', color='skyblue')
plt.title("Customer Registrations by Hour")
plt.xlabel("Hour")
plt.ylabel("Number of Registrations")
plt.show()
```



```
df['InvoiceDate'].dt.day_name().value_counts().plot(kind='bar', color='skyblue')
plt.title("Customer Registrations by Day of Week")
plt.xlabel("Day of Week")
plt.ylabel("Number of Registrations")
plt.show()
```



```
df['Country'].value_counts().plot(kind='bar', color='skyblue')
plt.title("Customer Registrations by Country")
plt.xlabel("Country")
plt.ylabel("Number of Registrations")
plt.show()
```



```
df['Country'].value_counts(normalize=True).plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Country")
plt.ylabel("")
plt.show()
```



```
df['InvoiceDate'].dt.year.value_counts().plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Year")
plt.ylabel("")
plt.show()
```



```
df['InvoiceDate'].dt.month.value_counts().plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Month")
plt.ylabel("")
plt.show()
```



```
df['InvoiceDate'].dt.hour.value_counts().plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Hour")
plt.ylabel("")
plt.show()
```



```
df['InvoiceDate'].dt.day_name().value_counts().plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Day of Week")
plt.ylabel("")
plt.show()
```



```
df['Country'].value_counts(normalize=True).plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Country")
plt.ylabel("")
plt.show()
```



```
df['InvoiceDate'].dt.year.value_counts().plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Year")
plt.ylabel("")
plt.show()
```



```
df['InvoiceDate'].dt.month.value_counts().plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Month")
plt.ylabel("")
plt.show()
```



```
df['InvoiceDate'].dt.hour.value_counts().plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Hour")
plt.ylabel("")
plt.show()
```



```
df['InvoiceDate'].dt.day_name().value_counts().plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Day of Week")
plt.ylabel("")
plt.show()
```


RFM Model Implementation for Specific Region

```
df['Country'].value_counts()
```

```
Country
United Kingdom          356728
Germany                   9025
France                    8326
EIRE                      7226
Spain                     2479
Netherlands               2359
Belgium                   2031
Switzerland               1957
Portugal                  1462
Australia                 1182
Norway                    1071
Italy                      758
Channel Islands            748
Finland                    685
Cyprus                     603
Sweden                     450
Unspecified                443
Austria                    398
Denmark                    380
Japan                      321
Poland                     329
Israel                     253
USA                        179
Hong Kong                  284
Singapore                  229
Iceland                    182
Canada                     151
Greece                     145
Malta                      112
United Arab Emirates        68
European Community          60
RSA                         57
Lebanon                     45
Lithuania                   35
Brazil                      32
Czech Republic              25
Bahrain                     19
Saudi Arabia                9
Name: count, dtype: int64
```

```
df['Country'].nunique()
```

```
37
```

```
df['InvoiceDate'].min(), df['InvoiceDate'].max()
```

```
(Timestamp('2010-12-01 08:26:00'), Timestamp('2011-12-09 12:50:00'))
```

```
df['InvoiceDate'].dtype
```

```
dtype('<M8[ns]')
```

```
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])
df['InvoiceDate'].dtype
```

```
dtype('<M8[ns]')
```

```
df['InvoiceDate'].min(), df['InvoiceDate'].max()
```

```
(Timestamp('2010-12-01 08:26:00'), Timestamp('2011-12-09 12:50:00'))
```

```
df['InvoiceDate'].dt.year.value_counts()
```

```
InvoiceDate
2011    356728
2010     12779
Name: count, dtype: int64
```

```
df['InvoiceDate'].dt.month.value_counts()
```

```
InvoiceDate
11    58185
12    38481
10    32770
9     30379
8     25361
7     25091
6     23494
5     26315
4     21229
3     26148
2     22460
1     20615
Name: count, dtype: int64
```

```
df['InvoiceDate'].dt.day.value_counts()
```

```
InvoiceDate
6     15500
13    14045
14    13844
20    13836
18    13666
12    13660
17    13475
5     13466
4     13190
11    12943
19    12943
3     12851
10    12491
7     12490
2     12113
9     11890
16    11890
15    11705
21    11358
8     10731
23     9854
22     9854
24     7354
27     6675
28     6675
26     6675
25     6675
29     5911
30     5911
31     3344
1      3344
Name: count, dtype: int64
```

```
df['InvoiceDate'].dt.hour.value_counts()
```

```
InvoiceDate
12    55424
13    46484
14    40890
15    38976
11    38909
10    30719
16    26126
9     20226
17    16813
18     5124
19     1400
20      500
8       400
7       101
Name: count, dtype: int64
```

```
df['InvoiceDate'].dt.day_name().value_counts()
```

```
InvoiceDate
Thursday     76674
Wednesday    74865
Tuesday      72204
Monday       66240
Sunday       46749
Friday       38960
Name: count, dtype: int64
```

```
df['InvoiceDate'].dt.day_name().value_counts().plot(kind='bar', color='skyblue')
plt.title("Customer Registrations by Day of Week")
plt.xlabel("Day of Week")
plt.ylabel("Number of Registrations")
plt.show()
```



```
df['InvoiceDate'].dt.month.value_counts().plot(kind='bar', color='skyblue')
plt.title("Customer Registrations by Month")
plt.xlabel("Month")
plt.ylabel("Number of Registrations")
plt.show()
```



```
df['InvoiceDate'].dt.hour.value_counts().plot(kind='bar', color='skyblue')
plt.title("Customer Registrations by Hour")
plt.xlabel("Hour")
plt.ylabel("Number of Registrations")
plt.show()
```



```
df['InvoiceDate'].dt.day_name().value_counts().plot(kind='bar', color='skyblue')
plt.title("Customer Registrations by Day of Week")
plt.xlabel("Day of Week")
plt.ylabel("Number of Registrations")
plt.show()
```



```
df['Country'].value_counts().plot(kind='bar', color='skyblue')
plt.title("Customer Registrations by Country")
plt.xlabel("Country")
plt.ylabel("Number of Registrations")
plt.show()
```



```
df['Country'].value_counts(normalize=True).plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Country")
plt.ylabel("")
plt.show()
```



```
df['InvoiceDate'].dt.year.value_counts().plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Year")
plt.ylabel("")
plt.show()
```



```
df['InvoiceDate'].dt.month.value_counts().plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Month")
plt.ylabel("")
plt.show()
```



```
df['InvoiceDate'].dt.hour.value_counts().plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Hour")
plt.ylabel("")
plt.show()
```



```
df['InvoiceDate'].dt.day_name().value_counts().plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("Customer Registrations by Day of Week")
plt.ylabel("")
plt.show()
```


RFM Model Implementation for All Regions

```
df.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalPrice
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34

```
df['InvoiceDate'].max()
```

```
Timestamp('2011-12-09 12:50:00')
```

```
import datetime as dt
snapshot_date = df['InvoiceDate'].max() + dt.timedelta(days=1)
snapshot_date
```

```
Timestamp('2011-12-10 12:50:00')
```

```
rfm = df.groupby('CustomerID').agg({
    'InvoiceDate': lambda x: (snapshot_date - x.max()).days,
    'InvoiceNo': 'count',
    'TotalPrice': 'sum'
})
rfm.head()
```

CustomerID	InvoiceDate	InvoiceNo	TotalPrice
12346.0	326	1	77183.60
12347.0	2	182	4310.00
12348.0	75	31	1797.24
12349.0	19	73	1757.55
12350.0	310	17	334.40

```
rfm.rename(columns={'InvoiceDate': 'Recency', 'InvoiceNo': 'Frequency', 'TotalPrice': 'Monetary'}, inplace=True)
rfm.head()
```

CustomerID	Recency	Frequency	Monetary
12346.0	326	1	77183.60
12347.0	2	182	4310.00
12348.0	75	31	1797.24
12349.0	19	73	1757.55
12350.0	310	17	334.40

```
rfm.describe()
```

	Recency	Frequency	Monetary
count	4338.000000	4338.000000	4338.000000
mean	92.536422	90.523744	2048.688081
std	100.014169	225.506968	8985.230220
min	1.000000	1.000000	3.750000
25%	18.000000	17.000000	306.482500
50%	51.000000	41.000000	668.570000
75%	142.000000	98.000000	1660.597500
max	374.000000	7676.000000	259657.300000

```
rfm['R_Score'] = pd.qcut(rfm['Recency'], 4, labels=[4,3,2,1])
rfm['F_Score'] = pd.qcut(rfm['Frequency'].rank(method='first'), 4, labels=[1,2,3,4])
rfm['M_Score'] = pd.qcut(rfm['Monetary'], 4, labels=[1,2,3,4])
rfm.head()
```

CustomerID	Recency	Frequency	Monetary	R_Score	F_Score	M_Score
12346.0	326	1	77183.60	1	1	4
12347.0	2	182	4310.00	4	4	4
12348.0	75	31	1797.24	2	2	3
12349.0	19	73	1757.55	3	3	3
12350.0	310	17	334.40	1	2	1

```
rfm['RFM_Segment'] = rfm['R_Score'].astype(str) + rfm['F_Score'].astype(str) + rfm['M_Score'].astype(str)
rfm.head()
```

CustomerID	Recency	Frequency	Monetary	R_Score	F_Score	M_Score	RFM_Segment
12346.0	326	1	77183.60	1	1	4	114
12347.0	2	182	4310.00	4	4	4	444
12348.0	75	31	1797.24	2	2	3	223
12349.0	19	73	1757.55	3	3	3	333
12350.0	310	17	334.40	1	2	1	121

```
rfm['RFM_Segment'].value_counts()
```

```
RFM_Segment
111    321
444    321
222    321
333    321
121    321
112    321
211    321
212    321
221    321
122    321
434    321
343    321
433    321
344    321
443    321
334    321
Name: count, dtype: int64
```

```
rfm['RFM_Segment'].value_counts().plot(kind='bar', color='skyblue')
plt.title("RFM Segment Distribution")
plt.xlabel("RFM Segment")
plt.ylabel("Number of Customers")
plt.show()
```



```
rfm['RFM_Segment'].value_counts(normalize=True).plot(kind='pie', autopct='%1.1f%%', figsize=(8,8))
plt.title("RFM Segment Distribution")
plt.ylabel("")
plt.show()
```



```
rfm.groupby('RFM_Segment')['Monetary'].mean().sort_values(ascending=False).plot(kind='bar', color='skyblue')
plt.title("Average Monetary Value by RFM Segment")
plt.xlabel("RFM Segment")
plt.ylabel("Average Monetary Value")
plt.show()
```



```
rfm.groupby('RFM_Segment')['Frequency'].mean().sort_values(ascending=False).plot(kind='bar', color='skyblue')
plt.title("Average Frequency by RFM Segment")
plt.xlabel("RFM Segment")
plt.ylabel("Average Frequency")
plt.show()
```



```
rfm.groupby('RFM_Segment')['Recency'].mean().sort_values(ascending=False).plot(kind='bar', color='skyblue')
plt.title("Average Recency by RFM Segment")
plt.xlabel("RFM Segment")
plt.ylabel("Average Recency")
plt.show()
```


Customer Lifetime Value and Segmentation

```
df['InvoiceDate'].max()
```

```
Timestamp('2011-12-09 12:50:00')
```

```
import datetime as dt
snapshot_date = df['InvoiceDate'].max() + dt.timedelta(days=1)
snapshot_date
```

```
Timestamp('2011-12-10 12:50:00')
```

```
clv = df.groupby('CustomerID').agg({
    'InvoiceDate': lambda x: (snapshot_date - x.max()).days,
    'InvoiceNo': 'count',
    'TotalPrice': 'sum'
})
clv.head()
```

CustomerID	InvoiceDate	InvoiceNo	TotalPrice
12346.0	326	1	77183.60
12347.0	2	182	4310.00
12348.0	75	31	1797.24
12349.0	19	73	1757.55
12350.0	310	17	334.40

```
clv.rename(columns={'InvoiceDate': 'Recency', 'InvoiceNo': 'Frequency', 'TotalPrice': 'Monetary'}, inplace=True)
clv.head()
```

CustomerID	Recency	Frequency	Monetary
12346.0	326	1	77183.60
12347.0	2	182	4310.00
12348.0	75	31	1797.24
12349.0	19	73	1757.55
12350.0	310	17	334.40

```
clv['AOV'] = clv['Monetary'] / clv['Frequency']
clv.head()
```

CustomerID	Recency	Frequency	Monetary	AOV
12346.0	326	1	77183.60	77183.600000
12347.0	2	182	4310.00	23.681319
12348.0	75	31	1797.24	57.975484
12349.0	19	73	1757.55	24.076027
12350.0	310	17	334.40	19.670588

```
clv['Purchase_Frequency'] = clv['Frequency'] / clv['Frequency'].max()
clv.head()
```

CustomerID	Recency	Frequency	Monetary	AOV	Purchase_Frequency
12346.0	326	1	77183.60	77183.600000	0.000130
12347.0	2	182	4310.00	23.681319	0.023711
12348.0	75	31	1797.24	57.975484	0.004038
12349.0	19	73	1757.55	24.076027	0.009511
12350.0	310	17	334.40	19.670588	0.002215

```
clv['Retention_Rate'] = 1 - (clv['Recency'] / 365)
clv.head()
```

CustomerID	Recency	Frequency	Monetary	AOV	Purchase_Frequency	Retention_Rate
12346.0	326	1	77183.60	77183.600000	0.000130	0.106849
12347.0	2	182	4310.00	23.681319	0.023711	0.994521
12348.0	75	31	1797.24	57.975484	0.004038	0.794521
12349.0	19	73	1757.55	24.076027	0.009511	0.947945
12350.0	310	17	334.40	19.670588	0.002215	0.150685

```
clv['Churn_Rate'] = 1 - clv['Retention_Rate']
clv.head()
```

CustomerID	Recency	Frequency	Monetary	AOV	Purchase_Frequency	Retention_Rate	Churn_Rate
12346.0	326	1	77183.60	77183.600000	0.000130	0.106849	0.893151
12347.0	2	182	4310.00	23.681319	0.023711	0.994521	0.005479
12348.0	75	31	1797.24	57.975484	0.004038	0.794521	0.205479
12349.0	19	73	1757.55	24.076027	0.009511	0.947945	0.052055
12350.0	310	17	334.40	19.670588	0.002215	0.150685	0.849315

```
clv['CLV'] = clv['AOV'] * clv['Purchase_Frequency'] * clv['Retention_Rate'] / clv
```