



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Τμήμα Ψηφιακών Συστημάτων

Π.Μ.Σ «Πληροφοριακά Συστήματα και Υπηρεσίες»

Τεχνικές Εξόρυξης Δεδομένων

και Στρατηγικές Διαχείρισης Πελατειακών Σχέσεων

για την Πρόβλεψη της Αξίας Ζωής Πελατών σε εταιρείες iGaming

Μακρίδου Μαρία

A.M: 2049

Επιβλέπων Καθηγητής: Φιλιππάκης Μιχαήλ

Πειραιάς, Ιούλιος 2023

Περίληψη

Τα τελευταία χρόνια, οι εταιρείες iGaming έχουν γνωρίσει σημαντική ανάπτυξη, καθώς όλο και περισσότεροι άνθρωποι ασχολούνται με τις διαδικτυακές πλατφόρμες τυχερών παιχνιδιών και στοιχημάτων. Αυτή η αύξηση της δημοτικότητας αναδεικνύει τη σημασία της αποτελεσματικής διαχείρισης των σχέσεων με τους πελάτες (Customer Relationship Management - CRM) και της ακριβούς πρόβλεψης της αξίας διάρκειας ζωής τους (Customer Lifetime Value - CLV).

Ο κλάδος του iGaming έχει γνωρίσει μια αξιοσημείωτη επέκταση, η οποία οφείλεται στην πρόοδο της τεχνολογίας, στην ευρεία πρόσβαση στο διαδίκτυο και στην αυξανόμενη αποδοχή των διαδικτυακών τυχερών παιχνιδιών σε διάφορες δικαιοδοσίες. Ο αυξανόμενος αριθμός χρηστών σε αυτές τις πλατφόρμες παρουσιάζει ευκαιρίες και προκλήσεις για τις εταιρείες iGaming. Για να αξιοποιήσουν αυτή την ανάπτυξη, είναι ζωτικής σημασίας για τις εταιρείες αυτές να υιοθετήσουν ισχυρές στρατηγικές CRM και να χρησιμοποιήσουν προηγμένες τεχνικές ανάλυσης, όπως μοντέλα μηχανικής μάθησης για την πρόβλεψη του CLV.

Αξιοποιώντας τα μοντέλα μηχανικής μάθησης, οι εταιρείες iGaming μπορούν να ξεκλειδώσουν πολύτιμες πληροφορίες και να ενισχύσουν την ικανότητά τους να μεγιστοποιούν τη δια βίου αξία των πελατών, οδηγώντας σε βιώσιμη ανάπτυξη και ανταγωνιστικό πλεονέκτημα στη δυναμική βιομηχανία iGaming.

Η παρούσα διπλωματική εργασία επικεντρώνεται στη σύγκλιση τριών βασικών τομέων: του iGaming, της διαχείρισης πελατειακών σχέσεων (CRM) και των μοντέλων μηχανικής μάθησης, που χρησιμοποιούνται για την πρόβλεψη της διάρκειας ζωής του πελάτη (CLV).

Για την υλοποίηση της εργασίας χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python σε περιβάλλον Jupyter Notebook, για την ανάπτυξη των μοντέλων μηχανικής μάθησης, επιτρέποντάς τη δημιουργία αλγορίθμων πρόβλεψης για την εκτίμηση του δείκτη CLV.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω ιδιαίτερα τον κ. Μιχαήλ Φιλιππάκη για τη βοήθεια και την καθοδήγησή του κατά τη διάρκεια της συγγραφής της διπλωματικής αυτής εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω όλους τους καθηγητές του μεταπτυχιακού τμήματος Πληροφοριακά Συστήματα και Υπηρεσίες για τις γνώσεις που απέκτησα κατά τη διάρκεια της φοίτησής μου .

Περιεχόμενα

Περίληψη	2
Ευχαριστίες	3
Περιεχόμενα.....	4
Κεφάλαιο 1: Μεγάλα Δεδομένα	6
1.1 Ο ορισμός των Μεγάλων Δεδομένων	6
1.2 Τα χαρακτηριστικά των Μεγάλων Δεδομένων	7
1.3 Σημασία των Μεγάλων Δεδομένων.....	10
1.4 Ταξινόμηση των Μεγάλων Δεδομένων	12
1.5 Αποθήκευση και Επεξεργασία των Μεγάλων Δεδομένων	16
1.6 Παραδείγματα Μεγάλων Δεδομένων.....	18
Κεφάλαιο 2: Εξόρυξη Γνώσης και Δεδομένων.....	21
2.1 Ορισμός Εξόρυξης Γνώσης και Δεδομένων.....	21
2.2 Εξόρυξη Δεδομένων και Ανεύρεση Γνώσης.....	22
2.3 Στόχοι της Εξόρυξης Δεδομένων	23
2.4 Που εφαρμόζεται η Εξόρυξη Γνώσης;.....	24
2.5 Πλεονεκτήματα και Μειονεκτήματα της Εξόρυξης Γνώσης	26
Κεφάλαιο 3: Διαχείριση Πελατειακών Σχέσεων (CRM)	28
3.1 Εισαγωγή	28
3.2 Έννοια και Ορισμοί.....	28
3.3 Οι τέσσερις διαστάσεις του CRM.....	29
3.4 Τα είδη του CRM	30
Κεφάλαιο 4: Ρόλος και Εφαρμογή του Δείκτη Customer Lifetime Value (CLV).....	34
4.1 Εισαγωγή	34
4.2 Ορισμοί.....	34
4.3 Υπολογισμός του δείκτη Customer Lifetime Value (CLV)	35
4.4 Μοντελοποίηση του δείκτη Customer Lifetime Value (CLV)	36
4.4.1 Μοντέλα RFM.....	36
4.4.2 Μοντέλα Πιθανοτήτων (Profitability Models)	37
4.4.3 Οικονομετρικά Μοντέλα (Econometric Models)	38
4.4.4 Μοντέλα Μηχανικής Μάθησης (Machine Learning Models)	39
4.4.4.1 XGBoost Classifier.....	39
Κεφάλαιο 5: Δημιουργία Μοντέλου Μηχανικής Μάθησης για την πρόβλεψη του Δείκτη CLV.....	42
5.1 Εισαγωγή	42

5.2	Τεχνικό Υπόβαθρο	42
5.3	Περιγραφή Δεδομένων	44
5.4	Προεπεξεργασία Δεδομένων	44
5.5	Διερεύνηση των Δεδομένων	45
5.6	Εξαγωγή Χαρακτηριστικών.....	49
5.7	Κατηγοριοποίηση των πελατών με χρήση της μεθόδου RFM	52
5.7.1	Εγγύτητα (Recency).....	54
5.7.2	Συχνότητα (Frequency)	55
5.7.3	Νομισματική Αξία (Monetary)	56
5.7.4	Συνολική Βαθμολόγηση	57
5.8	Μοντελοποίηση Customer Lifetime Value.....	59
5.8.1	Εφαρμογή K-means.....	61
5.8.2	Feature Engineering	61
5.8.3	Μοντέλο Μηχανικής Μάθησης.....	62
5.8.4	Σύγκριση Μοντέλων Κατηγοριοποίησης	63
Κεφάλαιο 6: Συμπεράσματα και Μελλοντικές Επεκτάσεις		65
6.1	Συμπεράσματα	65
6.2	Μελλοντικές Επεκτάσεις	65
Βιβλιογραφία - Αναφορές.....		67

Κεφάλαιο 1: Μεγάλα Δεδομένα

1.1 Ο ορισμός των Μεγάλων Δεδομένων

Η σύγχρονη εποχή χαρακτηρίζεται και ως εποχή των «Μεγάλων Δεδομένων» λόγω της πρωτοφανούς κλίμακας δεδομένων που παράγονται καθημερινά και τις ανάγκες ανάλυσής τους για την εξαγωγή πολύτιμης γνώσης.

Ο όρος «Μεγάλα Δεδομένα» χρησιμοποιήθηκε για πρώτη φορά το 1997, όταν οι ερευνητές της NASA, Michael Cox και David Ellsworth, ήθελαν να περιγράψουν μια πρόκληση που αντιμετώπιζαν. Οι ποσότητες πληροφοριών που παράγονταν ήταν τόσο τεράστιες, που δεν μπορούσαν να υποβληθούν σε επεξεργασία και οπτικοποίηση. Μάλιστα ανέφεραν ότι «Τα σύνολα δεδομένων είναι τόσο μεγάλα, ώστε η αποθήκευσή τους στην κύρια μνήμη, στον τοπικό δίσκο, ακόμη και σε απομακρυσμένο δίσκο καθίσταται αδύνατη». Αυτό το ονόμασαν «Πρόβλημα των μεγάλων δεδομένων».

Στην πραγματικότητα, τα μεγάλα δεδομένα δεν είναι μια εντελώς νέα έννοια. Από τις πρώτες απαρχές του πολιτισμού, οι άνθρωποι προσπάθησαν να χρησιμοποιήσουν «δεδομένα» για καλύτερη λήψη αποφάσεων ή για να αποκτήσουν ένα ανταγωνιστικό πλεονέκτημα. Αυτή η έρευνα μπορεί να χρονολογηθεί ακόμη και από τους αρχαίους Αιγύπτιους και τη Ρωμαϊκή Αυτοκρατορία. Η Βιβλιοθήκη της Αλεξάνδρειας, η οποία ιδρύθηκε γύρω στο 300 π.Χ., μπορεί να θεωρηθεί ως η πρώτη προσπάθεια των αρχαίων Αιγυπτίων να συλλάβουν όλα τα «δεδομένα» εντός της αυτοκρατορίας. Υπολογίζεται ότι η βιβλιοθήκη αποτελούνταν από 40.000 έως 400.000 ειλητάρια, που θα ισοδυναμούσαν με περίπου 100.000 βιβλία. Ο αρχαίος ρωμαϊκός στρατός χρησιμοποίησε πολύ λεπτομερή στατιστική ανάλυση για να προβλέψει σε ποια σύνορα θα ήταν πιο πιθανή μια εχθρική εξέγερση. Με βάση αυτές τις αναλύσεις, μπόρεσαν να αναπτύξουν τους στρατούς τους με τον πιο αποτελεσματικό δυνατό τρόπο και να έχουν ένα πλεονέκτημα έναντι άλλων στρατών. Έτσι, μπορούμε θεωρήσουμε αυτούς τους υπολογισμούς μία από τις πρώτες μορφές «προγνωστικής» ανάλυσης δεδομένων.

Επομένως, γίνεται αντιληπτό ότι οι περισσότερες τεχνικές που χρησιμοποιούνται σήμερα, από τους αλγόριθμους πρόβλεψης και τις τεχνικές ταξινόμησης, έχουν αναπτυχθεί πριν από αιώνες. Όμως, αυτό που έχει αλλάξει είναι η διαθεσιμότητα και η προσβασιμότητα σε τεράστιες ποσότητες δεδομένων. Ειδικά από τις αρχές του 21ου αιώνα, ο όγκος και η ταχύτητα με την

οποία παράγονται τα δεδομένα έχουν αλλάξει πέρα από τα μέτρα της ανθρώπινης κατανόησης. Ο συνολικός όγκος δεδομένων στον κόσμο ήταν 9 zettabytes το 2013, ο αριθμός το 2021 άγγιξε τα 79 zettabytes, που ισοδυναμούν με 79 τρισεκατομμύρια gigabyte. Μάλιστα υπολογίζεται ότι μέσα στα επόμενα 3 χρόνια, ο αριθμός αυτός θα έχει διπλασιαστεί.

Ο όρος «Μεγάλα Δεδομένα» χρησιμοποιείται τα τελευταία χρόνια, χωρίς να υπάρχει κάποιος σαφής ορισμός του.

Ο ορισμός που έχει επικρατήσει είναι αυτός της Gartner, της μεγαλύτερης επιχείρησης στον κόσμο που ασχολείται με την τεχνολογική έρευνα και τη συμβουλευτική. Η Gartner το 2012 έδωσε τον εξής ορισμό: «Τα μεγάλα δεδομένα είναι στοιχεία πληροφοριών μεγάλου όγκου (Volume), υψηλής ταχύτητας (Velocity) και μεγάλης ποικιλίας (Variety) που απαιτούν οικονομικά αποδοτικές, καινοτόμες μορφές επεξεργασίας πληροφοριών που επιτρέπουν βελτιωμένη διορατικότητα, λήψη αποφάσεων και αυτοματοποίηση διαδικασιών». Δηλαδή, η Gartner πρότεινε έναν ορισμό που περιλάμβανε τα "τρία Vs". Στη συνέχεια, ο ορισμός αυτός επεκτάθηκε σε 7 Vs.

1.2 Τα χαρακτηριστικά των Μεγάλων Δεδομένων

Προτού αναφερθούμε στη σημασία και τη σπουδαιότητα των μεγάλων δεδομένων, πρέπει να εστιάσουμε στα χαρακτηριστικά που διαθέτουν, τα οποία είναι γνωστά πλέον ως 7 Vs. Συγκεκριμένα:

1. Volume (Όγκος)

Ο όγκος είναι ένα προφανές χαρακτηριστικό των μεγάλων δεδομένων και αφορά κυρίως τη σχέση μεταξύ μεγέθους και ικανότητας επεξεργασίας. Ο χαρακτηρισμός «μέγας όγκος» ποικίλει ανάλογα με το αντικείμενο της ανάλυσης. Υπάρχουν επιχειρήσεις που διαχειρίζονται petabytes δεδομένων και άλλες που διαχειρίζονται zettabytes. Επιπλέον, ο όγκος των δεδομένων αλλάζει γρήγορα καθώς η συλλογή δεδομένων συνεχίζει να αυξάνεται. Ακριβώς όπως και η ικανότητα της πληροφορικής για την αποθήκευση και την επεξεργασία τους.

2. Velocity (Ταχύτητα)

Με τον όρο velocity αναφερόμαστε στην υψηλή ταχύτητα με την οποία παράγονται, συσσωρεύονται και αναλύονται τα δεδομένα. Μάλιστα, αυτή η ταχύτητα ακολουθεί μία

εκθετική συνάρτηση. Αν και τα περισσότερα δεδομένα αποθηκεύονται πριν από την ανάλυση, υπάρχει αυξανόμενη ανάγκη για επεξεργασία σε πραγματικό χρόνο αυτών των τεράστιων όγκων. Η επεξεργασία σε πραγματικό χρόνο μειώνει τις απαιτήσεις αποθήκευσης, ενώ παρέχει πιο ακριβείς και κερδοφόρες αποκρίσεις.

3. Variety (Ποικιλία)

Μια άλλη πρόκληση της επεξεργασίας των μεγάλων δεδομένων είναι και η μεγάλη ποικιλία δεδομένων που αποθηκεύονται και πρέπει να υποβληθούν σε επεξεργασία και ανάλυση. Τα δεδομένα χωρίζονται σε τρεις κατηγορίες, τα δομημένα, τα ημι-δομημένα και τα αδόμητα, που θα μελετηθούν εκτενέστερα σε επόμενη ενότητα.

Δομημένα: Τα δομημένα δεδομένα είναι τα δεδομένα που συμμορφώνονται με ένα μοντέλο δεδομένων, έχουν μια καλά καθορισμένη δομή, ακολουθούν μια συνεπή σειρά και μπορούν εύκολα να προσπελαστούν και να χρησιμοποιηθούν από ένα άτομο ή ένα πρόγραμμα υπολογιστή. Τα δομημένα δεδομένα αποθηκεύονται συνήθως σε καλά καθορισμένα σχήματα όπως οι βάσεις δεδομένων, που αποτελούνται από πίνακες με στήλες και γραμμές που ορίζουν με σαφήνεια τα χαρακτηριστικά τους. Για την αποθήκευση και την επεξεργασία τους χρησιμοποιείται η γλώσσα προγραμματισμού SQL (Structured Query Language).

Ημι-δομημένα: Τα δεδομένα αυτά δε συμμορφώνονται με ένα μοντέλο δεδομένων αλλά έχουν κάποια δομή. Είναι τα δεδομένα που δεν βρίσκονται σε μια βάση δεδομένων, αλλά έχουν ορισμένες οργανωτικές ιδιότητες που διευκολύνουν την ανάλυση. Με ορισμένες διεργασίες, μπορούμε να τις αποθηκεύσουμε στη σχεσιακή βάση δεδομένων.

Μη δομημένα: Τα δεδομένα αυτά δεν έχουν εύκολα αναγνωρίσιμη δομή έτσι ώστε να μην μπορούν να χρησιμοποιηθούν εύκολα από ένα πρόγραμμα υπολογιστή. Τα μη δομημένα δεδομένα δεν είναι οργανωμένα με προκαθορισμένο τρόπο, επομένως δεν είναι κατάλληλα για μια σχεσιακή βάση δεδομένων.

4. Veracity (Εγκυρότητα)

Η ανάλυση των μεγάλων δεδομένων είναι αρκετά σημαντική. Ωστόσο, είναι άχρηστη αν τα δεδομένα που αναλύονται είναι ανακριβή ή ελλιπή. Συχνά, καλούμαστε να αντιμετωπίσουμε δεδομένα που περιέχουν σφάλματα, ελλειπούσες τιμές και «θόρυβο», καθώς προέρχονται από διαφορετικές πηγές που παρουσιάζουν ποικιλία μορφών. Είναι αυτονόητο ότι η

ακρίβεια της τελικής ανάλυσης υποβαθμίζεται χωρίς πρώτα να καθαριστούν τα δεδομένα με τα οποία λειτουργεί. Για να είναι εφικτή η αξιόπιστη ανάλυση και η παραγωγή ορθών αποτελεσμάτων, έχουν αναπτυχθεί διάφορες μέθοδοι καθαρισμού δεδομένων (data cleansing).

5. Value (Αξία)

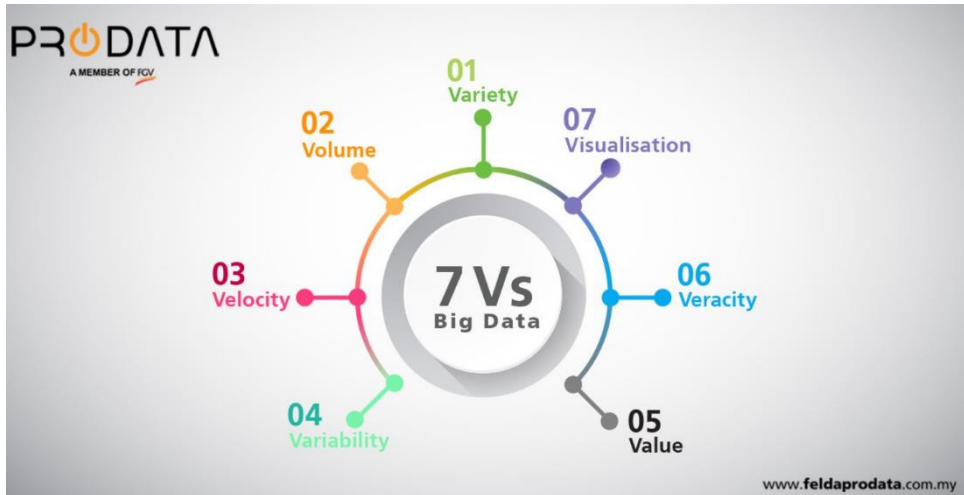
Ο τεράστιος όγκος δεδομένων είναι άχρηστος αν δε μπορούμε να τον μετατρέψουμε σε αξία. Η συλλογή και η πρόσβαση σε δεδομένα δεν έχουν καμία σημασία από μόνες τους. Για αυτό οι οργανισμοί καλούνται να τα αξιοποιήσουν αποτελεσματικά για την έγκαιρη και όσο τον δυνατόν ορθότερη λήψη αποφάσεων. Επομένως, ο όρος Value αναφέρεται στην αξία που μπορούν να αποκτήσουν τα δεδομένα ύστερα από τη σωστή διαχείρισή τους. Ως εκ τούτου, ορθώς ο όρος Value χαρακτηρίζεται ως το σημαντικότερο V των μεγάλων δεδομένων.

6. Variability (Μεταβλητότητα)

Ο όρος Variability αναφέρεται στο πόσο γρήγορα και σε ποιο βαθμό αλλάζει η δομή των δεδομένων. Είναι μια έννοια διαφορετική από την ποικιλία, καθώς η μεταβλητότητα αναφέρεται στα δεδομένα που αλλάζουν συνεχώς. Η απεριόριστη μεταβλητότητα των μεγάλων δεδομένων παρουσιάζει μια μοναδική πρόκληση αποκωδικοποίησης και μπορεί να έχει αντίκτυπο στην ποιότητα των δεδομένων. Επομένως, αν κάποιος θέλει να επωφεληθεί από την πλήρη αξία των δεδομένων, πρέπει να εστιάζει κυρίως στην κατανόηση και την ερμηνεία των σωστών νοημάτων των ακατέργαστων δεδομένων.

7. Visualization (Οπτικοποίηση)

Κάθε σύστημα επεξεργασίας δεδομένων, είναι σημαντικό, να μετατρέπει την τεράστια κλίμακα τους σε κάτι εύκολα κατανοητό και εφαρμόσιμο. Μία από τις καλύτερες μεθόδους για αυτό είναι η μετατροπή του σε μορφές γραφικών. Η οπτικοποίηση αναφέρεται στον τρόπο με τον οποίο μπορούν να αναπαρασταθούν τα δεδομένα ώστε να είναι εύκολα αναγνώσιμα, κατανοητά και προσβάσιμα. Επομένως, η οπτικοποίηση των δεδομένων είναι σημαντική για την λήψη καλύτερων αποφάσεων.



Εικόνα 1 - Τα χαρακτηριστικά των μεγάλων δεδομένων
FGV Prodata Systems

1.3 Σημασία των Μεγάλων Δεδομένων

Η σημασία των μεγάλων δεδομένων έγκειται στον τρόπο με τον οποίο μια εταιρεία αξιοποιεί τα συγκεντρωμένα δεδομένα. Κάθε εταιρεία χρησιμοποιεί τα δεδομένα που συλλέγει με τον δικό της τρόπο. Όσο πιο αποτελεσματικά χρησιμοποιεί τα δεδομένα της, τόσο πιο γρήγορα αναπτύσσεται.

Συγκεκριμένα, τα οφέλη από τη συλλογή και την ανάλυση των δεδομένων είναι:

1. Εξοικονόμηση κόστους (Cost Saving)

Οι επιχειρήσεις που χρησιμοποιούν εργαλεία διαχείρισης μεγάλων δεδομένων, μπορούν να πετύχουν εξοικονόμηση κόστους. Αυτά τα εργαλεία βοηθούν τις επιχειρήσεις να εντοπίσουν πιο αποτελεσματικούς τρόπους επιχειρηματικής δραστηριότητας.

2. Εξοικονόμηση χρόνου (Time Saving)

Η δυνατότητα των επιχειρήσεων να αποθηκεύουν τα δεδομένα τους σε πραγματικό χρόνο, βοηθά τις επιχειρήσεις να συλλέγουν δεδομένα ταυτόχρονα από διάφορες πηγές. Με αυτόν τον τρόπο, αναλύοντας αμέσως τα συγκεντρωμένα δεδομένα, επιτυγχάνεται και η λήψη γρήγορων αποφάσεων.

3. Ακρόαση των μέσων κοινωνικής δικτύωσης (Social media Listening)

Οι εταιρείες χρησιμοποιώντας εργαλεία ανάλυσης δεδομένων, έχουν τη δυνατότητα να λαμβάνουν σχόλια για την εταιρεία τους από τα μέσα κοινωνικής δικτύωσης και να αναλύουν αυτές τις κριτικές και τα συναισθήματα των πελατών. Με αυτόν τον τρόπο, μπορούν να βελτιώσουν την παρουσία τους στο διαδίκτυο.

4. Ενίσχυση της απόκτησης και της διατήρησης πελατών (Boosting Customer Acquisition and Retention)

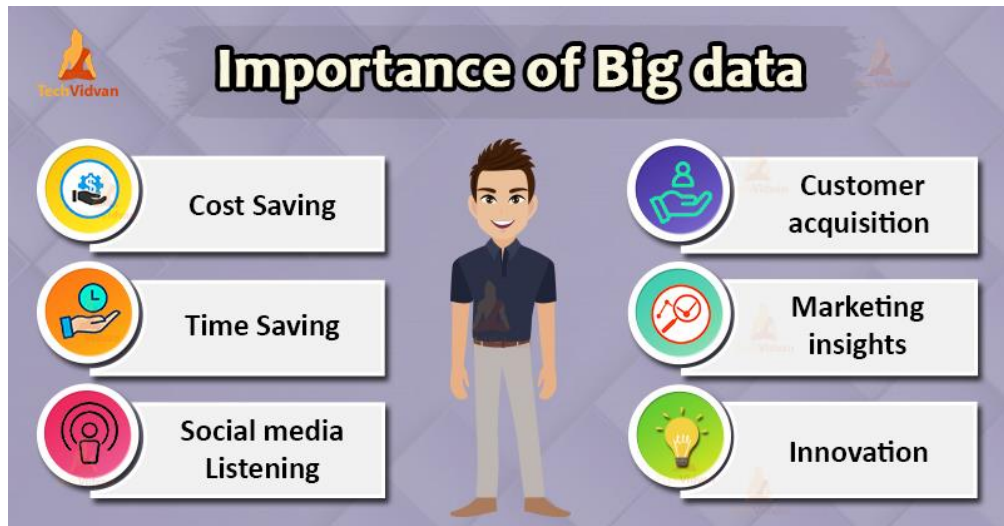
Οι πελάτες είναι ένα στοιχείο από το οποίο εξαρτάται κάθε επιχείρηση. Καμία επιχείρηση δεν μπορεί να επιτύχει χωρίς να δημιουργήσει μια ισχυρή πελατειακή βάση. Αλλά ακόμη και με μια σταθερή πελατειακή βάση, οι εταιρείες δεν μπορούν να αγνοήσουν τον ανταγωνισμό στην αγορά. Εάν μια εταιρεία δε γνωρίζει τι θέλουν οι πελάτες της, τότε υποβαθμίζεται η επιτυχία της. Αυτό έχει ως αποτέλεσμα την απώλεια πελατολογίου, πράγμα το οποίο δημιουργεί αρνητικές επιπτώσεις στην ανάπτυξη των επιχειρήσεων. Η ανάλυση μεγάλων δεδομένων βοηθά τις επιχειρήσεις να εντοπίσουν τάσεις και μοτίβα που σχετίζονται με τους πελάτες. Η ανάλυση της συμπεριφοράς των πελατών οδηγεί σε μια κερδοφόρα επιχείρηση.

5. Προσφορά πληροφοριών μάρκετινγκ (Offering Marketing Insights)

Η ανάλυση μεγάλων δεδομένων διαμορφώνει όλες τις επιχειρηματικές δραστηριότητες. Επιτρέπει στις εταιρείες να εκπληρώσουν τις προσδοκίες των πελατών και να κάνουν αλλαγές στις σειρές των προϊόντων τους. Με αυτόν τον τρόπο, εξασφαλίζονται και ισχυρές εκστρατείες μάρκετινγκ.

6. Κατανόηση των συνθηκών της αγοράς και Καινοτομία (Understanding the market conditions and Innovation)

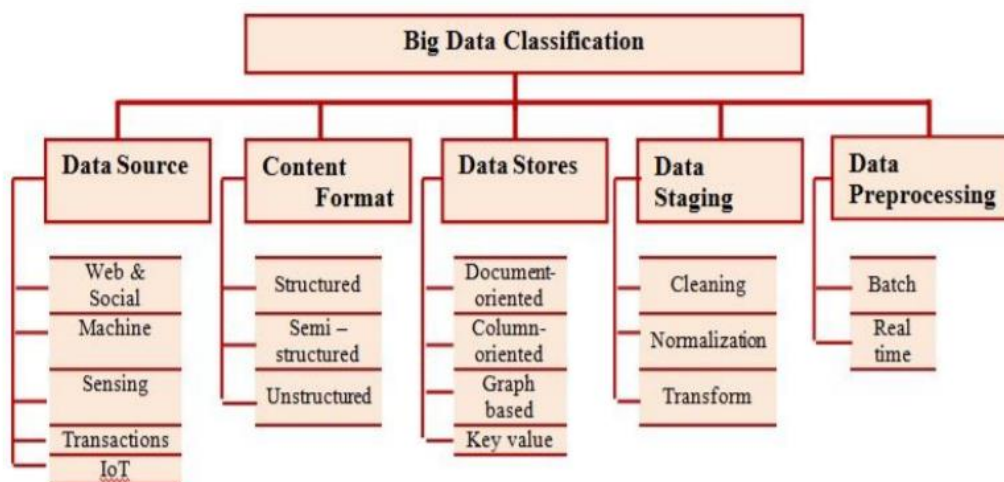
Η ανάλυση των μεγάλων δεδομένων βοηθά τις επιχειρήσεις να κατανοήσουν καλύτερα τις συνθήκες της αγοράς. Για παράδειγμα, η ανάλυση της αγοραστικής συμπεριφοράς των πελατών βοηθά τις εταιρείες να εντοπίσουν τα προϊόντα που πωλούνται περισσότερο. Έτσι, οι εταιρείες καθίστανται ικανές να παράγουν αυτά τα προϊόντα, αλλά και να καινοτομούν ανανεώνοντας τα προϊόντα τους, με σκοπό να ξεπεράσουν τους ανταγωνιστές τους.



Εικόνα 2 - Η σημασία των Μεγάλων Δεδομένων
TechVidvan

1.4 Ταξινόμηση των Μεγάλων Δεδομένων

Για την καλύτερη κατανόηση των μεγάλων δεδομένων και των χαρακτηριστικών τους, είναι σημαντικό να γίνει αναφορά στην Ταξινόμηση των Μεγάλων Δεδομένων (Big Data Classification). Οι πέντε κατηγορίες στις οποίες ταξινομούνται είναι: 1. Πηγές δεδομένων, 2. Δομή Περιεχομένου, 3. Αποθήκες Δεδομένων, 4. Σταδιοποίηση Δεδομένων και 5. Επεξεργασία Δεδομένων.



Εικόνα 3 - Ταξινόμηση των Μεγάλων Δεδομένων
Research India Publications

1. Πηγές δεδομένων (Data Sources)

- Διαδίκτυο και Κοινωνικά μέσα (Web & Social)

Καθημερινά δημιουργούνται στα μέσα κοινωνικής δικτύωσης (Twitter, Facebook, Instagram) εκατομμύρια GB δεδομένων. Μέσω διευθύνσεων URL, τα δεδομένα αυτά διαδίδονται σε όλον τον πλανήτη και ανταλλάσσονται πληροφορίες.

- Μηχανικά Δεδομένα (Machine-generated data)

Πρόκειται για δεδομένα που παράγονται αυτόματα από υλικό ή λογισμικό, χωρίς την παρέμβαση ανθρώπων. Τέτοιες πληροφορίες δημιουργούνται από υπολογιστές, ιατρικές συσκευές ή άλλα μηχανήματα.

- Δεδομένα Ανίχνευσης (Sensing)

Είναι δεδομένα που δημιουργούνται από μηχανές ανίχνευσης για τη μέτρηση των φυσικών μεγεθών και από αισθητήρες. Για παράδειγμα, οι αισθητήρες που ανιχνεύουν την κίνηση στα συστήματα συναγερμών.

- Συναλλαγές (Transactions)

Τα δεδομένα συναλλαγών, όπως είναι οι συναλλακτικές κινήσεις που πραγματοποιούνται μέσω χρεωστικών και πιστωτικών καρτών, περιέχουν και μία χρονική διάσταση για την περιγραφή των δεδομένων.

- Διαδίκτυο των πραγμάτων (Internet of Things)

Με τον όρο διαδίκτυο των πραγμάτων αναφερόμαστε στη διαδικασία με την οποία αντικείμενα που είναι συνδεδεμένα στο διαδίκτυο να μπορούν να διασυνδέονται. Τέτοια αντικείμενα μπορεί να είναι smartphones, tablets, ψηφιακές φωτογραφικές μηχανές ή ακόμα και οικιακές συσκευές και αυτοκίνητα. Μέσω αυτής της διαδικασίας, επιτυγχάνεται ο έλεγχος των αυτών συσκευών από τον υπολογιστή ή από το smartphone του κάθε χρήστη.

2. Δομή Περιεχομένου (Content Format)

- Δομημένα δεδομένα (Structured data)

Τα δομημένα δεδομένα είναι τα δεδομένα που συμμορφώνονται με ένα μοντέλο δεδομένων, έχουν μια καλά καθορισμένη δομή, ακολουθούν μια συνεπή σειρά και μπορούν εύκολα να προσπελαστούν και να χρησιμοποιηθούν από ένα άτομο ή ένα πρόγραμμα υπολογιστή. Τα

δομημένα δεδομένα αποθηκεύονται συνήθως σε καλά καθορισμένα σχήματα όπως οι βάσεις δεδομένων, που αποτελούνται από πίνακες με στήλες και γραμμές που ορίζουν με σαφήνεια τα χαρακτηριστικά τους. Για την αποθήκευση και την επεξεργασία τους χρησιμοποιείται η γλώσσα προγραμματισμού SQL (Structured Query Language).

- Ημι-δομημένα δεδομένα (Semi-structured data)

Τα δεδομένα αυτά δε συμμορφώνονται με ένα μοντέλο δεδομένων αλλά έχουν κάποια δομή. Είναι τα δεδομένα που δεν βρίσκονται σε μια βάση δεδομένων, αλλά έχουν ορισμένες οργανωτικές ιδιότητες που διευκολύνουν την ανάλυση. Με ορισμένες διεργασίες, μπορούμε να τις αποθηκεύσουμε στη σχεσιακή βάση δεδομένων.

- Αδόμητα δεδομένα (Unstructured data)

Τα δεδομένα αυτά δεν έχουν εύκολα αναγνωρίσιμη δομή έτσι ώστε να μην μπορούν να χρησιμοποιηθούν εύκολα από ένα πρόγραμμα υπολογιστή. Τα αδόμητα δεδομένα δεν είναι οργανωμένα με προκαθορισμένο τρόπο, επομένως δεν είναι κατάλληλα για μια σχεσιακή βάση δεδομένων.

3. Αποθήκες Δεδομένων (Data Stores)

Εκτός από το σχεσιακό μοντέλο, στο οποίο τα δεδομένα αποθηκεύονται σε πίνακες που αποτελούνται από γραμμές και στήλες, μπορούμε να αποθηκεύσουμε τα δεδομένα και σε μη σχεσιακές βάσεις δεδομένων, δηλαδή NoSQL (Not only SQL):

- Προσανατολισμένες σε έγγραφα (Document-oriented)

Οι αποθήκες δεδομένων που είναι προσανατολισμένες σε έγγραφα χρησιμοποιούνται για δεδομένα που ακολουθούν μορφή εγγράφου, όπως είναι οι συνταγές φαρμάκων ή μια σύμβαση εργασίας. Είναι σχεδιασμένες για να υποστηρίζουν σύνθετους τύπους δεδομένων, όπως XML, JSON, PDF, MS Word. Σημαντικό χαρακτηριστικό τους είναι ότι είναι πιο ευέλικτες και ταχύτερες στην ανάκτηση περιεχομένου από τις σχεσιακές βάσεις. Οι πιο γνωστές, τέτοιου είδους, αποθήκες είναι: MongoDB, Amazon DynamoDB, Microsoft Azure CosmosDB, Couchbase, Firebase Realtime Database.

- Προσανατολισμένες σε στήλες (Column-oriented)

Μια βάση δεδομένων με προσανατολισμό στις στήλες προτιμάται όταν τίθενται ερωτήματα που αφορούν μόνο λίγες από τις στήλες ενός μεγάλου μεγέθους πίνακα δεδομένων. Αυτές οι βάσεις δεδομένων αποθηκεύουν το περιεχόμενό τους σε στήλες εκτός από τις γραμμές, με τις τιμές χαρακτηριστικών που ανήκουν στην ίδια στήλη να αποθηκεύονται συνεχόμενα. Τέτοιες αποθήκες δεδομένων είναι οι MariaDB, CrateDB, ClickHouse, Greenplum Database, Apache Hbase.

- Βάση δεδομένων γραφημάτων (Graph based)

Μια βάση δεδομένων γραφημάτων είναι σχεδιασμένη για δεδομένα που χρησιμοποιούν ένα μοντέλο γραφημάτων με κόμβους, άκρες και ιδιότητες που σχετίζονται μεταξύ τους. Προτιμάται σε δεδομένα που έχουν μεγάλη συσχέτιση μεταξύ τους. Τέτοιου είδους βάσεις είναι οι Neo4j, ArangoDB, OrientDB, Ontotext GraphDB, FlockDB.

- Βάση δεδομένων κλειδιού- τιμής (Key-value)

Μια βάση δεδομένων κλειδιών-τιμών είναι ένας εναλλακτικό τύπος σχεσιακής βάσης που αποθηκεύει και αποκτά πρόσβαση σε δεδομένα που έχουν σχεδιαστεί για να φτάσουν πολύ μεγάλο μέγεθος. Μία τέτοια βάση αποθηκεύει τα δεδομένα ως μία συλλογή ζευγών κλειδιών-τιμών στα οποία ένα κλειδί χρησιμεύει ως μοναδικό αναγνωριστικό. Και τα κλειδιά και οι τιμές μπορεί να είναι είτε απλά είτε πολύπλοκα αντικείμενα. Ορισμένες βάσεις αυτού του είδους είναι οι Amazon DynamoDB, Amazon ElastiCache, Redis, Couchbase, Aerospike.

4. Σταδιοποίηση Δεδομένων (Data Staging)

- Καθαρισμός (Cleaning)

Είναι η διαδικασία με την οποία εντοπίζονται και αντικαθίστανται οι ελλιπείς τιμές της βάσης. Επιπλέον, αφαιρείται ο θόρυβος και οι μη εύλογες τιμές.

- Κανονικοποίηση (Normalization)

Είναι η διαδικασία αναδιοργάνωσης της βάσης δεδομένων έτσι ώστε να πληροί δύο βασικές απαιτήσεις, δηλαδή να μην υπάρχει πλεονασμός δεδομένων και όλα τα δεδομένα να αποθηκεύονται σε ένα μέρος.

- Μετασχηματισμός (Transform)

Είναι η διαδικασία με την οποία τα δεδομένα μετατρέπονται σε μία μορφή κατάλληλη για επεξεργασία και ανάλυση.

5. Επεξεργασία Δεδομένων (Data Processing)

- Κατά δεσμίδες (Batch)

Στην επεξεργασία δεδομένων κατά δεσμίδες, τα δεδομένα συλλέγονται, εισάγονται στη βάση και αποθηκεύονται για κάποιο χρονικό διάστημα, μέχρι να υποστούν επεξεργασία και ανάλυση. Τέτοια συστήματα επιτρέπουν την κλιμάκωση των εφαρμογών σε μεγάλες συστάδες μηχανών που περιλαμβάνουν χιλιάδες κόμβους.

- Σε πραγματικό χρόνο (Real time)

Η επεξεργασία δεδομένων σε πραγματικό χρόνο πραγματοποιείται καθώς τα δεδομένα εισάγονται στη βάση, σχεδόν αμέσως μετά από τη δημιουργία τους. Αυτά τα συστήματα χρησιμοποιούνται όταν υπάρχει μεγάλος αριθμός γεγονότων που πρέπει να υποβληθεί σε επεξεργασία σε σύντομο χρονικό διάστημα.

1.5 Αποθήκευση και Επεξεργασία των Μεγάλων Δεδομένων

Τα τελευταία χρόνια, η παραγωγή των δεδομένων έχει αυξηθεί εκθετικά. Έτσι, η αποθήκευση και η επεξεργασία τους αποτελούν πρώτιστη ανάγκη των επιχειρήσεων.

Από τη στιγμή που οι κυβερνήσεις άρχισαν να παρακολουθούν καλύτερα τα αρχεία και τα έγγραφα των πολιτών, έγινε εμφανής η ανάγκη για συστήματα κατάλληλα για την αποθήκευση και την επεξεργασία αυτών των δεδομένων. Επιπλέον, η εισαγωγή του Διαδικτύου των πραγμάτων (IoT) και του Διαδικτύου, έκανε αυτήν την ανάγκη ακόμα μεγαλύτερη.

Ωστόσο, δεν είναι μόνο οι μεγάλες επιχειρήσεις που πρέπει να τα χειριστούν μεγάλα δεδομένα, αλλά και οι μικρότερες επιχειρήσεις συλλέγουν καθημερινά πολλές πληροφορίες από πηγές όπως μηνύματα ηλεκτρονικού ταχυδρομείου, αλληλεπιδράσεις μέσω κοινωνικής δικτύωσης, πωλήσεις και μια ποικιλία άλλων πηγών. Ανεξάρτητα από το μέγεθος της εταιρείας ή τον κλάδο στον οποίο ανήκουν, τα δεδομένα πρέπει να αποθηκευτούν κάπου πριν μπορέσουν να ταξινομηθούν, να επεξεργαστούν και να αναλυθούν.

Η αποθήκευση μεγάλων δεδομένων δημιουργεί μεγάλες προκλήσεις, όπως είναι το κόστος, η επεκτασιμότητα, οι ρυθμοί μεταφοράς δεδομένων και η προστασία των δεδομένων. Τα μεγάλα δεδομένα αναπτύσσονται καθημερινά, απαιτώντας σημαντικό χώρο αποθήκευσης. Καθώς αυξάνονται οι πηγές δεδομένων, αυξάνονται και οι απαιτήσεις που πρέπει να ληφθούν υπόψη. Το σύστημα αποθήκευσης θα πρέπει να είναι σε θέση να κλιμακώνεται με τον ίδιο ρυθμό με τη συλλογή δεδομένων, ώστε να συμβαδίζει με την ανάπτυξη. Για να μεταφερθούν μεγάλοι όγκοι δεδομένων, οι υψηλοί ρυθμοί μεταφοράς είναι βασική απαίτηση. Σε περιβάλλοντα μεγάλων δεδομένων, θα πρέπει τα δεδομένα να μετακινούνται γρήγορα από τις πρωτογενείς πηγές στο περιβάλλον ανάλυσής τους. Τα μεγάλα δεδομένα περιέχουν συχνά ευαίσθητα δεδομένα, όπως προσωπικά στοιχεία ταυτοποίησης ή οικονομικά δεδομένα. Η προστασία τους είναι μια άλλη κρίσιμη πτυχή για κάθε επιχείρηση, επομένως θα πρέπει να λειτουργεί μόνο με κρυπτογραφημένα δεδομένα για προστασία της ιδιωτικής ζωής.

Επί του παρόντος, υπάρχουν δύο καθιερωμένες μέθοδοι αποθήκευσης μεγάλων δεδομένων, οι αποθήκες δεδομένων (Warehouse Storage) και η αποθήκευση σε cloud (Cloud Storage).

- Αποθήκες δεδομένων (Warehouse Storage):

Η κύρια λειτουργία μίας αποθήκης δεδομένων είναι η αποθήκευση και η επεξεργασία δεδομένων σε επίπεδο επιχείρησης. Αποτελεί ένα σημαντικό εργαλείο για την ανάλυση μεγάλων δεδομένων. Αυτές οι αποθήκες δεδομένων υποστηρίζουν τις διάφορες αναφορές, την επιχειρηματική ευφυΐα (Business Intelligence), την ανάλυση, την εξόρυξη δεδομένων, την έρευνα και άλλες σχετικές δραστηριότητες.

Τα εργαλεία αποθήκης δεδομένων καθιστούν δυνατή την αποτελεσματικότερη διαχείριση δεδομένων, καθώς επιτρέπουν τη δυνατότητα εύρεσης, πρόσβασης, οπτικοποίησης και ανάλυσης δεδομένων για τη λήψη καλύτερων επιχειρηματικών αποφάσεων και την επίτευξη πιο επιθυμητών επιχειρηματικών αποτελεσμάτων. Επιπλέον, κατασκευάζονται λαμβάνοντας υπόψη την εκθετική αύξηση των δεδομένων. Δεν υπάρχει κίνδυνος να γεμίσουν οι αποθήκες από τον αυξανόμενο όγκο δεδομένων που αποθηκεύονται.

Το μεγαλύτερο πλεονέκτημα των αποθηκών δεδομένων είναι η ικανότητα να μεταφράζουν ακατέργαστα δεδομένα σε πληροφορίες και διορατικότητα. Οι αποθήκες δεδομένων προσφέρουν έναν αποτελεσματικό τρόπο υποστήριξης ερωτημάτων, αναλύσεων, αναφορών, καθώς και παροχής προβλέψεων και τάσεων με βάση τα δεδομένα που συλλέγονται. Ο σχεδιασμός και ο καθαρισμός δεδομένων πρέπει να υποστηρίζονται από τη σωστή

αποθήκευση. Κανονικά, οι αποθήκες δεδομένων εξαρτώνται από μεγάλες χωρητικότητες αποθήκευσης που είναι ισχυρές, έχουν χαμηλότερο κόστος και καλές αποδόσεις.

- Αποθήκευση σε Cloud (Cloud Storage)

Με την αποθήκευση σε cloud, τα δεδομένα και οι πληροφορίες αποθηκεύονται ηλεκτρονικά στο διαδίκτυο. Με αυτήν την προσέγγιση, παρέχεται εύκολη προσβασιμότητα και χρηστικότητα, καθώς δεν είναι απαραίτητη η άμεση συνδεδεμένη πρόσβαση σε έναν σκληρό δίσκο ή σε έναν συγκεκριμένο ηλεκτρονικό υπολογιστή. Όταν θέλετε να αποκτήσετε πρόσβαση στα δεδομένα σας στο cloud, το μόνο που χρειάζεται είναι μία συσκευή με σύνδεση στο διαδίκτυο και η εισαγωγή των διαπιστευτηρίων που παρέχουν την πρόσβαση.

Επιπλέον, το cloud παρέχει όχι μόνο άμεσα διαθέσιμη υποδομή, αλλά και τη δυνατότητα να κλιμακώνεται γρήγορα αυτή η υποδομή για να διαχειριστεί μεγάλες αυξήσεις δεδομένων.

Η αποθήκευση στο cloud έχει βελτιώσει σημαντικά την παραγωγικότητα και την αποτελεσματικότητα των επιχειρήσεων, καθώς οι εργαζόμενοι μπορούν να μοιράζονται, να έχουν πρόσβαση και να επεξεργάζονται άμεσα αρχεία από απόσταση.

Εκτός από τα προηγούμενα οφέλη, η αποθήκευση στο cloud είναι επίσης σημαντικά φθηνότερη από τη φυσική αποθήκευση δεδομένων. Οι αποθήκες δεδομένων καταναλώνουν μεγάλες ποσότητες ενέργειας, χώρου, πόρων και συνεπάγονται μεγαλύτερο κίνδυνο. Ωστόσο, με την αποθήκευση σε cloud, εξοικονομείται ένα σημαντικό ποσό κόστους.

1.6 Παραδείγματα Μεγάλων Δεδομένων

Σε αυτήν την εποχή όπου κάθε πτυχή της καθημερινής μας ζωής περιστρέφεται γύρω από gadgets, όπως «έξυπνα» κινητά (smart phones), ρολόγια (smart watches), κάμερες (smart cameras), υπάρχει ένας τεράστιος όγκος δεδομένων που προέρχεται από διάφορες ψηφιακές πηγές.

Ορισμένοι βασικοί κλάδοι στους οποίους χρησιμοποιούνται και αξιοποιούνται κατάλληλα τα μεγάλα δεδομένα είναι:

- Στην Κυβέρνηση

Οι κυβερνήσεις αντιμετωπίζουν έναν εκτεταμένο όγκο δεδομένων σε καθημερινή βάση. Αυτό οφείλεται στις εις βάθος ενημερώσεις που πρέπει να διατηρούν για τα διαφορετικά αρχεία και τις βάσεις δεδομένων των πολιτών τους, την ανάπτυξή τους, τις γεωγραφικές έρευνες, τους ενεργειακούς πόρους κ.λπ. Τα δεδομένα αυτά καλούνται να εξεταστούν και να μελετηθούν, για την καλύτερη λειτουργία της κυβέρνησης. Τα δεδομένα αυτά χρησιμοποιούνται σε δύο κυρίως τομείς, στα προγράμματα πρόνοιας και στην περίπτωση της κυβερνοασφάλειας.

- Στην Υγεία

Τα μεγάλα δεδομένα διαδραματίζουν αναπόσπαστο ρόλο στην ενίσχυση των σύγχρονων λειτουργιών υγειονομικής περίθαλψης. Ορισμένες εφαρμογές της τεχνολογίας είναι στη χρήση Ηλεκτρονικών Μητρώων Υγείας (EHR), στην πρόβλεψη επιδημικών εστιών, στην αποφυγή ασθενειών που μπορούν να προληφθούν, στη βελτίωση της ποιότητας ζωής, καθώς ακόμη και στον αποτελεσματικό στρατηγικό σχεδιασμό για τη μείωση των απατών και των λαθών.

- Στην εκπαίδευση

Όσον αφορά τον κλάδο της εκπαίδευσης, συλλέγεται μεγάλος όγκος δεδομένων από τους φοιτητές, τους καθηγητές και τα μαθήματα. Η ανάλυση αυτών των δεδομένων μπορεί να δημιουργήσει πληροφορίες αποτελεσματικές για τη βελτίωση της μάθησης, τη βοήθεια των φοιτητών στον καθορισμό στόχων σταδιοδρομίας και γενικά της λειτουργίας των εκπαιδευτικών ιδρυμάτων.

- Στην Τραπεζική

Τα μεγάλα δεδομένα έχουν κάνει τις τράπεζες πιο αποτελεσματικές σε κάθε κλάδο. Η εφαρμογή της τεχνολογίας έχει απλοποιήσει την επεξεργασία των συναλλαγών, έχει βελτιώσει την κατανόηση των πελατών και έχει συνεισφέρει στον εντοπισμό της απάτης.

- Στα MME

Τα μεγάλα δεδομένα έχουν ανοίξει με επιτυχία το δρόμο στον κλάδο της διαφήμισης και των MME, τη στιγμή που οι σύγχρονες προσεγγίσεις κατανάλωσης περιεχομένου στο διαδίκτυο γίνονται η νέα τάση. Σήμερα, περισσότερο από ποτέ, μία διαφήμιση δεν εμφανίζεται σε ένα τυχαίο κοινό, αλλά σε κοινό που πιθανόν ενδιαφέρεται για το προϊόν που προβάλλεται. Επομένως, τα μεγάλα δεδομένα συνεισφέρουν στην πρόβλεψη του τι θέλει το κοινό, ανάλογα με την ηλικιακή του ομάδα, προσφέρουν πληροφορίες σχετικά με την απόσυρση πελατών και τη βελτιστοποίηση του προγράμματος ροής. Εξαιρετικό παράδειγμα αποτελεί η πλατφόρμα του Netflix, η οποία μέσω του ιστορικού προβολής του χρήστη, καθώς και συγκεκριμένων

σημείων στα οποία έχει σταματήσει ένα βίντεο επηρεάζει τα πάντα, από τις προσαρμοσμένες μικρογραφίες έως τα περιεχόμενα στην ενότητα «Δημοφιλή στο Netflix».

- Στο Marketing

Στο μάρκετινγκ, τα μεγάλα δεδομένα περιλαμβάνουν τη συλλογή, την ανάλυση και τη χρήση τεράστιων ποσοτήτων ψηφιακών πληροφοριών για τη βελτίωση των επιχειρηματικών λειτουργιών. Τέτοιες λειτουργίες είναι η ενίσχυση της απόκτησης και διατήρησης των πελατών (Acquisition & Retention), η έννοια του «γνωρίστε τον πελάτη σας» (KYC), η βελτιστοποίηση των καμπανιών και των προϊόντων, η αύξηση των εσόδων, η πιο στοχευμένη διαφήμιση και ο καλύτερος έλεγχος της επιχείρησης πραγματοποιώντας διαφορετικές δοκιμές. Η ανάλυση των μεγάλων δεδομένων και η επιχειρηματική ευφυΐα οδηγούν σε εξοικονόμηση χρόνου και κόστους, βελτιστοποιώντας την απόδοση του μάρκετινγκ και της επιχείρησης γενικότερα.

Κεφάλαιο 2: Εξόρυξη Γνώσης και Δεδομένων

2.1 Ορισμός Εξόρυξης Γνώσης και Δεδομένων

Η εξόρυξη γνώσης από δεδομένα (data mining) είναι μια σημαντική περιοχή δράσης των επιχειρήσεων, που τις βοηθάει να εστιάσουν στις πληροφορίες που βρίσκονται συγκεντρωμένες στις αποθήκες δεδομένων τους (data warehouses). Πρόκειται για μια διαδικασία ανάλυσης των δεδομένων και την εξαγωγή πολύτιμων πληροφοριών. Σκοπός της εξόρυξης δεδομένων είναι η αναζήτηση και ο εντοπισμός κρυφών μοτίβων (patterns) σε ομάδες δεδομένων και η ανακάλυψη σχέσεων που υπάρχουν μεταξύ αυτών. Αρχικά συλλέγονται τα δεδομένα και ακολουθεί η επιλογή του κατάλληλου αλγορίθμου για την αποκάλυψη τάσεων και συσχετίσεων για τη διευκόλυνση της έγκαιρης και αποτελεσματικής λήψης επιχειρηματικών αποφάσεων.

Σήμερα, οι επιχειρήσεις έχουν συνειδητοποιήσει ότι οι σχέσεις με τους πελάτες είναι πολύ σημαντικός παράγοντας για την επιτυχία τους. Η διαχείριση των σχέσεων των πελατών (Customer Relationship Management - CRM) είναι μια στρατηγική που μπορεί να βοηθήσει τις επιχειρήσεις να αποφύγουν την απώλεια πελατών, χτίζοντας μακροχρόνιες σχέσεις και αυξάνοντας τα κέρδη τους. Το CRM δεν εστιάζει στην απόκτηση νέων πελατών, αλλά στην παραμονή των υπαρχόντων. Το ζητούμενο αυτό πρέπει να επιτυγχάνεται βρίσκοντας τη βέλτιστη λύση με ταχύτητα, ευελιξία και όσο το δυνατόν χαμηλότερο κόστος. Προκειμένου να διασφαλιστεί η καλύτερη σχέση με τους πελάτες, είναι απαραίτητο να υπάρχουν οι κατάλληλες πληροφορίες και να είναι σωστά ομαδοποιημένες.

Η ανάπτυξη της τεχνολογίας πληροφοριών και επικοινωνιών (Information and communications technology - ICT) προσφέρει μια επιτυχημένη λύση σε αυτά τα προβλήματα. Η Επιχειρηματική ευφυΐα (Business Intelligence - BI) θεωρείται ένας ξεχωριστός κλάδος που συνδυάζει στοιχεία της τεχνολογίας των πληροφοριών, της στρατηγικής, της λογιστικής, της εταιρικής ανάλυσης και του μάρκετινγκ. Επιτρέπει τη συλλογή, την ανάλυση, τη μετάδοση και την αξιοποίηση της επιχειρηματικής πληροφορίας, με στόχο τη διευκόλυνση της επίλυσης των προβλημάτων διαχείρισης και τη λήψη καλύτερων επιχειρηματικών αποφάσεων.

Οι αποθήκες δεδομένων αποτελούν έναν τύπο βάσεων δεδομένων και δημιουργήθηκαν για να καλύψουν τις ανάγκες τέτοιων συστημάτων, καθώς τα δεδομένα είναι οργανωμένα με πρακτικό τρόπο για τη διεξαγωγή διαδικασιών ανάλυσης. Μία αποθήκη δεδομένων περιέχει αντίγραφα δεδομένων που απομονώνονται από επιχειρησιακές βάσεις και δομούνται ειδικά για αναφορές (reports) και αναλύσεις. Οι αποθήκες δεδομένων και η ηλεκτρονική αναλυτική επεξεργασία (OnLine Analytical Processing - OLAP) αποτελούν τη βάση για την εφαρμογή της επιχειρηματικής ευφυΐας.

Η εξόρυξη δεδομένων αποτελεί επίσης σημαντικό παράγοντα της επιχειρηματικής ευφυΐας καθώς περιλαμβάνει ένα ευρύ πεδίο υπολογιστικών μεθόδων που περιλαμβάνουν την στατιστική ανάλυση (statistical analysis), τα δένδρα αποφάσεων (decision trees), τα νευρωνικά δίκτυα (neural networks), την εξαγωγή κανόνων (rule induction) και την γραφική οπτικοποίηση (graphic visualization).

2.2 Εξόρυξη Δεδομένων και Ανεύρεση Γνώσης

Η εξόρυξη δεδομένων ορίζεται ως η εξαγωγή πληροφοριών από ένα τεράστιο σύνολο δεδομένων. Με άλλα λόγια, μπορούμε να πούμε ότι η εξόρυξη δεδομένων είναι η εξόρυξη της γνώσης από τα δεδομένα.

Στη βιβλιογραφία υπάρχει μια σύγχυση ανάμεσα στους όρους «Εξόρυξη Γνώσης» (Data mining) και «Ανεύρεση γνώσης στις βάσεις δεδομένων» (Knowledge discovery in data bases - KDD). Αν και σε ορισμένες περιπτώσεις αυτοί οι δύο όροι ταυτίζονται, στην πραγματικότητα η εξόρυξη δεδομένων αποτελεί τμήμα της ανεύρεσης γνώσης, αποτελώντας μάλιστα τον πυρήνα της.

Ο όρος KDD αναφέρεται στη διαδικασία ανακάλυψης γνώσης στα δεδομένα και δίνει έμφαση στις εφαρμογές υψηλού επιπέδου συγκεκριμένων τεχνικών Εξόρυξης Δεδομένων. Είναι ένα πεδίο ενδιαφέροντος για ερευνητές σε διάφορους τομείς, όπως η τεχνητή νοημοσύνη, η μηχανική μάθηση, η αναγνώριση προτύπων, οι βάσεις δεδομένων, τα στατιστικά στοιχεία, η απόκτηση γνώσεων για έμπειρα συστήματα και η οπτικοποίηση δεδομένων.

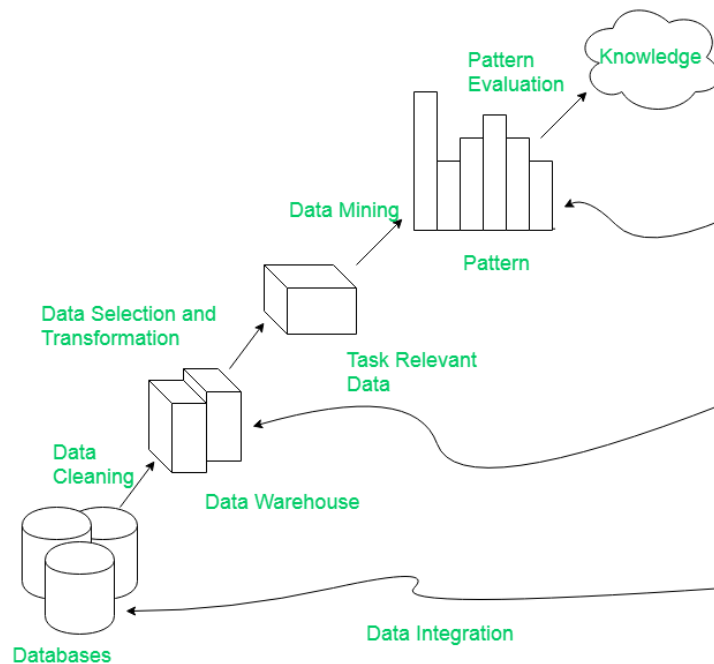
Η διαδικασία ανεύρεσης γνώσης αποτελείται από τα ακόλουθα βήματα:

1. Καθαρισμός δεδομένων (Data cleaning)
Σε αυτό το βήμα, βελτιώνεται η αξιοπιστία των δεδομένων. Περιλαμβάνει την εκκαθάριση των δεδομένων, τον χειρισμό των τιμών που λείπουν και την αφαίρεση θορύβου ή ακραίων τιμών.
2. Ενσωμάτωση δεδομένων (Data integration)
Η ενοποίηση δεδομένων είναι μια τεχνική προεπεξεργασίας δεδομένων που συγχωνεύει τα δεδομένα από πολλαπλές ετερογενείς πηγές δεδομένων σε ένα συνεκτικό χώρο αποθήκευσης, δηλαδή σε μία κοινή βάση δεδομένων.
3. Επιλογή δεδομένων (Data selection)
Είναι η διαδικασία κατά την οποία επιλέγονται τα δεδομένα που σχετίζονται με την ανάλυση που θα ακολουθήσει και είναι χρήσιμα για αυτήν.
4. Τροποποίηση δεδομένων (Data transformation)
Ο μετασχηματισμός δεδομένων ορίζεται ως η διαδικασία μετατροπής των δεδομένων που έχουν επιλεγεί στην κατάλληλη μορφή που απαιτείται από τη διαδικασία εξόρυξης.
5. Εξόρυξη δεδομένων (Data mining)
Η εξόρυξη δεδομένων αποτελεί το σημαντικότερο βήμα. Σε αυτό το στάδιο, εφαρμόζονται διάφορες μέθοδοι με στόχο την εξαγωγή χρήσιμων προτύπων.
6. Αξιολόγηση προτύπων (Pattern evaluation)
Στο βήμα αυτό, αξιολογούνται και ερμηνεύονται τα πρότυπα που εξήχθησαν βάσει συγκεκριμένων μέτρων αξιολόγησης (evaluation measures).

7. Αναπαράσταση γνώσης (Knowledge representation)

Η αναπαράσταση γνώσης ορίζεται ως τεχνική που χρησιμοποιεί εργαλεία οπτικοποίησης για την αναπαράσταση των αποτελεσμάτων εξόρυξης δεδομένων.

Αξίζει να σημειωθεί ότι, καθώς η διαδικασία ανεύρεσης γνώσης είναι μία επαναληπτική διαδικασία, ο χρήστης έχει τη δυνατότητα να τροποποιήσει και να βελτιώσει τα μέτρα αξιολόγησης, να τελειοποιήσει τη διαδικασία εξόρυξης, να μετασχηματίσει τα υπάρχοντα δεδομένα ή να ενσωματώσει νέα, προκειμένου να ληφθούν διαφορετικά και πιο κατάλληλα αποτελέσματα.



Εικόνα 4 - Στάδια της διαδικασίας Ανεύρεσης Γνώσης
KDD Process in Data Mining – GeeksforGeeks

2.3 Στόχοι της Εξόρυξης Δεδομένων

Η εξόρυξη δεδομένων είναι ένας καινοτόμος τρόπος για να αποκτηθούν νέες και χρήσιμες επιχειρηματικές γνώσεις αναλύοντας τις πληροφορίες που βρίσκονται στη βάση δεδομένων μιας εταιρείας. Αυτές οι πληροφορίες υποστηρίζουν και διευκολύνουν τη λήψη βέλτιστων επιχειρηματικών αποφάσεων. Ουσιαστικά, η εξόρυξη δεδομένων είναι ένας πρωτοποριακός τρόπος αξιοποίησης των πληροφοριών που έχει ήδη η εταιρεία προκειμένου να βελτιώσει τις διαδικασίες, να αυξήσει την απόδοση της ή να βελτιστοποιήσει τη χρήση των πόρων.

Με την εξόρυξη δεδομένων, παρέχονται πληροφορίες που μπορούν να οδηγήσουν σε τεράστια βελτίωση στην ποιότητα και την αξιοπιστία της λήψης επιχειρηματικών αποφάσεων, σε αντίθεση με τις συμβατικές μεθόδους αναζήτησης.

Οι συμβατικές μέθοδοι μπορούν να πουν σε μια τράπεζα, για παράδειγμα, ποιος από τους τύπους τραπεζικών λογαριασμών που παρέχει είναι ο πιο κερδοφόρος. Ωστόσο, μόνο η εξόρυξη δεδομένων επιτρέπει στην τράπεζα να δημιουργεί προφίλ πελατών που έχουν ήδη αυτόν τον τύπο λογαριασμού. Στη συνέχεια, η τράπεζα μπορεί να χρησιμοποιήσει την εξόρυξη

δεδομένων για να βρει άλλους πελάτες που ταιριάζουν με αυτό το προφίλ, ώστε να μπορεί να στοχεύσει με ακρίβεια μια καμπάνια μάρκετινγκ σε αυτούς.

Η εξόρυξη δεδομένων μπορεί να εντοπίσει μοτίβα στα δεδομένα της εταιρείας, για παράδειγμα, στα αρχεία αγορών σούπερ μάρκετ. Εάν, για παράδειγμα, οι πελάτες αγοράσουν το προϊόν Α και το προϊόν Β, ποιο προϊόν Γ είναι πιο πιθανό να αγοράσουν επίσης; Οι ακριβείς απαντήσεις σε ερωτήσεις όπως αυτή είναι πολύτιμα βοηθήματα στις στρατηγικές μάρκετινγκ.

Τα εργαλεία εξόρυξης δεδομένων διευκολύνουν και αυτοματοποιούν τη διαδικασία ανακάλυψης αυτού του είδους πληροφοριών σε μεγάλες βάσεις δεδομένων.

Οι βασικοί στόχοι της εξόρυξης δεδομένων είναι:

1. Πρόβλεψη (Prediction)

Οι διαδικασίες πρόβλεψης της εξόρυξης δεδομένων προσπαθούν να κάνουν εκτιμήσεις αξιοποιώντας τα διαθέσιμα δεδομένα. Η πρόβλεψη μελλοντικών συμπεριφορών έχει ως στόχο να ληφθούν αποφάσεις οι οποίες οδηγούν στη μεγιστοποίηση του κέρδους αλλά και στην αποφυγή δυσάρεστων καταστάσεων. Τα αποτελέσματα αυτά μπορεί να είναι πληροφορίες σχετικές με το ύψος των πωλήσεων ενός καταστήματος για μια δεδομένη χρονική περίοδο, αλλά και αν η κατάργηση κάποιου προϊόντος θα είχε θετική επίδραση στις πωλήσεις.

2. Αναγνώριση (Identification)

Σε αυτό το στάδιο γίνεται ο προσδιορισμός μοτίβων στα δεδομένα. Για παράδειγμα, τα νιόπαντρα ζευγάρια τείνουν να ξοδεύουν περισσότερα χρήματα αγοράζοντας έπιπλα.

3. Ταξινόμηση (Classification)

Σε αυτή τη φάση τα δεδομένα διαχωρίζονται σε κλάσεις. Για παράδειγμα, οι πελάτες ενός σούπερ μάρκετ μπορούν να ταξινομηθούν σε διαφορετικές κατηγορίες με βάση τη συμπεριφορά τους στις αγορές. Κατά την ανάλυση, η κατηγοριοποίηση αυτή μπορεί να βοηθήσει στη λήψη καλύτερων αποφάσεων, ώστε να προσελκύονται περισσότεροι πελάτες.

4. Βελτιστοποίηση (Optimization)

Η βέλτιστη χρήση περιορισμένων πόρων αποτελεί επίσης στόχο της εξόρυξης δεδομένων. Τέτοιοι πόροι είναι ο χρόνος, ο χώρος, τα χρήματα και η μεγιστοποίηση κάποιων μεγεθών, όπως τα κέρδη και οι πωλήσεις. Για παράδειγμα, πώς να χρησιμοποιηθεί καλύτερα η διαφήμιση για την αύξηση των πωλήσεων.

2.4 Που εφαρμόζεται η Εξόρυξη Γνώσης;

Στις μέρες μας συγκεντρώνονται μεγάλες ποσότητες δεδομένων. Ο όγκος των δεδομένων που συλλέγονται σχεδόν διπλασιάζεται κάθε χρόνο. Για την εξαγωγή δεδομένων ή την αναζήτηση γνώσης από αυτά τα τεράστια δεδομένα, χρησιμοποιούνται τεχνικές εξόρυξης δεδομένων. Η εξόρυξη δεδομένων χρησιμοποιείται σχεδόν σε όλα τα μέρη όπου αποθηκεύεται και υφίσταται επεξεργασία μεγάλος όγκος δεδομένων.

Διάφοροι τομείς στους οποίους βρίσκει εφαρμογή η εξόρυξη δεδομένων είναι:

1. Χρηματοοικονομικός τομέας

Ο ανταγωνισμός μεταξύ των επιχειρήσεων στον χρηματοοικονομικό τομέα αυξάνεται διαρκώς. Οι επιχειρήσεις μπορούν να χρησιμοποιήσουν αυτοματοποιημένους αλγόριθμους για να κατανοήσουν καλύτερα τους πελάτες τους καθώς και τις δεσεκατομύρια συναλλαγές που αποτελούν το χρηματοπιστωτικό σύστημα. Οι εταιρείες χρηματοοικονομικών υπηρεσιών μπορούν να χρησιμοποιήσουν την εξόρυξη δεδομένων για να κατανοήσουν καλύτερα τους κινδύνους της αγοράς, να εντοπίσουν πιο γρήγορα την απάτη, να διαχειριστούν τους κανονισμούς και τις υποχρεώσεις συμμόρφωσης και να μεγιστοποιήσουν την απόδοση των επενδύσεών τους στο μάρκετινγκ.

2. Ασφάλειες

Οι ασφαλιστικές εταιρείες μπορούν να επιλύσουν πολύπλοκα προβλήματα όπως η απάτη, η συμμόρφωση, η διαχείριση κινδύνου και η απώλεια πελατών χρησιμοποιώντας αναλυτική τεχνογνωσία. Οι εταιρείες έχουν χρησιμοποιήσει τεχνικές εξόρυξης δεδομένων για καλύτερες τιμές των προϊόντων σε όλους τους επιχειρηματικούς τομείς και ανακαλύπτουν νέους τρόπους για να προσφέρουν ανταγωνιστικά προϊόντα στην υπάρχουσα πελατειακή τους βάση.

3. Εκπαίδευση

Η εξόρυξη δεδομένων επιτρέπει στους εκπαιδευτικούς να αποκτήσουν πρόσβαση στα δεδομένα των μαθητών, να προβλέψουν τα επίπεδα επιτυχίας και να εντοπίσουν μαθητές που χρειάζονται πρόσθετη υποστήριξη. Επιπλέον, χρησιμοποιείται για να κατευθύνει τους μαθητές στην κατάλληλη επιλογή σπουδών και επαγγέλματος, σύμφωνα με τα δημογραφικά και προσωπικά χαρακτηριστικά τους.

4. Τηλεπικοινωνίες, Μέσα & Τεχνολογία

Σε μια πολυσύχναστη αγορά με έντονο ανταγωνισμό, οι απαντήσεις βρίσκονται συχνά στα δεδομένα των πελατών. Τα αναλυτικά μοντέλα μπορούν να βοηθήσουν τις εταιρείες τηλεπικοινωνιών, μέσων και τεχνολογίας να κατανοήσουν τα δεδομένα τους, επιτρέποντάς τους να κατηγοριοποιούν τις συνήθειες και τις απαιτήσεις των πελατών τους, ώστε να προβλέπουν τη συμπεριφορά τους και να προσφέρουν εξαιρετικά στοχευμένες και σχετικές καμπάνιες.

5. Λιανικό εμπόριο

Το λιανικό εμπόριο είναι ένας αρκετά ανταγωνιστικός κλάδος. Η εξόρυξη γνώσης μπορεί να κάνει εφικτή μια προσωποποιημένη σχέση με κάθε πελάτη χωριστά. Έτσι, τέτοιοι είδους πληροφορίες μπορούν να βελτιστοποιήσουν τις πωλήσεις, καθώς οι έμποροι λιανικής μπορούν να προσφέρουν πιο στοχευμένες καμπάνιες και να βρουν την προσφορά που έχει τον μεγαλύτερο αντίκτυπο στους πελάτες.

6. Υγεία

Στον χώρο της υγείας συλλέγεται και αποθηκεύεται καθημερινά μεγάλος όγκος δεδομένων από καρτέλες ασθενώς, διάφορες ασθένειες, δοκιμές νέων φαρμάκων και πολλά άλλα. Η εξόρυξη δεδομένων στην υγειονομική περίθαλψη έχει αποδειχθεί αποτελεσματική σε τομείς όπως η προγνωστική ιατρική, η διαχείριση της υγειονομικής περίθαλψης και η μέτρηση της αποτελεσματικότητας ορισμένων θεραπειών.

2.5 Πλεονεκτήματα και Μειονεκτήματα της Εξόρυξης Γνώσης

Η εξόρυξη δεδομένων είναι η διαδικασία εξαγωγής προτύπων και τάσεων από μεγάλες ποσότητες δεδομένων. Χρησιμοποιείται για διάφορους σκοπούς, όπως για τη βελτίωση της εμπειρίας των πελατών, την αύξηση της κερδοφορίας και τη μείωση των κινδύνων. Παρόλο που η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί για να αναλύσει τα δεδομένα σας και να σας δώσει μερικές πολύτιμες πληροφορίες, μπορεί επίσης να κάνει κατάχρηση των δεδομένων και να παραβιάσει το απόρρητο.

Πλεονεκτήματα:

1. Η εξόρυξη δεδομένων βοηθά τις επιχειρήσεις να κάνουν επικερδείς παραγωγικές και λειτουργικές προσαρμογές. Η εξόρυξη δεδομένων μπορεί να χρησιμοποιηθεί για την εύρεση συσχετισμών μεταξύ προϊόντων, καταναλωτών, προμηθευτών και άλλων πτυχών της επιχείρησης. Αυτό μπορεί να βοηθήσει μια εταιρεία να εντοπίσει τάσεις που μπορεί να μην είχαν εντοπιστεί πριν, ή τουλάχιστον να τη βοηθήσει να κάνει πιο ακριβείς προβλέψεις.
2. Ανεξάρτητα από τον όγκο των πληροφοριών, το λογισμικό εξόρυξης δεδομένων τις λαμβάνει όλες και τις επεξεργάζεται σε σύντομο χρονικό διάστημα. Στο παρελθόν, η επεξεργασία τέτοιων τεράστιων ποσοτήτων δεδομένων ήταν αδύνατη, επομένως ήταν πολύ δύσκολο να κατανοηθούν τα δεδομένα. Ωστόσο σήμερα η επεξεργασία των δεδομένων πραγματοποιείται από κάθε εταιρεία σε καθημερινή βάση.
3. Η εξόρυξη δεδομένων χρησιμοποιείται, κατά κύριο λόγο, για την εύρεση κρυφών μοτίβων και τάσεων στη συμπεριφορά των χρηστών. Αυτό το κάνει, ελέγχοντας τη συχνότητα του πόσο συχνά συμβαίνει κάτι, και με αυτόν τον τρόπο μπορούν να σχεδιαστούν διαφορετικές στρατηγικές. Αυτές οι πληροφορίες μπορούν να χρησιμοποιηθούν για την κατανόηση των τάσεων, την ανακάλυψη κρυφών μοτίβων και την πρόταση στρατηγικών που μπορούν να δοκιμάσουν οι επιχειρήσεις.
4. Η εξόρυξη δεδομένων μπορεί να βοηθήσει στον εντοπισμό κινδύνων και απάτης, που ενδέχεται να μην είναι ανιχνεύσιμα μέσω των παραδοσιακών μέσων ανάλυσης δεδομένων. Με αυτόν τον τρόπο, επιτρέπει τον διαχωρισμό του αυθεντικού υλικού από το επικίνδυνο. Για παράδειγμα, ο διακομιστής email διαχωρίζει τα email που έχουν σταλεί από ασαφή αναγνωριστικά και τα μεταφέρει στα ανεπιθύμητα (spam) ώστε να προστατευθούμε από εξαπάτηση.

Μειονεκτήματα:

1. Η ανάλυση δεδομένων είναι μια περίπλοκη διαδικασία και συχνά απαιτεί από άτομα με εκπαίδευση και εξειδίκευση να χρησιμοποιούν τα εργαλεία. Αυτή η δυσκολία για ανάλυση δεδομένων μπορεί να αποθαρρύνει τις μικρές επιχειρήσεις από τη χρήση αυτής της τεχνολογίας.
2. Η εξόρυξη δεδομένων δεν παρέχει πάντα ακριβείς πληροφορίες, επομένως δε μπορούμε να βασιστούμε εξ ολοκλήρου στα αποτελέσματά της. Επιπλέον, πρέπει να σκεφτούμε ότι είναι

μόνο ένα μηχάνημα που λαμβάνει αποφάσεις, επομένως προφανώς, δεν θα λάμβανε υπόψη το συναίσθημα των ανθρώπων που θα επηρεάσει η απόφαση που θα λάβει.

3. Ένα από τα σημαντικότερα μειονεκτήματα της εξόρυξης δεδομένων είναι τα δεδομένα και τα ζητήματα απορρήτου. Υπήρξαν πολλές περιπτώσεις σε πολλούς διαφορετικούς ιστοτόπους και εφαρμογές επειδή συνέλεξαν και έκαναν κακή χρήση των πληροφοριών των χρηστών τους. Ορισμένοι προγραμματιστές χρησιμοποιούν τα δεδομένα που συλλέγονται και τα πωλούν στη μαύρη αγορά στο διαδίκτυο, και αν όχι, τα χρησιμοποιούν για να εκβιάσουν άτομα ή να διαπράξουν κλοπή ταυτότητας.
4. Η εξόρυξη δεδομένων είναι μια πολύ δαπανηρή διαδικασία. Οι εταιρείες πρέπει να προσλάβουν υπαλλήλους και ειδικούς τεχνολογίας για να διασφαλίσουν ότι η εξόρυξη δεδομένων πραγματοποιείται σωστά. Επίσης, πρέπει να επενδύσουν σε προηγμένο λογισμικό εξόρυξης δεδομένων, το οποίο μπορεί να είναι ακριβό. Για τις μικρές επιχειρήσεις, που δεν παράγουν αρκετές πολύτιμες πληροφορίες, το κόστος της εξόρυξης δεδομένων υπερτερεί των οφελών.
5. Ορισμένες αποφάσεις που προκύπτουν από την εξόρυξη δεδομένων μπορεί να είναι ηθικά λανθασμένες, επομένως δε μπορούν να χρησιμοποιηθούν για σκοπούς μάρκετινγκ. Για παράδειγμα, η χρήση της ιατρικής κατάστασης κάποιου ασθενή για την πώληση ενός προϊόντος δεν αποτελεί ηθική πρακτική.

Κεφάλαιο 3: Διαχείριση Πελατειακών Σχέσεων (CRM)

3.1 Εισαγωγή

Η ανάγκη για περαιτέρω βελτιωμένη διαχείριση των σχέσεων εταιρείας – πελατών έχει ως αποτέλεσμα την ανάπτυξη συστημάτων διαχείρισης πελατειακών σχέσεων (CRM). Σε αυτό το κεφάλαιο, δίνεται η ερμηνεία και ο ορισμός της διαχείρισης πελατειακών σχέσεων. Επίσης, περιγράφονται οι τέσσερις διαστάσεις του CRM. Στη συνέχεια, πραγματοποιείται ο διαχωρισμός του CRM σε Λειτουργικό, Συνεργατικό και Αναλυτικό, τρία υποσυστήματα τα οποία είναι απαραίτητα για την επιχείρηση τόσο μεμονωμένα, όσο και συνδεδεμένα μεταξύ τους.

3.2 Έννοια και Ορισμοί

Η διαχείριση πελατειακών σχέσεων αποτελεί μια επιχειρηματική προσέγγιση με στόχο την κατανόηση, τη γνώση και την επίδραση στη συμπεριφορά του πελάτη μέσω τεχνικών επικοινωνίας, ώστε να βελτιωθεί η απόκτηση νέων πελατών (acquisition), η διατήρηση υπαρχόντων πελατών (retention), η βελτίωση της εικόνας της επιχείρησης στον πελάτη (loyalty) και επομένως η κερδοφορία της επιχείρησης.

Παρόλο που το CRM είναι ευρέως αναγνωρισμένο ως ένα από τα πιο σημαντικά εργαλεία των επιχειρήσεων, δεν υπάρχει ένας γενικά αποδεκτός ορισμός. Διάφοροι ορισμοί που έχουν δοθεί είναι:

Κατά τον Bose (2002), το CRM αποτελεί την ενοποίηση τεχνολογιών και επιχειρηματικών διαδικασιών που χρησιμοποιούνται για την ικανοποίηση των αναγκών ενός πελάτη κατά τη διάρκεια κάθε δεδομένης αλληλεπίδρασης μεταξύ πελάτη και επιχείρησης. Το CRM περιλαμβάνει την απόκτηση, ανάλυση και χρήση γνώσεων σχετικά με έναν πελάτη προκειμένου να μεγιστοποιηθούν και να γίνουν πιο αποτελεσματικές οι πωλήσεις αγαθών ή υπηρεσιών.

Οι Day και Van Den Bulte (2002) όρισαν το CRM ως μια διαλειτουργική διαδικασία για την επίτευξη συνεχούς διαλόγου με τους πελάτες, σε όλα τα σημεία επαφής και πρόσβασής τους, με εξατομικευμένη διαχείριση των πιο πολύτιμων πελατών, για την αύξηση της διατήρησης των πελατών και της αποτελεσματικότητας των πρωτοβουλιών του marketing.

Κατά τον Ko (2004), το CRM ορίζεται ως η ολοκληρωμένη στρατηγική διαχείρισης πελατών μιας επιχείρησης για την αποτελεσματική διαχείριση των πελατών παρέχοντας εξατομικευμένα αγαθά και υπηρεσίες και μεγιστοποιώντας τις αξίες διάρκειας ζωής των πελατών (customers' lifetime value).

Σύμφωνα με τους Berger και Bechwati (2000), στο επίκεντρο της διαχείρισης πελατειακών σχέσεων βρίσκεται η ανάπτυξη και η διατήρηση των μακροχρόνιων σχέσεων με τους πελάτες, και όχι μόνο μία σειρά από ξεχωριστές συναλλαγές. Υποστηρίζουν, ακόμα, ότι η επιχείρηση θα πρέπει να εξετάσει την αξία του πελάτη σύμφωνα με τη συνολική του σχέση με την εταιρεία, και όχι με το κέρδος από κάθε ξεχωριστή συναλλαγή. Από τις πιο σημαντικές λειτουργίες του

marketing είναι η συλλογή και η συσσώρευση πληροφοριών για τους πελάτες και η παροχή υπηρεσιών σε αυτούς. Η συλλογή των απαιτούμενων πληροφοριών καθίσταται δύσκολη, εάν η εταιρεία δεν εκμεταλλευθεί κατάλληλα την τεχνολογία των πληροφοριών (Information Technology). Οι ραγδαίες εξελίξεις της τεχνολογίας, όπως οι βάσεις δεδομένων, η αποθήκευση των δεδομένων (data warehouses) και η εξόρυξη δεδομένων παίζουν κυρίαρχο ρόλο στις νέες στρατηγικές marketing (Ahn et al. 2003).

3.3 Οι τέσσερις διαστάσεις του CRM

Η χρήση ενός συστήματος διαχείρισης πελατών (CRM) έχει πολλά οφέλη, τόσο για τον πελάτη, όσο και για την ίδια την επιχείρηση.

Το CRM αποτελείται από τέσσερις διαστάσεις:

1. Εντοπισμός πελατών (Customer Identification)

Το CRM ξεκινά με τον εντοπισμό πελατών, ο οποίος αναφέρεται ως απόκτηση πελάτη (Customer Acquisition). Αυτή η φάση περιλαμβάνει τον εντοπισμό των ατόμων που είναι πιθανότερο είτε να γίνουν πελάτες είτε να γίνουν οι πιο επικερδείς για την επιχείρηση. Επιπλέον, περιλαμβάνει την ανάλυση πελατών που έχουν στραφεί σε ανταγωνιστές και πώς μπορούν να ανακτηθούν πίσω. Σε αυτό το στάδιο, γίνεται και η τμηματοποίηση των πελατών σε μικρότερες ομάδες.

2. Προσέλκυση πελατών (Customer Attraction)

Αυτή είναι η φάση που ακολουθεί τον εντοπισμό των πελατών. Μετά τον εντοπισμό των τμημάτων (segments) των δυνητικών πελατών, οι επιχειρήσεις μπορούν να στραφούν στον εντοπισμό των κατάλληλων πόρων και στην προσπάθεια προσέλκυσης των πελατών – στόχων. Το άμεσο marketing, δηλαδή η διαδικασία προώθησης του προϊόντος ή της υπηρεσίας μέσω διαφόρων καναλιών, αποτελεί στοιχείο της προσέλκυσης των πελατών.

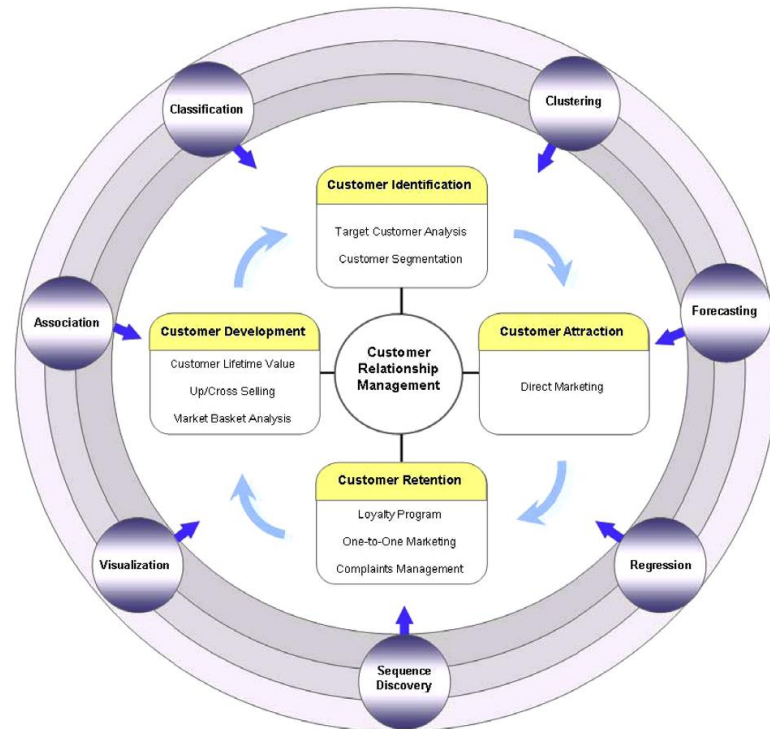
3. Διατήρηση πελατών (Customer Retention)

Αυτό είναι το κεντρικό μέλημα για ένα σύστημα CRM. Η ικανοποίηση του πελάτη αποτελεί τη βασική προϋπόθεση για τη διατήρησή του. Εργαλεία της διατήρησης πελατών αποτελούν το εξατομικευμένο marketing (one-to-one-marketing), τα προγράμματα πιστότητας (loyalty programs) και η διαχείριση παραπόνων των πελατών (customer complaints). Το εξατομικευμένο marketing αναφέρεται στις καμπάνιες που υποστηρίζονται από την ανάλυση της συμπεριφοράς των πελατών και την πρόβλεψη των αλλαγών σε αυτές. Τα προγράμματα πιστότητας περιλαμβάνουν τις διαφημιστικές καμπάνιες και τις δραστηριότητες που έχουν ως στόχο τη διατήρηση μιας μακροπρόθεσμης σχέσης με τους πελάτες.

4. Ανάπτυξη πελατών (Customer Development)

Στα στοιχεία ανάπτυξης πελατών περιλαμβάνονται η ανάλυση της αξίας του πελάτη κατά τη διάρκεια ζωής του (lifetime value analysis), οι σταυροειδείς πωλήσεις (up/cross selling) και η ανάλυση καλάθιού marketing (market basket analysis). Η ανάλυση της αξίας του πελάτη κατά τη διάρκεια ζωής του ορίζεται ως η πρόβλεψη των καθαρών εσόδων που αναμένεται να αποφέρει ο πελάτης στην εταιρεία. Οι σταυροειδείς πωλήσεις αφορούν τις

προωθητικές ενέργειες που έχουν ως στόχο τη διερεύνηση των συνδεδεμένων προϊόντων/υπηρεσιών που χρησιμοποιούνται από τον πελάτη. Η ανάλυση του καλαθιού marketing στοχεύει στη μεγιστοποίηση του αριθμού των συναλλαγών των πελατών και εν συνεχεία της αξίας τους, μέσω της καταναλωτικής τους συμπεριφοράς τους.



Εικόνα 5 - Διαστάσεις του CRM
Data Mining Techniques for Customer Relationship Management - ResearchGate

3.4 Τα είδη του CRM

Τα συστήματα CRM σε μια επιχείρηση έχουν διάφορες λειτουργίες και ικανοποιούν πολλούς στόχους. Το Λειτουργικό, το Συνεργατικό και το Αναλυτικό είναι τα τρία υποσυστήματα του CRM. Η ύπαρξη συνδέσμων ανάμεσα στα τρία αυτά υποσυστήματα είναι πολύ σημαντική για την ολοκλήρωση των δεδομένων και λειτουργιών.

- **Λειτουργικό CRM (Operational CRM)**

Οι επιχειρησιακές διαδικασίες "Front office" περιλαμβάνονται στο λειτουργικό CRM. Αυτές οι διαδικασίες περιλαμβάνουν όλη την επαφή με τους πελάτες (πωλήσεις, marketing, εξυπηρέτηση πελατών). Οι δραστηριότητες που έχουν με τους πελάτες καταγράφονται στο «ιστορικό των πελατών», ώστε να μπορούν να ανακτηθούν για περαιτέρω ανάλυση, όταν αυτό κριθεί απαραίτητο.

Ένα από τα σημαντικότερα οφέλη του Λειτουργικού CRM είναι ότι κάθε πελάτης μπορεί να επικοινωνεί με διαφορετικά άτομα ή μέσω διαφορετικών καναλιών με την επιχείρηση, χωρίς να χρειάζεται να εξηγήσει κάθε φορά όλο το ιστορικό των ενεργειών του. Σημαντικό πλεονέκτημα αποτελεί επίσης η ευκολότερη διαχείριση των προωθητικών ενεργειών

(campaigns), καθώς επίσης και η αυτοματοποίηση πολλών λειτουργιών marketing και πωλήσεων.

- **Αυτοματοποίηση πωλήσεων (Sales force automation – SFA)**
Τα εργαλεία SFA αυτοματοποιούν ορισμένες από τις σημαντικότερες πωλήσεις της επιχείρησης και τομέων της διοίκησης πωλήσεων, όπως για παράδειγμα το account management, το lead management, το επικοινωνιακό management, τα δημογραφικά και ψυχογραφικά χαρακτηριστικά, καθώς και την απόδοση του management. Τα εργαλεία αυτά έχουν σχεδιαστεί ώστε να συμβάλλουν στη βελτίωση την παραγωγικότητας των πωλήσεων.
- **Πελατειακές υπηρεσίες και υποστήριξη (Customer Service and support – CSS)**
Το CSS συμβάλλει στην αυτοματοποίηση ορισμένων υπηρεσιών παραπόνων, επιστροφής προϊόντων και αναζήτησης πληροφοριών. Ωστόσο, συνεχίζουν αν υπάρχουν παραδοσιακά εσωτερικά εργαλεία βοήθειας και παραδοσιακά τηλεφωνικά κέντρα υποστήριξης για τις ερωτήσεις των πελατών συμπεριλαμβάνονται στο Customer interaction center – CIC, χρησιμοποιώντας διάφορα κανάλια επικοινωνίας.
- **Αυτοματοποίηση επιχειρησιακού marketing (Enterprise marketing automation– EMA)**
Τα εργαλεία EMA προσφέρουν πληροφορίες σχετικά με το επιχειρησιακό περιβάλλον, συμπεριλαμβανομένων της γενικής κατεύθυνσης της επιχείρησης και των ανταγωνιστών. Είναι η εκτελεστική πλευρά της καμπάνιας και του Lead management. Σκοπός μιας εφαρμογής EMA είναι η βελτίωση της αποτελεσματικότητας της εκστρατείας marketing.
- **Συνεργατικό CRM (Collaborative CRM)**
Το Συνεργατικό CRM χρησιμοποιείται για την εναρμόνιση των πολυ-κάναλων υπηρεσιών και την ανταπόκριση και αποτελεσματική υποστήριξη που παρέχεται στους πελάτες όσον αφορά ερωτήσεις, παράπονα, πελατειακά θέματα κλπ. Το Συνεργατικό CRM είναι χρήσιμο για διάφορα τμήματα μέσα στην επιχείρηση όπως είναι τα τμήματα πωλήσεων, τεχνικής υποστήριξης και marketing, ώστε να διαμοιράζονται οι σημαντικές πληροφορίες που συλλέγονται από κάθε αλληλεπίδραση με τους πελάτες. Ο βασικότερος στόχος ενός Συνεργατικού CRM αποτελεί το να είναι εφικτή η χρήση όλων των πληροφοριών που έχουν συγκεντρωθεί από όλα τα τμήματα της επιχείρησης για κάθε πελάτη, ώστε να βελτιωθεί η ποιότητα των πελατειακών υπηρεσιών. Επιπλέον, βασικό στόχο αποτελεί η αύξηση της αποδοτικότητας, του εισοδήματος και της ικανοποίησης του πελάτη.

Το πλέον σημαντικό τμήμα ενός Συνεργατικού CRM είναι η εγκατάσταση λογισμικού σε συνδυασμό με τις πελατοκεντρικές στρατηγικές. Η τεχνολογία είναι ένα δυναμικό εργαλείο για την υποστήριξη του CRM, αλλά χωρίς τον συνδυασμό με τους πελάτες, το εργαλείο αυτό δε μπορεί να προσφέρει το μέγιστο δυνατό πλεονέκτημα στην επιχείρηση.

Το Συνεργατικό CRM υποστηρίζει διαδικασίες που διεξάγονται σε επίπεδο back-office. Τέτοιου είδους διαδικασίες επηρεάζουν τις δραστηριότητες των πελατών και την διατήρηση των πελατειακών σχέσεων. Ορισμένες τέτοιες διαδικασίες είναι οι Collaborative εσωτερικές λειτουργίες της διεύθυνσης IT, η διαφήμιση, η τιμολόγηση, ο σχεδιασμός, το marketing.

Παρόλο που η χρήση ενός Συνεργατικού CRM είναι μια αξιοθαύμαστη ιδέα, οι περισσότερες επιχειρήσεις δυσκολεύονται να δημιουργήσουν ένα σύστημα που να

λειτουργεί. Το πρόβλημα προέρχεται από το γεγονός ότι το κάθε τμήμα της επιχείρησης λειτουργεί ανεξάρτητα από τα υπόλοιπα και υπάρχουν μόνο κάποιες μικρές φυσικές αλληλεπιδράσεις. Η χρήση της τεχνολογίας μπορεί να γεφυρώσει την επικοινωνία μεταξύ των διαφόρων τμημάτων, ενθαρρύνοντας την ροή των πληροφοριών.

- **Αναλυτικό CRM (Analytical CRM)**

Το Αναλυτικό CRM αποτελεί τη λογική συνέχεια του Λειτουργικού και του Συνεργατικού CRM. Κάθε επιχείρηση η οποία έχει υλοποιήσει Λειτουργικό και Συνεργατικό CRM, έχοντας ως σκοπό την καθημερινή καταγραφή, την αυτοματοποίηση των διαδικασιών και την διαχείριση των σχέσεων με τους πελάτες, πρέπει συνεχώς να εμπλουτίζει και να ενημερώνει τη βάση δεδομένων της. Αυτή τη βάση δεδομένων καλείται να αναλύσει το τμήμα του marketing χρησιμοποιώντας το Αναλυτικό CRM ώστε να προκύψουν χρήσιμα και πολύτιμα συμπεράσματα.

Το Αναλυτικό CRM πραγματοποιεί:

- Εξειδικευμένες καμπάνιες marketing, με σκοπό τις σταυροειδείς πωλήσεις (up/cross selling)
- Ανάλυση της συμπεριφοράς των πελατών, ώστε να λαμβάνονται οι βέλτιστες αποφάσεις σχετικά με τα προϊόντα και τις προσφερόμενες υπηρεσίες.
- Προβλέψεις των μελλοντικών εσόδων και εξόδων της επιχείρησης.
- Ανάλυση κερδοφορίας, τόσο σε επίπεδο επιχείρησης, όσο και σε επίπεδο πελάτη.

Για το Αναλυτικό CRM είναι απαραίτητη η χρήση του εργαλείου εξόρυξης δεδομένων (data mining).

Το Αναλυτικό CRM παρέχει μια πλήρη εικόνα των πελατών και έτσι συνεισφέρει ώστε η επιχείρηση να γνωρίζει ποιοι είναι οι πελάτες της, τι χρειάζονται και κυρίως τι μπορεί να κάνουν στο μέλλον.

Σύμφωνα με τον Steve Bonnadío, οι επιχειρήσεις οι οποίες δεν εφαρμόζουν στρατηγικές Αναλυτικού CRM, ταυτόχρονα με τις στρατηγικές των Λειτουργικών και Συνεργατικών, συνήθως αποτυγχάνουν στο CRM.

Ανάλογα με τις ανάγκες της κάθε επιχείρησης, οι σκοποί της εφαρμογής ενός Αναλυτικού CRM είναι:

- Η σταδιακή δημιουργία ενός πελατοκεντρικού data warehouse
Ενοποιώντας τα δεδομένα των πελατών, γίνεται εφικτή η διασπορά τους μέσα στην επιχείρηση με την βοήθεια διαφόρων μεθόδων εξαγωγής, μεταφοράς και επιλογής. Αυτά τα δεδομένα σχετίζονται με τον υπολογισμό της αξίας του πελάτη, της πιστότητας του πελάτη κλπ.
- Ποσοτική ανάλυση της πιστότητας του πελάτη

Είναι σημαντικό για κάθε επιχείρηση να γνωρίζει τον βαθμό αφοσίωσης του κάθε πελάτη. Επιπλέον, έχει ως βασικό μέλημα να δημιουργεί όσο το δυνατόν πιο πιστούς πελάτες. Ο υπολογισμός της αφοσίωσης κάθε πελάτη μπορεί να γίνει με χρήση επιλεγμένων αλγορίθμων.

- Αποτελεσματική ταξινόμηση των πελατών

Η ταξινόμηση των πελατών γίνεται σύμφωνα με τα χαρακτηριστικά όπως η αξία του πελάτη, οι υπηρεσίες ή τα προϊόντα που επιλέγει, κλπ. Με αυτόν τον τρόπο, οι πελάτες τοποθετούνται σε διαφορετικές ομάδες. Σκοπός της επιχείρησης με αυτήν την ταξινόμηση είναι να μπορεί να καταλάβει καλύτερα τις ανάγκες της κάθε διαφορετικής ομάδας πελατών και να εφαρμόσει διαφορετικές στρατηγικές.

- Ανάλυση των αιτιών απώλειας πελάτη

Με τη χρήση διαφόρων αλγορίθμων, αναλύονται χαρακτηριστικά των χαμένων ομάδων πελατών. Μέσα από αυτήν την ανάλυση βρίσκουν τους λόγους που αυτό συνέβη, καθώς επίσης και αν υπάρχουν τρόποι ώστε να ανακτηθούν αυτοί οι πελάτες.

- Σύστημα αξιολόγησης.

Αναλύονται τα αποτελέσματα του υπολογισμού της αξίας και της πιστότητας του πελάτη, της ταξινόμησης των πελατών, καθώς και των χαμένων πελατών. Παρατηρούνται τα πλεονεκτήματα και τα μειονεκτήματα των επιλεγμένων αλγορίθμων και με αυτόν τον τρόπο βελτιώνονται οι μετρήσεις.

Κεφάλαιο 4: Ρόλος και Εφαρμογή του Δείκτη Customer Lifetime Value (CLV)

4.1 Εισαγωγή

Επειδή οι σημερινές επιχειρήσεις αντιμετωπίζουν αυξημένη πολυπλοκότητα και ανταγωνισμό, οι εταιρείες πρέπει να αναπτύξουν καινοτόμες δραστηριότητες μάρκετινγκ που να ανταποκρίνονται και να υπερβαίνουν τις ανάγκες των πελατών και να βελτιώνουν την ικανοποίηση και τη διατήρησή τους. Το επίκεντρο του CRM είναι η δημιουργία στενότερων και βαθύτερων σχέσεων με τους πελάτες. Επομένως, οι επιχειρήσεις θα πρέπει να είναι πρόθυμες και ικανές να αλλάζουν τη συμπεριφορά τους προς κάθε μεμονωμένο πελάτη με βάση όσα γνωρίζουν για τον καθένα.

Λόγω ότι η προσέλκυση νέων πελατών (acquisition) είναι δαπανηρή, η διατήρηση των υφιστάμενων πελατών (retention) είναι απαραίτητη. Η διατήρηση των πελατών μπορεί να μεγιστοποιηθεί με τον προσδιορισμό των προϊόντων και υπηρεσιών πιο κοντά στις προτιμήσεις τους. Έτσι, οι υφιστάμενοι πελάτες μπορούν να αποβούν πιο κερδοφόροι από τους νέους. Επομένως, ο κεντρικός στόχος του CRM είναι η μεγιστοποίηση της αξίας ζωής ενός πελάτη (Customer Lifetime Value) για τον οργανισμό.

Στην ουσία, το παραδοσιακό στατικό CRM αφορά την ανάλυση των πληροφοριών των πελατών για επιχειρηματικές αποφάσεις με στόχο να βοηθήσει τους οργανισμούς να κατανοήσουν τις ανάγκες των πελατών. Ωστόσο, οι διαφορές μεταξύ των πελατών μπορούν να προσδιοριστούν μέσω τμηματοποίησης της αγοράς, η οποία υποδηλώνει ότι ορισμένοι πελάτες είναι πιο κερδοφόροι από άλλους, καθώς αναλύει την πίστη (loyalty), την κερδοφορία (profitability) και την απόρριψη (churn) των πελατών και διοχετεύει την αποτελεσματικότητα και την απόδοση της καμπάνιας.

4.2 Ορισμοί

Οι εταιρείες χρειάζεται να αξιολογούν με ακρίβεια την αξία των πελατών τους και να χτίζουν στρατηγικές για να διατηρήσουν τους κερδοφόρους πελάτες. Για την εταιρεία, η αξία που παράγεται από ένα τέτοιο αποτέλεσμα πρέπει να υπολογίζεται μέσω της αξίας ζωής του πελάτη.

Βάσει των ανωτέρω, έχουν προκύψει οι εξής ορισμοί της Αξίας Διάρκειας Ζωής του πελάτη:

- **Lifetime Value – LTV:**
Κατά τον Novo (2001), η αξία διάρκειας ζωής (LTV) του πελάτη είναι μια πρόβλεψη του συνολικού καθαρού κέρδους που μια εταιρεία μπορεί να αναμένει σε μια συνεχή σχέση μεταξύ πελάτη και προϊόντος ή υπηρεσίας.
- **Customer Lifetime Value – CLV:**
Κατά τους Kumar και Rajan (2009), η αξία διάρκειας ζωής πελάτη (CLV) ορίζεται ως το άθροισμα των σωρευμένων ταμειακών ροών-μειούμενο με χρήση του σταθμισμένου μέσου

κόστους κεφαλαίου (Weighted Average Cost of Capital - WACC)- ενός πελάτη καθ' όλη τη διάρκεια της ζωής του στην εταιρεία.

Ως εκ τούτου, οι εταιρείες χρειάζεται να αξιολογούν με ακρίβεια την αξία των πελατών τους και να χτίζουν στρατηγικές για να διατηρούν κερδοφόρους πελάτες.

4.3 Υπολογισμός του δείκτη Customer Lifetime Value (CLV)

Ο τύπος υπολογισμού της αξίας διάρκειας ζωής πελάτη (CLV) είναι η αξία πελάτη (Customer Value) πολλαπλασιαζόμενη με τη μέση διάρκεια ζωής του πελάτη (Average Customer Lifespan). Το αποτέλεσμα δίνει τα έσοδα που η εταιρεία μπορεί εύλογα να περιμένει να δημιουργήσει ένας μέσος πελάτης σε όλη τη διάρκεια της σχέσης του μαζί της.

$$CLV = \text{Customer Value} \times \text{Average Customer Lifespan}$$

Για τον υπολογισμό της αξίας πελάτη (Customer Value), πρέπει να υπολογιστεί η μέση αξία αγοράς (Average Purchase Value) και, στη συνέχεια, να πολλαπλασιαστεί αυτός ο αριθμός με τη μέση συχνότητα αγορών (Average Purchase Frequency Rate).

$$\text{Customer Value} = \text{Average Purchase Value} \times \text{Average Purchase Frequency Rate}$$

Η μέση αξία αγοράς (Average Purchase Value) υπολογίζεται διαιρώντας τα συνολικά έσοδα της εταιρείας (Total Revenue) σε μια περίοδο (συνήθως ενός έτους) με τον αριθμό των αγορών (Number of Orders) κατά την ίδια περίοδο.

$$\text{Average Purchase Value} = \frac{\text{Total Revenue}}{\text{Number of Orders}}$$

Η μέση συχνότητα αγορών (Average Purchase Frequency Rate) υπολογίζεται διαιρώντας τον αριθμό των αγορών (Number of Purchases) με τον αριθμό των μοναδικών πελατών (Number of Unique Customers) που πραγματοποίησαν αγορές κατά τη διάρκεια αυτής της περιόδου.

$$\text{Average Purchase Frequency Rate} = \frac{\text{Number of Purchases}}{\text{Number of Unique Customers}}$$

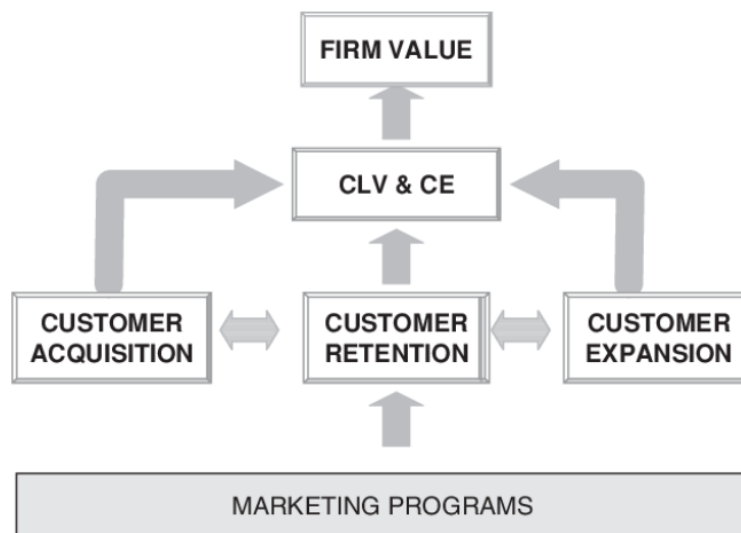
Για τον υπολογισμό της μέσης διάρκειας ζωής του πελάτη (Average Customer Lifespan), πρέπει να υπολογιστεί ο μέσος όρος των ετών που ένας πελάτης συνεχίζει να αγοράζει από την εταιρεία.

$$\text{Average Customer Lifespan} = \frac{\text{Sum of Customer Lifespans}}{\text{Number of Customers}}$$

4.4 Μοντελοποίηση του δείκτη Customer Lifetime Value (CLV)

Το πρώτο επίσημο μοντέλο για το CLV αναπτύχθηκε στις αρχές της δεκαετίας του 1980 από τους Roland Rust, Valarie Zeithaml και Katherine Lemon, οι οποίοι πρότειναν ένα πλαίσιο για την εκτίμηση της αξίας ενός πελάτη με βάση τις αναμενόμενες μελλοντικές του αγορές. Αυτό το μοντέλο βασίστηκε στην υπόθεση ότι οι πελάτες έχουν μια συγκεκριμένη πιθανότητα να κάνουν επαναλαμβανόμενες αγορές και ότι αυτή η πιθανότητα μπορεί να εκτιμηθεί με βάση την προηγούμενη συμπεριφορά τους.

Με την έλευση των ψηφιακών τεχνολογιών και την ευρεία διαθεσιμότητα των δεδομένων πελατών, έχει σημειωθεί έκρηξη στον αριθμό και την ποικιλία των προσεγγίσεων στη μοντελοποίηση CLV. Αυτά περιλαμβάνουν παραδοσιακά στατιστικά μοντέλα, αλγόριθμους μηχανικής μάθησης και άλλες προηγμένες τεχνικές που χρησιμοποιούν την ανάλυση δεδομένων και την τεχνητή νοημοσύνη για την ανάλυση μεγάλου όγκου δεδομένων πελατών και τον εντοπισμό μοτίβων και τάσεων που μπορούν να βοηθήσουν τις επιχειρήσεις να βελτιστοποιήσουν το CLV τους.



Εικόνα 6 - Μοντελοποίηση Αξίας Διάρκειας Ζωής Πελάτη
(CLV = Διάρκεια ζωής πελάτη, CE = Καθαρή θέση πελάτη)
Conceptual Framework for Modeling Customer Lifetime Value - Gupta et al., 2006

Υπάρχουν πολλές διαφορετικές προσεγγίσεις για τη μοντελοποίηση του CLV, καθεμία από τις οποίες μπορεί να είναι χρήσιμη σε διαφορετικά πλαίσια. Ορισμένες από αυτές θα αναλυθούν στις επόμενες παραγράφους:

4.4.1 Μοντέλα RFM

Η ανάλυση RFM (Recency, Frequency, Monetary) είναι μια μέθοδος που χρησιμοποιείται στο μάρκετινγκ και στη διαχείριση των σχέσεων με τους πελάτες για τον εντοπισμό των πιο πολύτιμων πελατών με βάση την αγοραστική τους συμπεριφορά. Είναι μια απλή αλλά αποτελεσματική προσέγγιση που κατηγοριοποιεί τους πελάτες σε διαφορετικά τμήματα με βάση τρία κριτήρια:

- Εγγύτητα (Recency):
Πότε ήταν η τελευταία φορά που ένας πελάτης έκανε μια αγορά;
- Συχνότητα (Frequency):
Πόσο συχνά κάνουν αγορές οι πελάτες σε μία συγκεκριμένη χρονική περίοδο;
- Monetary (Νομισματική Αξία):
Πόσα χρήματα ξοδεύουν οι πελάτες σε κάθε αγορά σε μία χρονική περίοδο;

Η ανάλυση RFM εκχωρεί σε κάθε πελάτη μια βαθμολογία για καθένα από αυτά τα τρία κριτήρια και στη συνέχεια συνδυάζει αυτές τις βαθμολογίες σε μια ενιαία βαθμολογία RFM. Οι πελάτες με υψηλότερα σκορ RFM θεωρούνται πιο πολύτιμοι για την επιχείρηση και μπορούν να στοχευθούν με συγκεκριμένες καμπάνιες μάρκετινγκ ή στρατηγικές διατήρησης.

Το μοντέλο RFM αναπτύχθηκε για πρώτη φορά τη δεκαετία του 1950 από μια ομάδα ερευνητών στο Direct Marketing Association, οι οποίοι αναζητούσαν έναν τρόπο να εντοπίσουν τους πελάτες που ανταποκρίνονται περισσότερο στις καμπάνιες. Έκτοτε, το μοντέλο έχει υιοθετηθεί ευρέως από επιχειρήσεις όλων των τύπων και μεγεθών και έχει προσαρμοστεί για να ταιριάζει σε μια σειρά διαφορετικών εφαρμογών μάρκετινγκ και διαχείρισης πελατών.

Ένα από τα βασικά πλεονεκτήματα του μοντέλου RFM είναι η απλότητα και η ευκολία χρήσης του. Μπορεί να εφαρμοστεί χρησιμοποιώντας βασικό λογισμικό υπολογιστικών φύλλων και δεν απαιτεί προηγμένα στατιστικά ή αναλυτικά εργαλεία. Επιπλέον, παρέχει ένα σαφές και διαισθητικό πλαίσιο για την κατανόηση της συμπεριφοράς των πελατών και τον εντοπισμό προτύπων και τάσεων στα δεδομένα πελατών.

Ωστόσο, υπάρχουν ορισμένοι περιορισμοί στο μοντέλο RFM. Για παράδειγμα, δεν λαμβάνει υπόψη άλλους σημαντικούς παράγοντες που μπορεί να επηρεάσουν την αξία των πελατών, όπως δημογραφικά στοιχεία πελατών, ψυχογραφικά στοιχεία ή κοινωνική επιρροή. Επιπλέον, μπορεί να μην είναι κατάλληλο για επιχειρήσεις με μεγάλο αριθμό πελατών χαμηλής αξίας, καθώς αυτοί οι πελάτες μπορεί να μην είναι εύκολα διακριτοί με βάση τις βαθμολογίες RFM τους.

Συνολικά, το μοντέλο RFM παραμένει μια δημοφιλής και αποτελεσματική προσέγγιση για την ανάλυση της συμπεριφοράς των πελατών και τον εντοπισμό πολύτιμων πελατών, ιδιαίτερα για επιχειρήσεις με έμφαση στο άμεσο μάρκετινγκ ή τη διατήρηση πελατών.

4.4.2 Μοντέλα Πιθανοτήτων (Profitability Models)

Τα μοντέλα πιθανοτήτων μπορούν να είναι ένα χρήσιμο εργαλείο στη μοντελοποίηση της αξίας ζωής του πελάτη (CLV), καθώς παρέχουν έναν τρόπο εκτίμησης της κερδοφορίας μεμονωμένων πελατών ή τμημάτων πελατών. Συνδυάζοντας μοντέλα πιθανοτήτων με άλλα δεδομένα, όπως το ιστορικό αγорών και η συμπεριφορά των πελατών, οι εταιρείες μπορούν να αποκτήσουν μια πιο ολοκληρωμένη κατανόηση της αξίας σχέσης με κάθε πελάτη.

Μια προσέγγιση για την ενσωμάτωση μοντέλων πιθανοτήτων στη μοντελοποίηση CLV είναι η εκτίμηση της καθαρής παρούσας αξίας (Net Present Value - NPV) κάθε πελάτη. Η NPV λαμβάνει υπόψη τόσο τα προσδοκώμενα μελλοντικά έσοδα που παράγονται από έναν πελάτη όσο και το κόστος που σχετίζεται με την εξυπηρέτηση αυτού του πελάτη με την πάροδο του χρόνου. Αφαιρώντας το εκτιμώμενο κόστος από τα εκτιμώμενα έσοδα και προεξοφλώντας τις προκύπτουσες ταμειακές ροές στην παρούσα αξία τους, οι εταιρείες μπορούν να εκτιμήσουν την NPV κάθε πελάτη.

Μια πρόκληση για την ενσωμάτωση μοντέλων πιθανοτήτων στη μοντελοποίηση CLV είναι η ανάγκη να ληφθούν υπόψη μακροπρόθεσμα στρατηγικά ζητήματα, όπως η δημιουργία επωνυμίας και η έρευνα και ανάπτυξη. Αυτοί οι παράγοντες μπορεί να μην συνδέονται άμεσα με τη βραχυπρόθεσμη κερδοφορία, αλλά μπορούν να έχουν σημαντικό αντίκτυπο στην αξία διάρκειας ζωής των πελατών με την πάροδο του χρόνου.

Συνοπτικά, τα μοντέλα πιθανοτήτων επιδιώκουν να κάνουν προβλέψεις σχετικά με το αν ένας πελάτης θα συνεχίσει μελλοντικά να είναι ενεργός. Εφόσον ισχύει αυτή η προϋπόθεση, τότε εξετάζεται η αγοραστική του συμπεριφορά.

4.4.3 Οικονομετρικά Μοντέλα (Econometric Models)

Τα οικονομετρικά μοντέλα χρησιμοποιούνται ευρέως στη μοντελοποίηση της Αξίας διάρκειας ζωής πελατών (CLV) και μπορούν να βοηθήσουν τις επιχειρήσεις να εκτιμήσουν τον αντίκτυπο διαφορετικών στρατηγικών μάρκετινγκ στη συμπεριφορά των πελατών. Αυτά τα μοντέλα μπορούν να χρησιμοποιηθούν για τη βελτιστοποίηση των στρατηγικών απόκτησης (acquisition), διατήρησης (retention) και επέκτασης (expansion), οι οποίες αποτελούν τους τρεις βασικούς τρόπους με τους οποίους οι επιχειρήσεις μπορούν να αυξήσουν την αξία διάρκειας ζωής των πελατών.

Οι στρατηγικές απόκτησης (acquisition) εστιάζονται στην προσέλκυση νέων πελατών σε μια επιχείρηση. Τα οικονομετρικά μοντέλα μπορούν να βοηθήσουν τις επιχειρήσεις να εκτιμήσουν τη διάρκεια ζωής των νέων πελατών και να εντοπίσουν τα πιο αποτελεσματικά κανάλια και τακτικές για την προσέγγισή τους. Για παράδειγμα, οι επιχειρήσεις μπορούν να χρησιμοποιούν στοχευμένες καμπάνιες μάρκετινγκ για να προσελκύσουν πελάτες που είναι πιθανό να έχουν υψηλό CLV, με βάση τα δημογραφικά στοιχεία, τα ενδιαφέροντα και τη συμπεριφορά τους στο παρελθόν.

Οι στρατηγικές διατήρησης (retention) επικεντρώνονται στη διατήρηση των υπάρχοντων πελατών πιστών σε μια επιχείρηση. Τα οικονομετρικά μοντέλα μπορούν να βοηθήσουν τις επιχειρήσεις να εκτιμήσουν την πιθανότητα απόρριψης (churn) ή φθοράς (attrition) πελατών, με βάση την προηγούμενη συμπεριφορά του πελάτη και τις παρεμβάσεις μάρκετινγκ που έχουν στοχεύσει σε αυτούς. Εντοπίζοντας τους παράγοντες που είναι πιο πιθανό να προκαλέσουν την αποχώρηση των πελατών, οι επιχειρήσεις μπορούν να λάβουν μέτρα για να αντιμετωπίσουν αυτά τα ζητήματα και να βελτιώσουν τη διατήρηση των πελατών.

Οι στρατηγικές επέκτασης (expansion) επικεντρώνονται στην αύξηση της αξίας των υπαρχόντων πελατών, ενθαρρύνοντάς τους να κάνουν περισσότερες αγορές ή αυξάνοντάς τους τις πωλήσεις σε προϊόντα ή υπηρεσίες υψηλότερης αξίας. Τα οικονομετρικά μοντέλα μπορούν να βοηθήσουν τις επιχειρήσεις να εκτιμήσουν τον αντίκτυπο διαφορετικών τύπων προσφορών και προωθήσεων στη συμπεριφορά των πελατών και να εντοπίσουν τους πελάτες που είναι πιο πιθανό να ανταποκριθούν θετικά σε αυτές τις προσφορές.

Συνολικά, τα οικονομετρικά μοντέλα είναι ένα πολύτιμο εργαλείο για τις επιχειρήσεις για τη βελτιστοποίηση των στρατηγικών CLV τους. Με την εκτίμηση του αντίκτυπου των διαφορετικών παρεμβάσεων μάρκετινγκ στη συμπεριφορά των πελατών, οι επιχειρήσεις μπορούν να λάβουν τεκμηριωμένες αποφάσεις σχετικά με τον τρόπο κατανομής των πόρων τους και τη βελτίωση της μακροπρόθεσμης κερδοφορίας τους.

4.4.4 Μοντέλα Μηχανικής Μάθησης (Machine Learning Models)

Τα μοντέλα μηχανικής μάθησης έχουν αναδειχθεί ως ένα ισχυρό εργαλείο για τη μοντελοποίηση CLV.

Για τη δημιουργία αυτών των μοντέλων χρησιμοποιούνται τεχνικές ταξινόμησης και κατάταξης, νευρωνικά δίκτυα και μηχανές υποστήριξης αποφάσεων. Τα μοντέλα αυτά χρησιμοποιούν αλγόριθμους για την ανάλυση μεγάλου όγκου δεδομένων πελατών και τον εντοπισμό μοτίβων και τάσεων που μπορούν να χρησιμοποιηθούν για την πρόβλεψη της συμπεριφοράς των πελατών και την εκτίμηση της αξίας ζωής τους.

Σε αντίθεση με τα παραδοσιακά στατιστικά μοντέλα, τα μοντέλα μηχανικής μάθησης μπορούν να μαθαίνουν από δεδομένα με επαναληπτικό τρόπο, επιτρέποντάς τους να προσαρμόζονται και να βελτιώνουν τις προβλέψεις τους με την πάροδο του χρόνου.

4.4.4.1 XGBoost Classifier

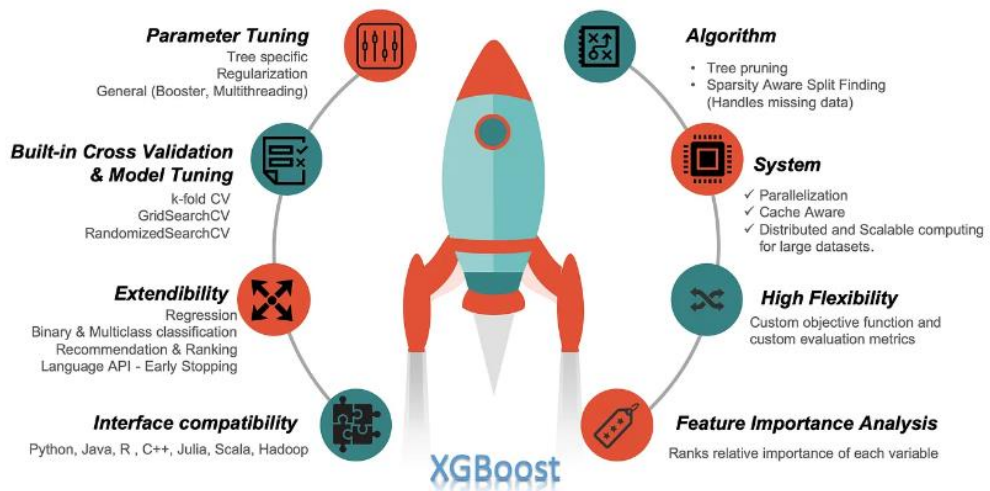
Η ονομασία XGBoost προέρχεται από την έκφραση «Extreme Gradient Boosting». Ο αλγόριθμος XGBoost λειτουργεί εκπαιδεύοντας επαναληπτικά μια σειρά από δέντρα αποφάσεων (decision trees) που κάνουν προβλέψεις για τη μεταβλητή στόχο. Κάθε δέντρο εκπαιδεύεται στα υπόλοιπα (residuals) των προηγούμενων δέντρων, με στόχο την ελαχιστοποίηση της αντικειμενικής συνάρτησης, η οποία είναι ένα μέτρο της διαφοράς μεταξύ των προβλεπόμενων και των πραγματικών τιμών.

Ένα από τα βασικά χαρακτηριστικά του XGBoost είναι η ικανότητά του να χειρίζεται μεγάλα σύνολα δεδομένων με χαρακτηριστικά υψηλών διαστάσεων, γεγονός που το καθιστά ελκυστική επιλογή για πολλές εφαρμογές του πραγματικού κόσμου. Επιπλέον, ο αλγόριθμος XGBoost υποστηρίζει διάφορες τεχνικές συστηματοποίησης (regularization techniques) για την αποφυγή υπερπροσαρμογής (overfitting) και τη βελτίωση της απόδοσης γενίκευσης (generalization) του μοντέλου.

Στο πλαίσιο της μοντελοποίησης CLV, ο XGBoost μπορεί να χρησιμοποιηθεί ως ένα ισχυρό εργαλείο για την πρόβλεψη της συμπεριφοράς των πελατών και την εκτίμηση της αξίας ζωής τους. Αξιοποιώντας το πλούσιο σύνολο δυνατοτήτων που είναι διαθέσιμες στα δεδομένα πελατών, ο XGBoost μπορεί να μάθει πολύπλοκα μοτίβα και σχέσεις που είναι δύσκολο να αποτυπωθούν με παραδοσιακά στατιστικά μοντέλα.

Πλεονεκτήματα XGBoost:

- **Επεκτασιμότητα (Scalability):**
Ο XGBoost έχει σχεδιαστεί για να διαχειρίζεται μεγάλα σύνολα δεδομένων με χαρακτηριστικά υψηλών διαστάσεων. Ένα από τα χαρακτηριστικά του αποτελεί το out-of-core computing, το οποίο βελτιστοποιεί τον διαθέσιμο αποθηκευτικό χώρο και μεγιστοποιεί τη χρήση του.
- **Συστηματοποίηση (Regularization):**
Ο XGBoost υποστηρίζει μια σειρά τεχνικών συστηματοποίησης, οι οποίες βοηθούν στην αποφυγή υπερπροσαρμογής, καθώς δίνεται η δυνατότητα αρνητικής βαθμολόγησης των περίπλοκων μοντέλων.
- **Σημασία των χαρακτηριστικών (Feature Importance):**
Ο XGBoost παρέχει μια ενσωματωμένη βαθμολογία σημασίας χαρακτηριστικών που μπορεί να βοηθήσει στον εντοπισμό των πιο σημαντικών χαρακτηριστικών στο σύνολο δεδομένων. Αυτές οι πληροφορίες μπορούν να χρησιμοποιηθούν για τη βελτίωση της απόδοσης και της ερμηνείας του μοντέλου.
- **Ευελιξία (Flexibility):**
Ο XGBoost είναι ένας ισχυρός και ευέλικτος αλγόριθμος μηχανικής μάθησης που μπορεί να εφαρμοστεί σε ένα ευρύ φάσμα προβλημάτων ταξινόμησης (classification) και παλινδρόμησης (regression), συμπεριλαμβανομένης της μοντελοποίησης CLV.
- **Χειρισμός ημιτελών δεδομένων (Handling Incomplete Data):**
Ο XGBoost διαθέτει ενσωματωμένη υποστήριξη για το χειρισμό τιμών που λείπουν στα δεδομένα εισόδου. Μπορεί να μάθει αυτόματα την καλύτερη στρατηγική για την απόδοση τιμών στα ημιτελή δεδομένα (missing values imputation) με βάση τα δεδομένα και το μοντέλο, κάτι που μπορεί να βελτιώσει την ακρίβεια των προβλέψεων.
- **Επίγνωση προσωρινής μνήμης (Cache Awareness):**
Ο XGBoost έχει σχεδιαστεί για να έχει επίγνωση της κρυφής μνήμης, πράγμα που σημαίνει ότι μπορεί να χρησιμοποιήσει αποτελεσματικά τη διαθέσιμη προσωρινή μνήμη της CPU για να επιταχύνει τον υπολογισμό. Αυτό μπορεί να βελτιώσει σημαντικά την απόδοση του αλγορίθμου, ειδικά για μεγάλα σύνολα δεδομένων.



Εικόνα 7 - Χαρακτηριστικά του αλγορίθμου XGBoost
 XGBoost: A Deep Dive Into Boosting – Dzone

Κεφάλαιο 5: Δημιουργία Μοντέλου Μηχανικής Μάθησης για την πρόβλεψη του Δείκτη CLV

5.1 Εισαγωγή

Σε αυτό το κεφάλαιο περιγράφεται η πειραματική διαδικασία υλοποίησης ενός μοντέλου μηχανικής μάθησης, το οποίο τροφοδοτείται με δεδομένα από μία εταιρεία iGaming και πραγματοποιεί προβλέψεις σχετικά με τις τιμές του δείκτη Customer Lifetime Value, όσον αφορά τις καταθέσεις των πελάτων.

Στις επόμενες ενότητες, αναλύεται πρωτίστως η συλλογή, η προεπεξεργασία και η ανάλυση των δεδομένων που χρησιμοποιούνται. Στη συνέχεια, περιγράφεται η διαδικασία δημιουργίας και εφαρμογής μοντέλων μηχανικής μάθησης.

5.2 Τεχνικό Υπόβαθρο

Για τη συλλογή των δεδομένων έχει χρησιμοποιηθεί κώδικας SQL σε περιβάλλον PostgreSQL, όπου βρίσκονται αποθηκευμένα τα δεδομένα της εταιρείας. Η ανάλυση των δεδομένων και η ανάπτυξη των αλγορίθμων μηχανικής μάθησης έχει υλοποιηθεί με χρήση της γλώσσας προγραμματισμού Python σε περιβάλλον Jupyter Notebook.

1. SQL:

Η SQL (Structured Query Language) είναι μια τυπική γλώσσα προγραμματισμού που χρησιμοποιείται για τη διαχείριση και το χειρισμό σχεσιακών βάσεων δεδομένων. Χρησιμοποιείται για την εισαγωγή, ενημέρωση, διαγραφή και ανάκτηση δεδομένων από βάσεις δεδομένων. Η SQL είναι μια δηλωτική γλώσσα, που σημαίνει ότι προσδιορίζετε τι θέλετε να κάνετε με τα δεδομένα και όχι πώς να το κάνετε.

Η SQL αναπτύχθηκε για πρώτη φορά τη δεκαετία του 1970 από τους ερευνητές της IBM Donald D. Chamberlin και Raymond F. Boyce. Έκτοτε έχει γίνει μια ευρέως χρησιμοποιούμενη γλώσσα για εργασία με σχεσιακές βάσεις δεδομένων και υποστηρίζεται από πολλά συστήματα διαχείρισης βάσεων δεδομένων, συμπεριλαμβανομένων των MySQL, PostgreSQL, Oracle, Microsoft SQL Server και SQLite, μεταξύ άλλων.

2. Python:

Η Python είναι μια γλώσσα προγραμματισμού υψηλού επιπέδου, γενικής χρήσης, σχεδιασμένη να δίνει έμφαση στην αναγνωσιμότητα και την απλότητα του κώδικα. Κυκλοφόρησε για πρώτη φορά το 1991 από τον Guido van Rossum και έκτοτε έχει γίνει μια από τις πιο δημοφιλείς γλώσσες προγραμματισμού στον κόσμο.

Η Python χρησιμοποιείται συχνά για ανάπτυξη ιστού, ανάλυση δεδομένων, τεχνητή νοημοσύνη και αυτοματισμό. Ένα από τα πλεονεκτήματά της είναι ότι διαθέτει πολλές βιβλιοθήκες, οι οποίες παρέχουν υποστήριξη για πολλές κοινές εργασίες προγραμματισμού.

Παρακάτω περιγράφονται οι βιβλιοθήκες που χρησιμοποιήθηκαν στην παρούσα διπλωματική εργασία:

- **Pandas:**
Η ονομασία Pandas προήλθε από τη φράση «Panel Data», με την οποία εννοούμε τα δεδομένα που είναι οργανωμένα σε μορφή πίνακα με σειρές και στήλες, όπου κάθε γραμμή αντιπροσωπεύει μια παρατήρηση και κάθε στήλη αντιπροσωπεύει μια μεταβλητή. Η βιβλιοθήκη Pandas είναι μια ισχυρή βιβλιοθήκη που χρησιμοποιείται για διαχείριση και ανάλυση δεδομένων. Παρέχει δομές δεδομένων όπως Series και DataFrame για χειρισμό δεδομένων και εκτελεί διάφορες λειτουργίες σε αυτά όπως ομαδοποίηση, φιλτράρισμα και συγχώνευση. Μπορεί να διαβάσει δεδομένα από διάφορες πηγές όπως CSV, Excel, βάσεις δεδομένων SQL και άλλα.
- **NumPy:**
Ο όρος NumPy προέρχεται από τη φράση «Numerical Python». Η NumPy είναι μια βιβλιοθήκη που χρησιμοποιείται για επιστημονικούς υπολογισμούς και αριθμητικές πράξεις. Από τις βασικές της ιδιότητες είναι η διαχείριση μεγάλων πολυδιάστατων πινάκων και η ευκολία στις μαθηματικές πράξεις στα στοιχεία του πίνακα. Παρέχει επίσης συναρτήσεις για γραμμική άλγεβρα, ανάλυση Fourier και δημιουργία τυχαίων αριθμών.
- **Matplotlib:**
Το όνομα «Matplotlib» προέρχεται από το «MATLAB» και το «plotting». Η MATLAB είναι μια γλώσσα προγραμματισμού που χρησιμοποιείται συνήθως για αριθμητική ανάλυση, οπτικοποίηση δεδομένων και άλλες επιστημονικές εργασίες υπολογιστών. Η Matplotlib είναι μια βιβλιοθήκη χρήσιμη για την οπτικοποίηση των δεδομένων, μέσω της δημιουργίας γραφημάτων.
- **Seaborn:**
Η βιβλιοθήκη Seaborn παρέχει μια σειρά από λειτουργίες οπτικοποίησης που βασίζονται στη Matplotlib, διευκολύνοντας τη δημιουργία σύνθετων στατιστικών απεικονίσεων με λιγότερες γραμμές κώδικα. Η βιβλιοθήκη παρέχει λειτουργίες για τη δημιουργία διάφορων τύπων γραφημάτων, συμπεριλαμβανομένων των χαρτών θερμότητας (heatmaps), των διαγραμμάτων διασποράς (scatterplots), των διαγραμμάτων γραμμών (line plots), των γραφημάτων ράβδων (bar plots) και άλλων. Παρέχει επίσης υποστήριξη για την εξερεύνηση στατιστικών δεδομένων, όπως τη δημιουργία γραφημάτων για την εξέταση των κατανομών, τις σχέσεις μεταξύ των μεταβλητών και άλλες πτυχές της ανάλυσης δεδομένων.
- **Scikit-learn:**
Το όνομα «Scikit-learn» είναι ένας συνδυασμός του «SciPy» (Scientific Python) and «toolkit». Η Scikit-learn είναι μια βιβλιοθήκη που χρησιμοποιείται για τη μηχανική μάθηση και την εξόρυξη δεδομένων. Παρέχει εργαλεία ταξινόμησης, παλινδρόμησης, ομαδοποίησης και μείωσης διαστάσεων. Παρέχει επίσης υποστήριξη για την επιλογή μοντέλου, τη διασταυρωμένη επικύρωση (cross-validation) και τις μετρήσεις απόδοσης.

5.3 Περιγραφή Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν για τους σκοπούς της εργασίας είναι τρία αρχεία τύπου .csv, τα οποία περιέχουν πληροφορίες για δημογραφικά χαρακτηριστικά, τις καταθέσεις και τις συναλλαγές των πελατών μίας εταιρείας iGaming. Πρόκειται για δεδομένα που προήλθαν με εξαγωγή από τη βάση δεδομένων της εταιρείας με χρήση κώδικα SQL. Τα τρία αρχεία έχουν μια κοινή μεταβλητή «Customer ID» βάση της οποίας γίνονται οι απαραίτητες συνενώσεις, για την εφαρμογή των επιλεχθέντων αλγορίθμων.

Παρακάτω περιγράφονται τα δεδομένα των τριών αρχείων:

1. customers_data.csv

Το παρόν αρχείο περιέχει δεδομένα για ορισμένα δημογραφικά χαρακτηριστικά των πελατών. Αποτελείται από τις στήλες:

- Customer ID: Το Id κάθε πελάτη, το οποίο είναι ξεχωριστό για τον καθένα (τύπου integer)
- Date of Birth: Η ημερομηνία γέννησης κάθε πελάτη (τύπου datetime)
- Gender: Το φύλο κάθε πελάτη, με τιμές MALE ή FEMALE (τύπου string)
- Country ISO: Η χώρα εγγραφής κάθε πελάτη σε μορφή ISO (τύπου string)

2. deposits_data.csv

Το αρχείο αυτό περιέχει αναλυτικά όλες τις καταθέσεις που έχουν πραγματοποιήσει οι πελάτες. Αποτελείται από τις στήλες:

- Customer ID: Το Id κάθε πελάτη (τύπου int)
- Date: Η ημερομηνία και ώρα κατάθεσης (τύπου datetime)
- Deposit: Το ποσό της κατάθεσης σε ευρώ (τύπου float)

3. transactions_data.csv

Το αρχείο αυτό περιέχει τα συνολικά ποσά των συναλλαγών των πελατών ανά ημέρα. Αποτελείται από τις στήλες:

- Customer ID: Το Id κάθε πελάτη (τύπου int)
- Date: Η ημερομηνία της συναλλαγής (τύπου datetime)
- Stake: Το ποσό πονταρίσματος σε ευρώ (τύπου float)
- Winnings: Το ποσό των κερδών σε ευρώ (τύπου float)

5.4 Προεπεξεργασία Δεδομένων

Ένα σημαντικό μέρος της προεπεξεργασίας των δεδομένων πραγματοποιήθηκε στο σύστημα διαχείρισης της βάσης δεδομένων PostgreSQL, μέσω των SQL ερωτημάτων που χρησιμοποιήθηκαν για να δημιουργηθούν τα τρία αρχεία .csv που εισάγονται και αναλύονται στον κώδικα της Python.

Ωστόσο, κατά την ανάλυση των δεδομένων, φάνηκε χρήσιμο να δημιουργηθούν ηλικιακές ομάδες των πελατών. Αρχικά, υπολογίζεται η ηλικία των πελατών, βάσει της ημερομηνίας γεννήσεώς τους και μετέπειτα δημιουργούνται τα κατάλληλα ηλικιακά γκρουπ.

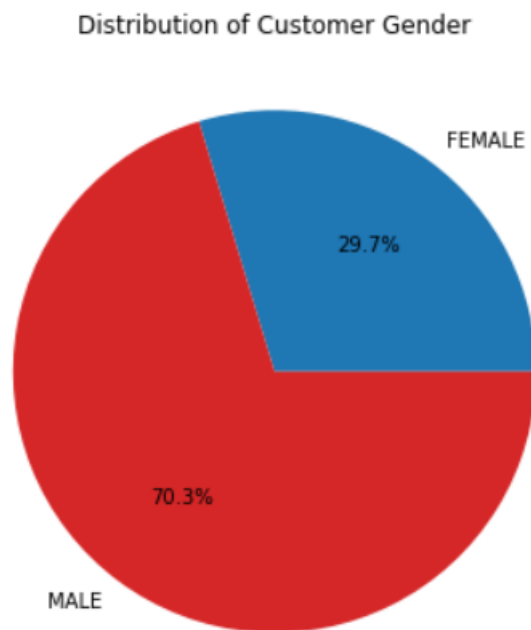
Επιπλέον, προστίθεται μία νέα στήλη με τον όνομα της χώρας εγγραφής του κάθε πελάτη, με βάση τον κωδικό ISO που υπάρχει στα αρχικά δεδομένα.

Τέλος, στον πίνακα των συναλλαγών προστίθεται μία στήλη με όνομα GGR (Gross Gaming Revenue) που είναι η διαφορά του πονταρίσματος (Stake) με τα κέρδη (Winnings) των πελατών ανά ημέρα. Πράγμα το οποίο σημαίνει ότι όταν το πρόσημο αυτής της μεταβλητής είναι θετικό τότε ο πελάτης είναι χαμένος, ενώ όταν το πρόσημο είναι αρνητικό τότε ο πελάτης είναι κερδισμένος, σε ποσό τόσο όσο η απόλυτη τιμή του αριθμού.

5.5 Διερεύνηση των Δεδομένων

Πραγματοποιήθηκε μια διερευνητική εργασία στις πληροφορίες της βάσης δεδομένων με σκοπό να αναλυθούν αποτελεσματικά τα δεδομένα της μελέτης. Με αυτόν τον τρόπο δημιουργήθηκε μια πλήρης εικόνα σχετικά με τα χαρακτηριστικά και τα μεγέθη των δεδομένων. Αυτό το τελικό σύνολο δεδομένων θα χρησιμοποιηθεί στη συνέχεια για την εκπαίδευση και αξιολόγηση των αλγορίθμων που έχουν επιλεγεί.

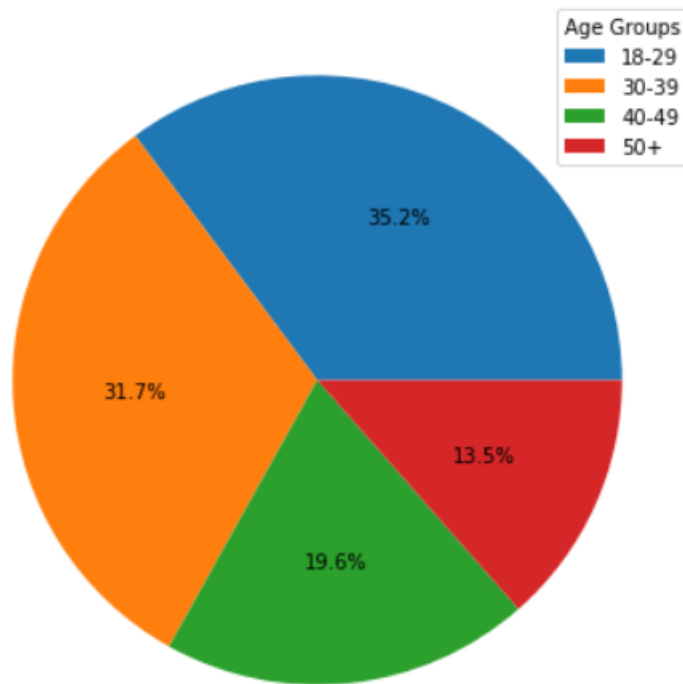
Με βάση τα δημογραφικά στοιχεία των πελατών, παρατηρείται ότι η πλειοψηφία των πελατών είναι άνδρες.



Εικόνα 8 - Πελάτες ανά φύλο

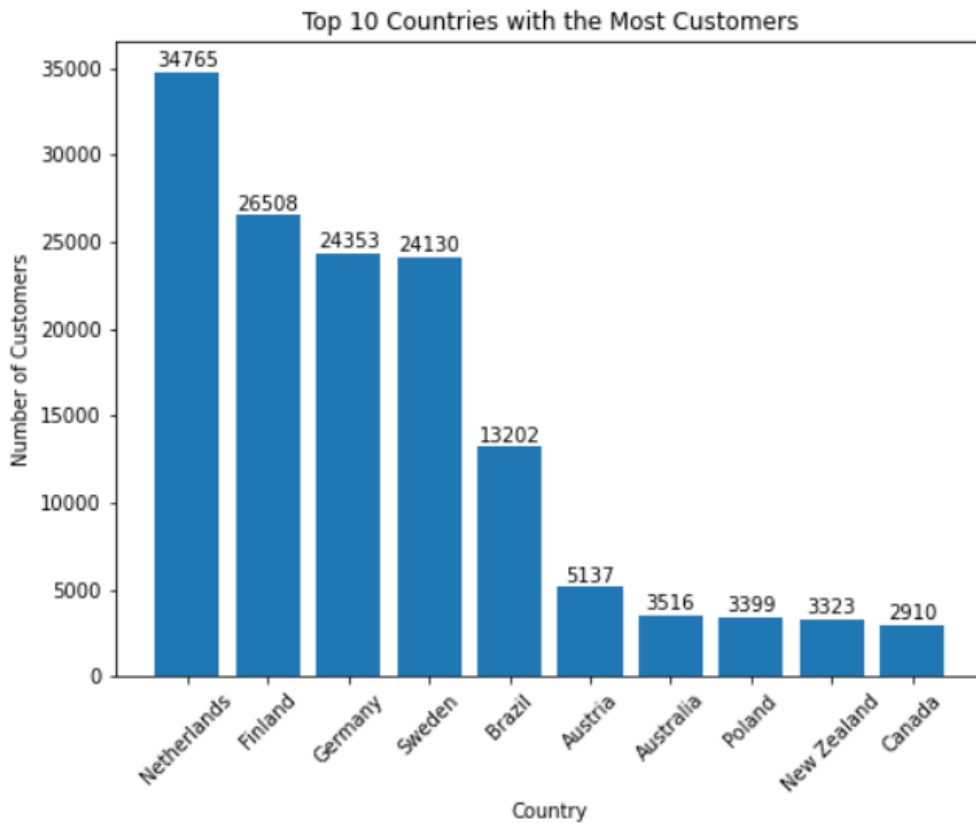
Εξετάζοντας τους πελάτες ως προς την ηλικία τους, και μάλιστα βάσει των ηλικιακών ομάδων που δημιουργήθηκαν, παρατηρείται ότι οι περισσότεροι πελάτες είναι νέοι και συγκεκριμένα από 18 μέχρι 39 ετών.

Customer Age Group Distribution



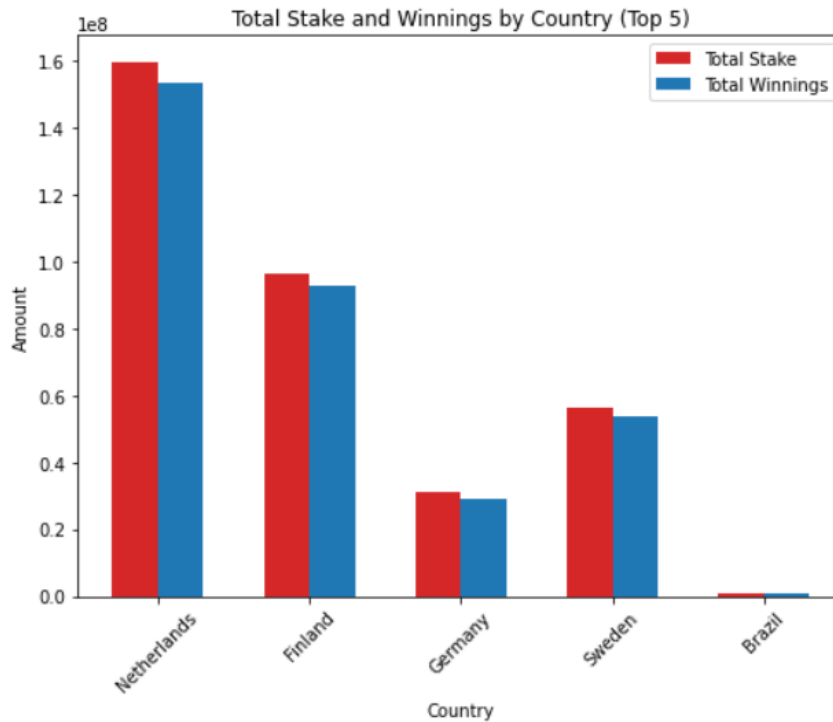
Εικόνα 9 - Ηλικιακές ομάδες

Στη συνέχεια, εξετάζοντας τη χώρα εγγραφής των πελατών και συγκεκριμένα παρατηρώντας τις 10 χώρες με τις περισσότερες εγγραφές, βλέπουμε ότι η πλειοψηφία των πελατών είναι από τις χώρες Φινλανδία, Σουηδία, Ολλανδία και Γερμανία.



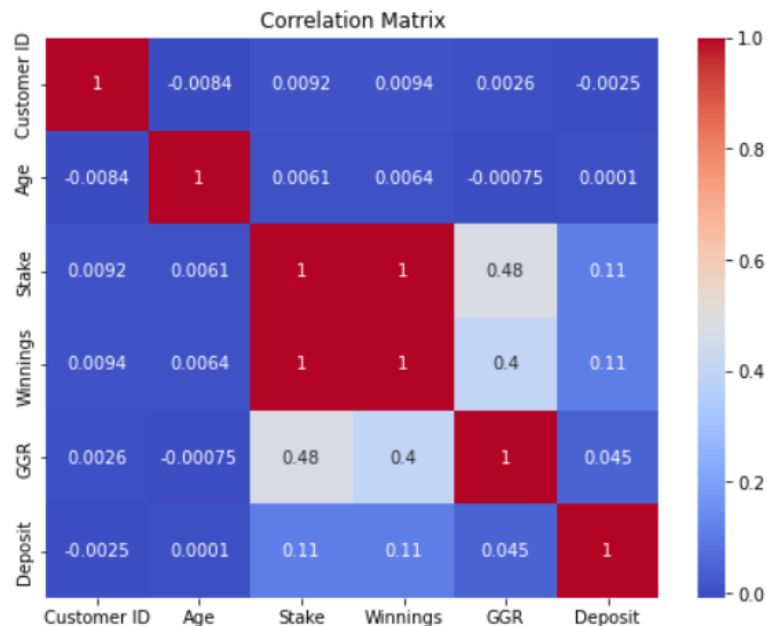
Εικόνα 10 - Πελάτες ανά χώρα εγγραφής

Η Φινλανδία είναι πρώτη σε αριθμό πελατών, με αριθμό λίγο μεγαλύτερο από τη Σουηδία, αλλά αρκετά μεγαλύτερο από αυτόν την Ολλανδίας. Ωστόσο, παρατηρείται ότι τόσο η Σουηδία, όσο και η Ολλανδία έχουν αρκετά μεγαλύτερα ποσά πονταρίσματος και κερδών. Αυτό σημαίνει ότι οι πελάτες από τη Φινλανδία συνηθίζουν να παίζουν με μικρότερα ποσά.



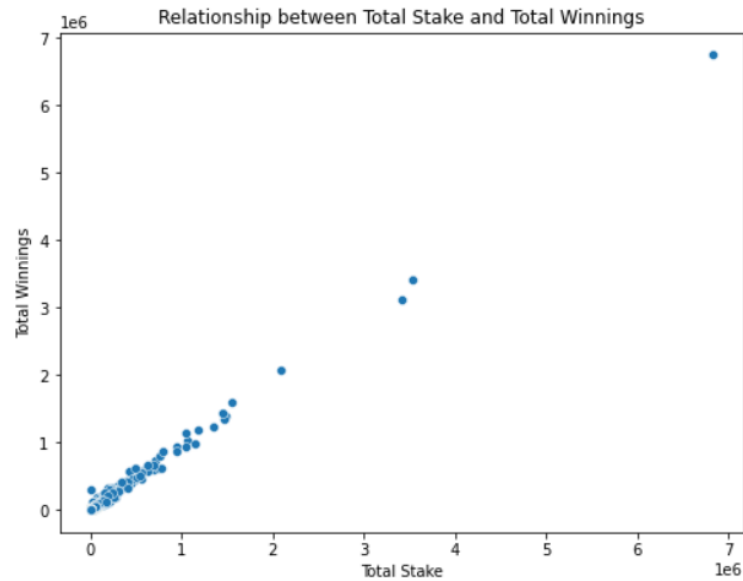
Εικόνα 11 - Συνολικά ποσά πονταρίσματος και κερδών ανά χώρα

Εν συνεχεία, δημιουργήθηκε ένας πίνακας συσχετίσεων (correlation matrix) μεταξύ των αριθμητικών μεταβλητών, όπως φαίνεται παρακάτω. Όπως ήταν αναμενόμενο, τα ποσά Stake και Winnings έχουν τη μέγιστη συσχέτιση και τα μεγέθη Stake και GGR έχουν μία μέτρια συσχέτιση μεταξύ τους.

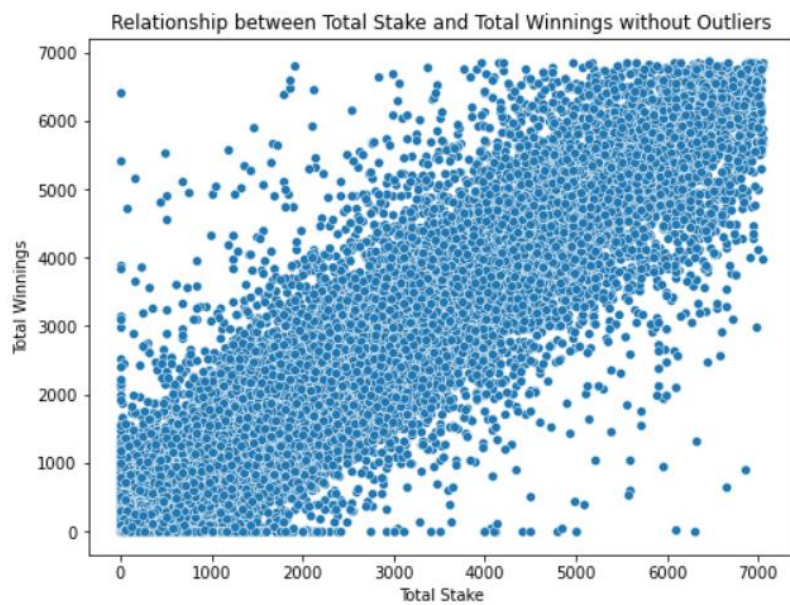


Εικόνα 12 - Πίνακας συσχετίσεων αριθμητικών μεταβλητών

Επομένως, αναλύθηκε παραπάνω η συσχέτιση μεταξύ Stake και Winnings των πελατών. Για τον λόγο αυτό, δημιουργήθηκε ένα scatterplot, το οποίο εν συνεχεία αναλύθηκε περισσότερο, αγνοώντας τις ακραίες τιμές.

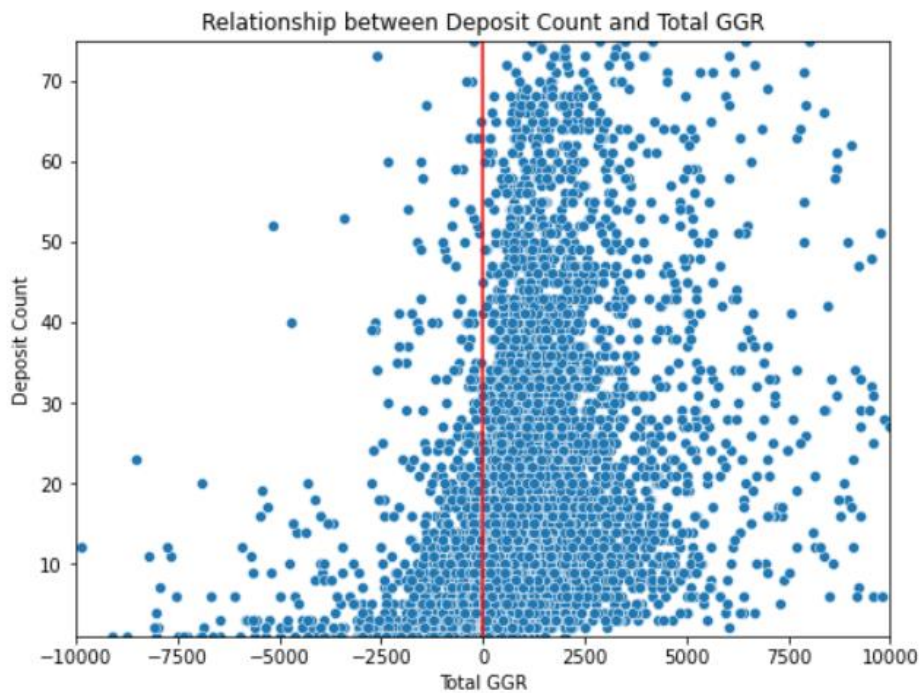


Εικόνα 13 - Σχέση μεταξύ Stake και Winnings



Εικόνα 14 - Σχέση μεταξύ Stake και Winnings χωρίς τις ακραίες τιμές

Παρακάτω παρατίθεται το scatterplot που δημιουργήθηκε για τη σχέση μεταξύ του αριθμού των καταθέσεων που πραγματοποιεί ένας πελάτης σε σχέση με το συνολικό GGR του.



Εικόνα 15 - Σχέση μεταξύ συνολικού GGR και αριθμού καταθέσεων

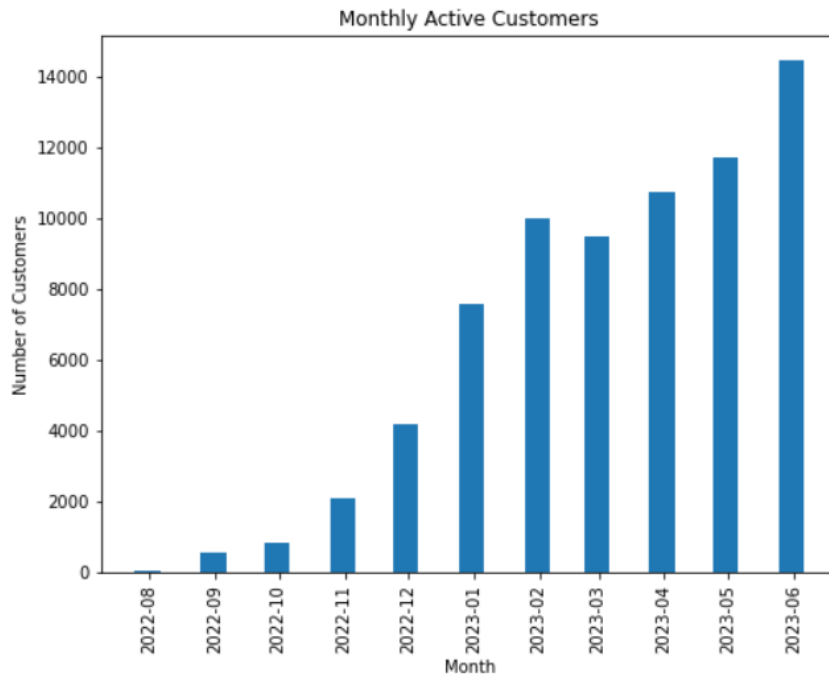
Παρατηρώντας το ανωτέρω scatterplot, είναι εμφανές ότι για τις θετικές τιμές του GGR, πραγματοποιούνται περισσότερες καταθέσεις από τους πελάτες. Συμπέρασμα το οποίο ήταν αναμενόμενο, καθώς όσο οι παίκτες χάνουν τα χρήματά τους παίζοντας, δηλαδή έχουν θετικό GGR, πραγματοποιούν νέες καταθέσεις. Από την άλλη, οι πελάτες οι οποίοι είναι κερδισμένοι, δηλαδή αυτοί οι οποίοι έχουν αρνητικό GGR, πραγματοποιούν στη συνέχεια λιγότερες καταθέσεις, είτε λόγω ότι έχουν χρήματα στον λογαριασμό τους για να συνεχίσουν να παίζουν, είτε λόγω ότι πραγματοποιούν ανάληψη των κερδών τους και σταματούν να παίζουν.

5.6 Εξαγωγή Χαρακτηριστικών

Μετά την πλήρη αναγνώριση των δεδομένων, δημιουργήθηκε μια σφαιρική εικόνα της επιχειρησιακής κατάστασης της εταιρείας και των δεδομένων που περιλαμβάνονται σε αυτήν.

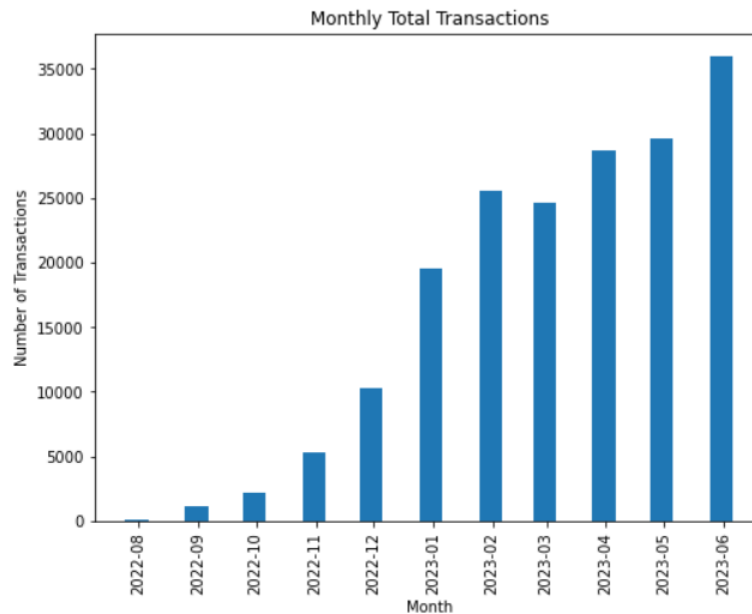
Ως εκ τούτου, είναι δυνατό να εξάγουμε χρήσιμα χαρακτηριστικά που απαιτούνται για τη δημιουργία ενός αξιόπιστου και σωστού μοντέλου. Σε αυτήν την περίπτωση, η ανάλυση επικεντρώνεται στις καταθέσεις που πραγματοποιούνται ως μέτρο ανάπτυξης των εσόδων της εταιρείας.

Η πληροφορία σχετικά με τους πελάτες που είχαν ενεργή δραστηριότητα κάθε μήνα είναι σημαντική για την ανάλυση της κινητικότητας της πελατείας της εταιρείας. Αυτή η μετρική μπορεί να παρέχει χρήσιμη πληροφορία.



Εικόνα 16 – Ενεργοί πελάτες ανά μήνα

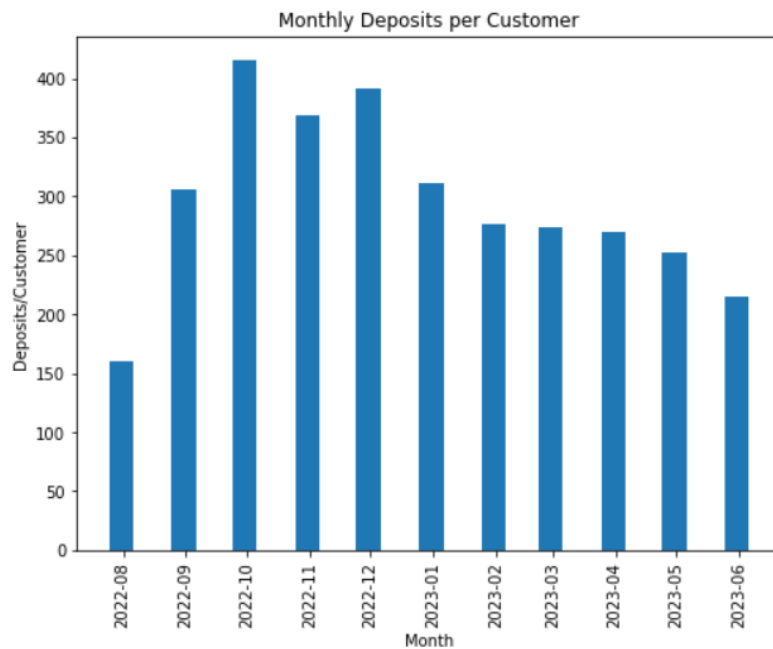
Στη συνέχεια, αναπτύχθηκε μία μετρική για την παρακολούθηση του βασικού αντικειμένου ανάλυσης, δηλαδή των καταθέσεων. Αυτή η μετρική είναι ο αριθμός των μηνιαίων συναλλαγών. Χρησιμοποιώντας αυτόν τον αριθμό συναλλαγών ανά μήνα, σε συνδυασμό με τους ενεργούς πελάτες ανά μήνα, μπορεί να γίνει εντοπισμός των πελατών που προσφέρουν μεγαλύτερη αξία στην εταιρεία όσον αφορά το κέρδος.



Εικόνα 17 – Πλήθος συναλλαγών ανά μήνα

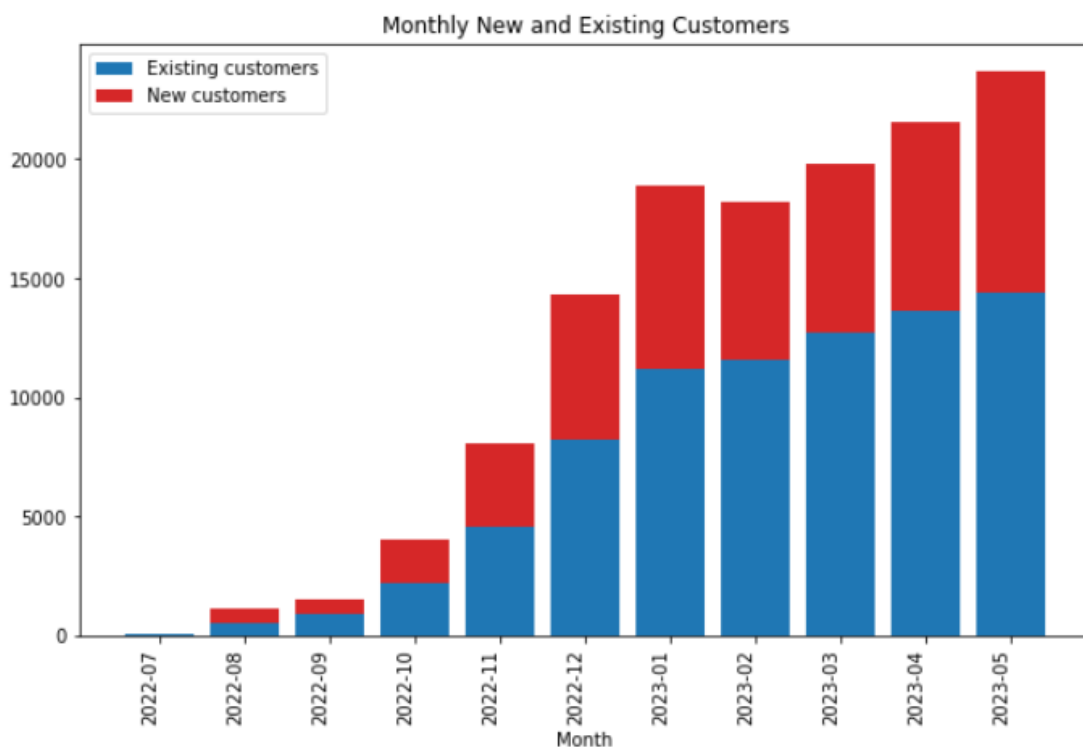
Από τα ανωτέρω γραφήματα, παρατηρούμε ότι το πλήθος των συναλλαγών αυξάνεται αναλογικά με το πλήθος των ενεργών πελατών.

Η επόμενη μετρική υπολογίζει τη μέση μηνιαία αξία των καταθέσεων ανά πελάτη.



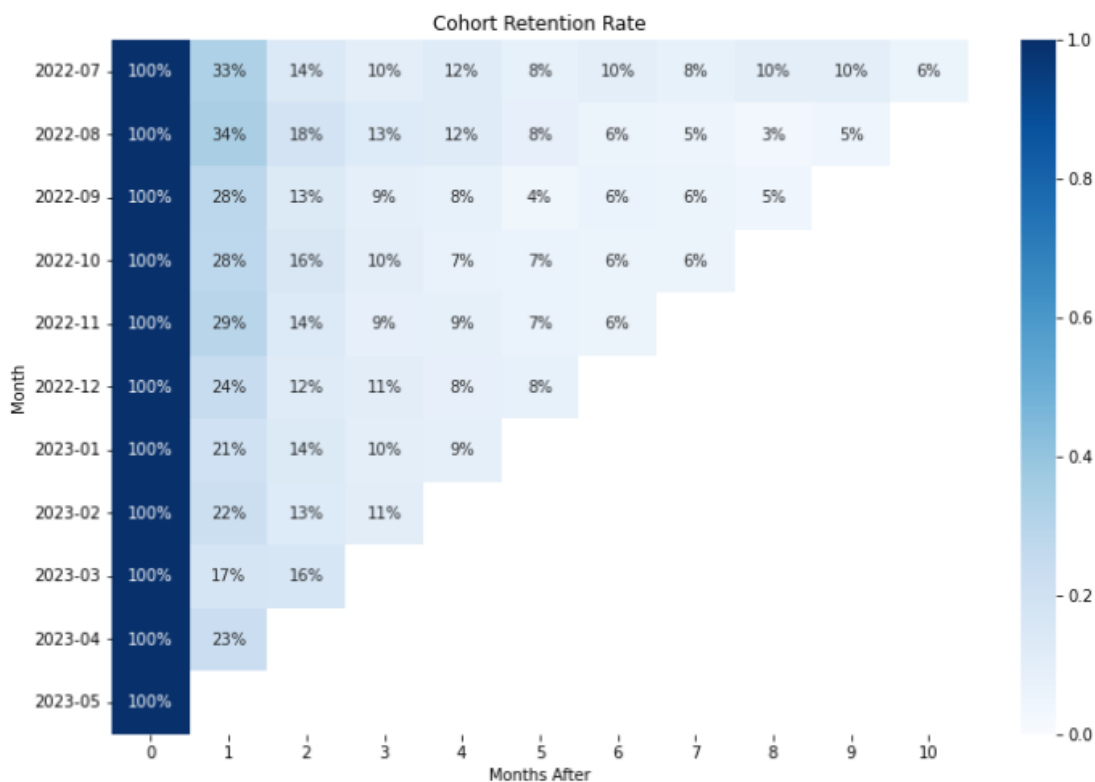
Εικόνα 18 – Μέση μηνιαία κατάθεση ανά πελάτη

Η επόμενη μετρική στοχεύει στην αξιολόγηση της αύξησης του πελατολογίου της εταιρείας και τη σύγκριση των νέων πελατών με τους υφιστάμενους. Αυτό επιτυγχάνεται μέσω της μέτρησης του ποσοστού των νέων πελατών σε σχέση με τον συνολικό αριθμό πελατών της εταιρείας. Η μετρική αυτή είναι χρήσιμη για την αξιολόγηση της αποτελεσματικότητας των προσπαθειών της εταιρείας στην προσέλκυση νέων πελατών και στη διατήρηση των υφιστάμενων.



Εικόνα 19 – Νέοι και υφιστάμενοι πελάτες ανά μήνα

Εν συνεχεία, θα χρησιμοποιηθεί μια προσέγγιση για τον υπολογισμό του ποσοστού διατήρησης πελατών, η οποία βασίζεται στη μέτρηση του ποσοστού πελατών που επανέλαβαν καταθέσεις στον επόμενο μήνα μετά την αρχική τους κατάθεση (Cohort Based Retention Rate). Ωστόσο, από τα αποτελέσματα που φαίνονται στο παρακάτω διάγραμμα, παρατηρείται ότι το ποσοστό των πελατών που διατηρούν τη σχέση τους με την εταιρεία κυμαίνεται στο 25-30% στον δεύτερο μήνα, το οποίο μειώνεται στο 10-15% στον τρίτο μήνα. Πράγμα το οποίο σημαίνει ότι το μεγαλύτερο μέρος των καταθέσεων πραγματοποιείται από νέους πελάτες.



Εικόνα 20 – Retention Rate

5.7 Κατηγοριοποίηση των πελατών με χρήση της μεθόδου RFM

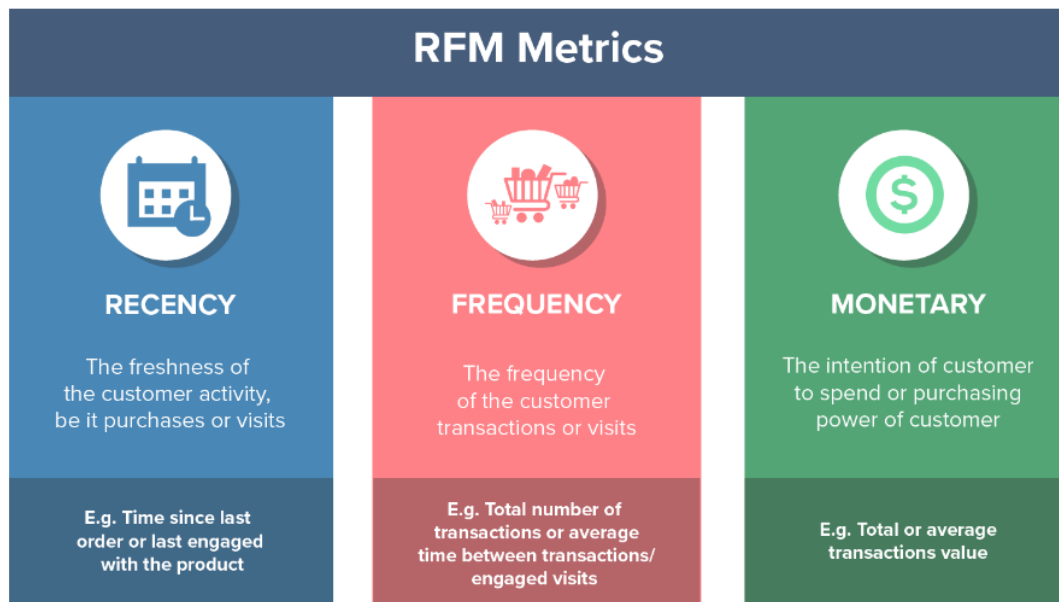
Στο προηγούμενο κεφάλαιο, αναλύθηκαν οι βασικοί παράγοντες που επηρεάζουν την κερδοφορία και τη λειτουργία της εταιρείας. Μέσα από αυτή την ανάλυση, δημιουργήθηκε μια εικόνα για τα δεδομένα, που μας δίνει τη δυνατότητα να αναλύσουμε σε βάθος το πελατολόγιο της εταιρείας.

Στο παρόν κεφάλαιο θα αναλυθεί η σημασία των πελατών στην κερδοφορία μιας επιχείρησης. Θα προσεγγιστεί η εταιρεία ως μια εμπορική επιχείρηση και θα γίνει κατηγοριοποίηση των πελατών βάσει της κερδοφορίας που αποφέρουν και της αξίας που προσδίδουν στην εταιρεία. Μέσα από αυτή την ανάλυση, θα αποκτηθεί μια πιο εμπειριστατωμένη εικόνα για το πελατολόγιο της εταιρείας και τους τρόπους με τους οποίους μπορεί να αυξηθεί η κερδοφορία της.

Η πελατοκεντρική επικεντρώνεται στους πελάτες ως τον καθοριστικό παράγοντα για την επίτευξη κερδών και επιδιώκει να αναγνωρίσει τους διαφορετικούς τύπους πελατών που

υπάρχουν σε μία επιχείρηση. Αυτό σημαίνει ότι δεν μπορεί να υπάρχει μία γενική στρατηγική που να εφαρμόζεται σε όλους τους πελάτες, αλλά απαιτείται μία προσαρμοσμένη στρατηγική που θα λαμβάνει υπόψη τις διαφορές μεταξύ τους.

Για την υλοποίηση αυτής της προσαρμοσμένης στρατηγικής, απαιτείται η κατηγοριοποίηση των πελατών με βάση την αξία που προσφέρουν στην επιχείρηση και τα χαρακτηριστικά τους. Με αυτόν τον τρόπο, μπορεί να δημιουργηθεί μία πιο αποτελεσματική στρατηγική που θα εστιάζει στους πελάτες που παρέχουν τη μεγαλύτερη αξία στην επιχείρηση και θα προσφέρει εξατομικευμένες υπηρεσίες σε αυτούς.



Εικόνα 21 – Βασικές μετρικές μεθόδου RFM
Customer Segmentation with RFM Analysis & K-Means Clustering – Medium.com

Σε αυτήν την ανάλυση, θα χρησιμοποιηθεί η μέθοδος RFM για να διαχωρίσουμε τους πελάτες σε τρεις κατηγορίες: χαμηλής αξίας, μεσαίας αξίας και υψηλής αξίας. Η μέθοδος RFM θεωρείται μία από τις πιο ασφαλείς και αποδοτικές μεθόδους ανάλυσης. Οι τρεις βασικές κατηγορίες, στις οποίες διαχωρίζονται οι πελάτες είναι:

- **Χαμηλής αξίας:**
Οι πελάτες που ανήκουν στην κατηγορία χαμηλής αξίας είναι λιγότερο ενεργοί και έχουν χαμηλή συχνότητα συναλλαγών και αμελητέο κέρδος.
- **Μεσαίας αξίας:**
Οι πελάτες μεσαίας αξίας έχουν μέτριες επιδόσεις όσον αφορά την αλληλεπίδρασή τους με την επιχείρηση και μεσαίο αποδιδόμενο κέρδος.
- **Υψηλής αξίας:**
Οι πελάτες υψηλής αξίας αντιπροσωπεύουν τους πιο σημαντικούς πελάτες μιας επιχείρησης, με συχνές και επικερδείς συναλλαγές που αποτελούν κεφάλαιο για την επιχείρηση και συμβάλλουν στην ανάπτυξή της.

Όπως αναλύθηκε και στο προηγούμενο κεφάλαιο, η μέθοδος RFM χρησιμοποιεί τρεις βασικούς παράγοντες - την εγγύτητα (Recency), τη συχνότητα (Frequency) και τη νομισματική αξία (Monetary) – για να κατηγοριοποιήσει τους πελάτες σε διαφορετικά επίπεδα αξίας. Η υλοποίηση της μεθόδου αυτής γίνεται με χρήση εργαλείων και τεχνικών μηχανικής μάθησης, όπως ο αλγόριθμος K-means, ο οποίος χρησιμοποιείται για την αναγνώριση των βασικών ομάδων (clusters) για κάθε ένα από τα τρία μεγέθη (εγγύτητα, συχνότητα και νομισματική αξία). Από αυτές τις κατηγορίες μπορεί να προκύψει μια συνολική εικόνα των πελατών και να ληφθούν αποφάσεις για τη βελτίωση της στρατηγικής πωλήσεων και μάρκετινγκ.

Για την εφαρμογή της μεθόδου RFM και τον υπολογισμό του δείκτη Customer Lifetime Value (CLV), δημιουργήθηκε ένα υποσύνολο των δεδομένων, το οποίο περιλαμβάνει τις μεταβλητές εκείνες που αντιπροσωπεύουν τα μεγέθη τα οποία σχετίζονται τόσο με τις απαραίτητες πληροφορίες για τους πελάτες όσο και με τα ποσοτικά και ποιοτικά μεγέθη των συναλλαγών που πραγματοποιούν.

5.7.1 Εγγύτητα (Recency)

Για να υπολογιστεί ο δείκτης Εγγύτητας (Recency) για κάθε πελάτη, χρειάζεται να βρεθεί η πιο πρόσφατη ημερομηνία συναλλαγής του και να υπολογιστούν οι ημέρες αδράνειας από την προηγούμενη συναλλαγή. Στη συνέχεια, θα χρησιμοποιηθεί ο αλγόριθμος συσταδοποίησης K-means για να ομαδοποιηθούν οι πελάτες σε διαφορετικές ομάδες βάσει του αριθμού των ημερών αδράνειας. Από αυτές τις ομάδες θα προκύψει η βαθμολογία για τον δείκτη Recency.

count	47094.000000
mean	89.634518
std	66.770614
min	0.000000
25%	31.000000
50%	81.000000
75%	138.000000
max	309.000000

Εικόνα 22 – Περιγραφική στατιστική Recency

Στα πλαίσια του υπολογισμού της Εγγύτητας, προέκυψε ότι η πλειοψηφία των πελατών διατηρούν πολύ μικρή σχέση με την εταιρεία, καθώς η μέση τιμή των ημερών που μεσολαβούν μεταξύ των συναλλαγών ενός πελάτη φτάνει τις 90 ημέρες.

Στη συνέχεια, ακολούθησε η εφαρμογή του αλγορίθμου K-means για τη δημιουργία μιας βαθμολογίας Εγγύτητας μέσω της ομαδοποίησης των πελατών. Για να γίνει αυτό, απαιτείται ο προσδιορισμός του αριθμού των συστάδων που θα δημιουργηθούν. Για τον σκοπό αυτό, χρησιμοποιήθηκε η μέθοδος "Elbow Rule".

Ο «κανόνας του αγκώνα» είναι μία προσέγγιση προσδιορισμού του αριθμού των συστάδων αξιοποιώντας το κέρδος της πληροφορίας που προστίθεται στις κλάσεις και το ποσοστό της εξηγούμενης διακύμανσης.

Βάσει δοκιμών και ελέγχων που πραγματοποιήθηκαν, προέκυψε ότι με την χρήση 4 συστάδων επιτυγχάνεται η καλύτερη απόδοση του αλγορίθμου.

	count	mean	std	min	25%	50%	75%	max
RecencyCluster								
0	11825.0	13.229767	9.614854	0.0	4.0	12.0	21.0	31.0
1	11726.0	53.892802	14.868172	32.0	40.0	53.0	67.0	81.0
2	11910.0	110.129807	16.235522	82.0	96.0	111.0	124.0	138.0
3	11633.0	182.344451	34.854503	139.0	155.0	176.0	200.0	309.0

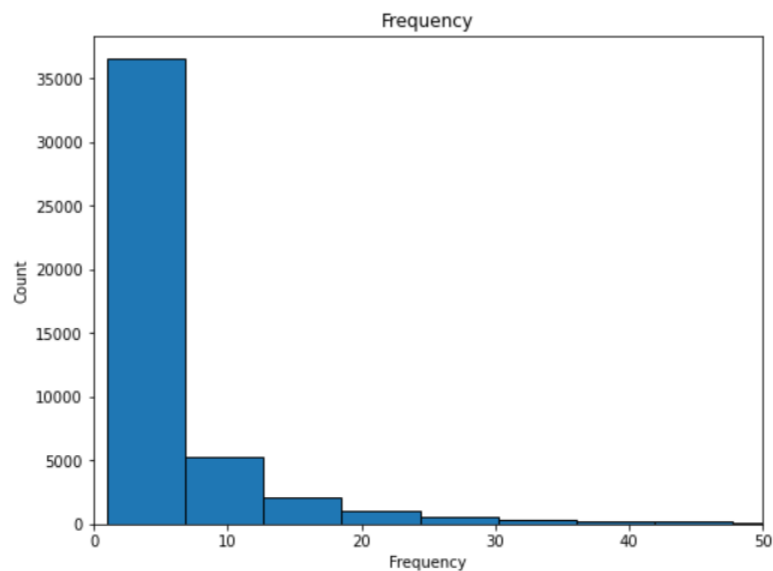
Εικόνα 23 – Recency K-means

Παρατηρούμε ότι οι πελάτες μοιράζονται σχεδόν ισόποσα μεταξύ των 4 clusters.

Ωστόσο και οι 4 συστάδες έχουν σχετικά μεγάλο μέσο όρο των ημερών που μεσολαβούν από την τελευταία φορά που ένας πελάτης έκανε κατάθεση.

5.7.2 Συχνότητα (Frequency)

Για να υπολογιστεί η συχνότητα (Frequency) των συναλλαγών ενός πελάτη, θα ακολουθηθεί η ίδια διαδικασία συσταδοποίησης όπως για τον προσδιορισμό της Εγγύτητας. Συγκεκριμένα, θα δημιουργηθούν αντίστοιχα clusters που θα περιλαμβάνουν τον συνολικό αριθμό των συναλλαγών του εκάστοτε πελάτη. Αξίζει να σημειωθεί ότι η καταμέτρηση των συναλλαγών θα πραγματοποιηθεί επί των ημερομηνιών της κάθε συναλλαγής, δηλαδή θα μετρηθεί σε πόσες διαφορετικές ημερομηνίες ένας πελάτης αλληλοεπιδρούσε με την εταιρεία.



Εικόνα 24 – Διάγραμμα Frequency

Αναλύοντας τον συνολικό αριθμό των συναλλαγών των πελατών, παρατηρούμε ότι η πλειονότητα των πελατών πραγματοποιεί μικρό αριθμό συναλλαγών, οπότε η συχνότητά τους είναι χαμηλή. Αυτό καθιστά τη διάκριση των πελατών με βάση την κερδοφορία λίγο πιο δύσκολη, αλλά μπορούμε να χρησιμοποιήσουμε έναν αλγόριθμο ομαδοποίησης για να τους κατατάξουμε σε διαφορετικά clusters. Παρατηρούμε στον παρακάτω πίνακα ότι το cluster 2

έχει τις χαμηλότερες συχνότητες, υποδηλώνοντας ότι υπάρχει περιορισμένος αριθμός πελατών στην υψηλότερη κατηγορία κερδοφορίας.

	count	mean	std	min	25%	50%	75%	max
FrequencyCluster								
0	25860.0	1.259474	0.438355	1.0	1.0	1.0	2.0	2.0
1	10678.0	4.129519	1.083991	3.0	3.0	4.0	5.0	6.0
2	10556.0	20.815271	28.255842	7.0	9.0	12.0	21.0	585.0

Εικόνα 25– Frequency K-means

5.7.3 Νομισματική Αξία (Monetary)

Για τον υπολογισμό της νομισματικής αξίας (Monetary Value) του πελάτη, ή αλλιώς το κέρδος που επιφέρει ένας πελάτης στην επιχείρηση, χρησιμοποιήθηκε το ποσό πονταρίσματος των πελατών και το ποσό των καταθέσεών τους.

Χρησιμοποιώντας την ίδια λογική με αυτήν που έχει εφαρμοστεί ο αλγόριθμος συσταδοποίησης στο Recency και στο Frequency, εφαρμόστηκε και στο Revenue.

	count	mean	std	min	25%	50%	75%	max
MonetaryCluster								
0	11958.0	20.734461	7.751512	1.67864	14.0	20.404682	25.0	35.0
1	14297.0	67.434524	21.932982	35.29420	50.0	60.000000	90.0	100.0
2	9103.0	178.313275	48.307578	100.05000	140.0	175.000000	210.0	280.0
3	11736.0	1428.160648	3109.688495	281.00000	410.0	665.000000	1355.0	150820.0

Εικόνα 26 – Monetary Value K-means

Ως εκ τούτου, παρατηρώντας των ανωτέρω πίνακα προκύπτει ότι το cluster 3 είναι αυτό που περιλαμβάνει τους πελάτες, οι οποίοι αποδίδουν το μεγαλύτερο κέρδος στην επιχείρηση.

5.7.4 Συνολική Βαθμολόγηση

Η μέθοδος RFM χρησιμοποιείται για την αξιολόγηση της σημαντικότητας των πελατών μιας επιχείρησης βασιζόμενη στα μεγέθη της συχνότητας των συναλλαγών, του ποσού των αγορών και του χρόνου που έχει περάσει από την τελευταία αγορά τους.

Μετά τον υπολογισμό των τριών αυτών μεγεθών, μπορεί να υπολογιστεί μια συνολική βαθμολογία για τον κάθε πελάτη, βασιζόμενη στις τρεις διακριτές εφαρμογές του αλγορίθμου K-means στα μεγέθη της μεθόδου RFM (Recency – Frequency – Monetary Value).

Η διαδικασία αυτή παράγει έναν αριθμό σκορ, ο οποίος αντιστοιχεί στο άθροισμα του αριθμού των clusters στα οποία ανήκει ο πελάτης.

Όσο πιο μικρός είναι ο αριθμός του cluster στον οποίο ανήκει ο πελάτης, τόσο λιγότερο σημαντική είναι η αξία που αποδίδει στην επιχείρηση.

Η συνολική βαθμολογία κατανέμεται σε βαθμίδες, με τους πελάτες που έχουν τη μέγιστη βαθμολογία να είναι αυτοί με τη μεγαλύτερη σημασία, ενώ οι πελάτες με την ελάχιστη βαθμολογία να έχουν την χαμηλότερη σημασία για την επιχείρηση.

Για σκοπούς καλύτερης μοντελοποίησης του ανωτέρω συστήματος, οι βαθμίδες που προέκυψαν ομαδοποιήθηκαν στις ακόλουθες τέσσερις:

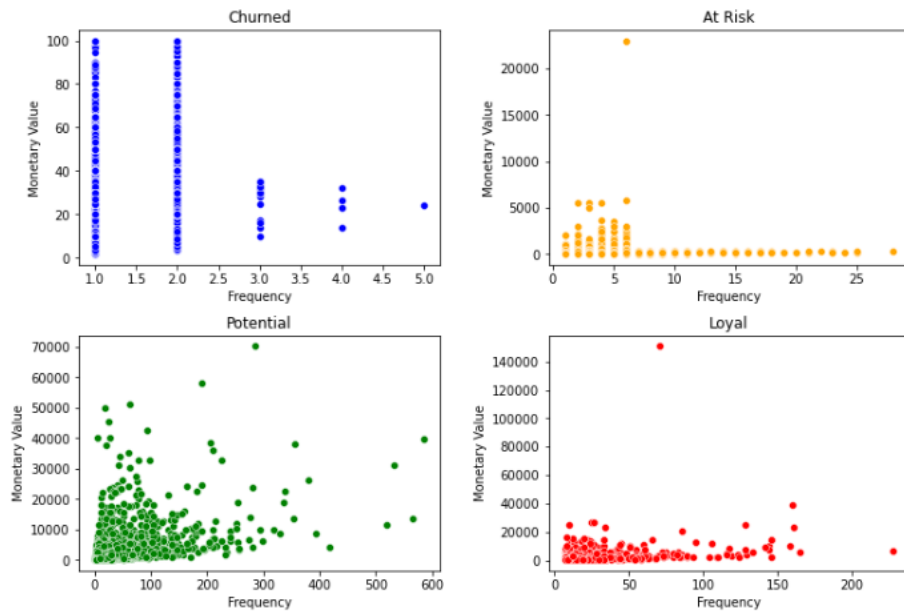
- Churned
- At Risk
- Potential
- Loyal

Η ομαδοποίηση αυτή βοηθάει στην τελική τμηματοποίηση των πελατών βάσει της συνολικής τους βαθμολογίας.

	Recency	Frequency	MonetaryValue	RFM Score
Segment				
At Risk	97.70	2.24	98.20	2.98
Churned	29.18	1.20	33.39	0.67
Loyal	182.42	18.56	1682.48	8.00
Potential	99.49	14.22	991.67	5.74

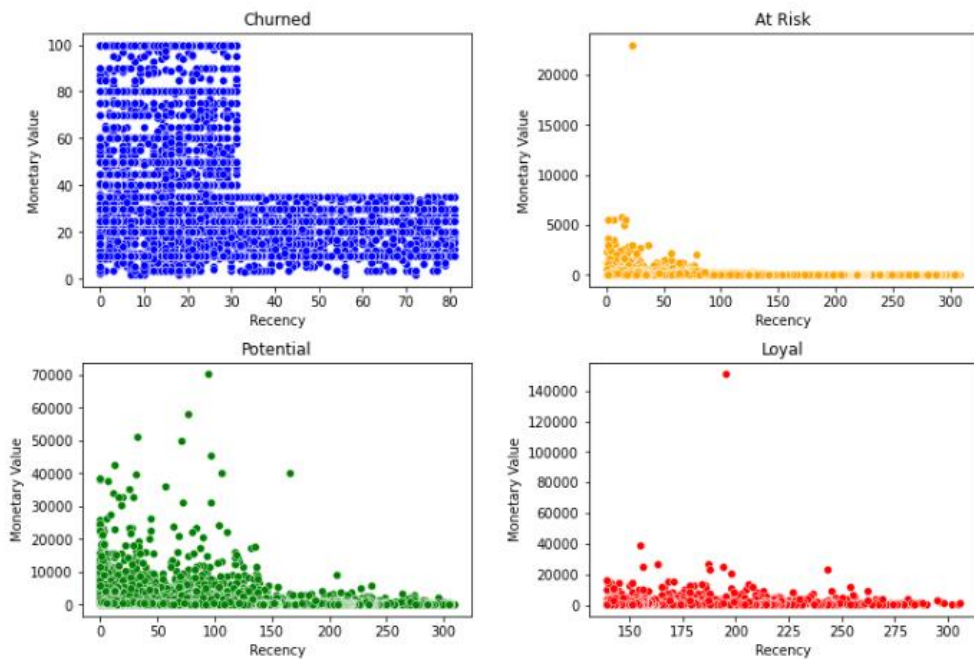
Εικόνα 27 – Τμηματοποίηση πελατών βάσει RFM

Τα ακόλουθα γραφήματα είναι αρκετά χρήσιμα στην κατηγοριοποίηση των πελατών, με βάση την αξία τους.



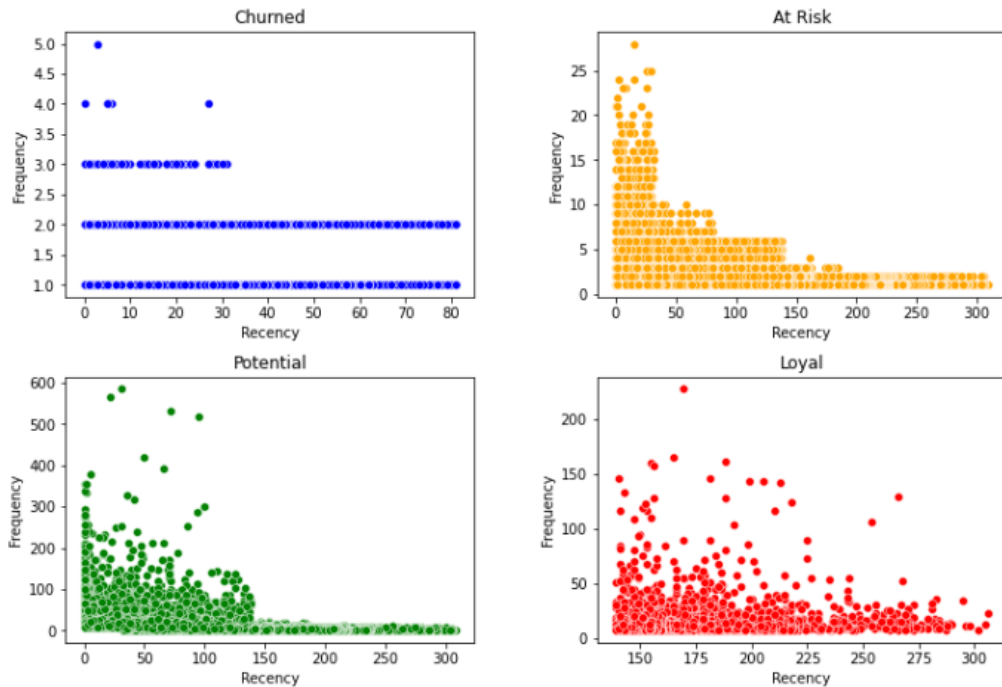
Εικόνα 28 – Monetary Value VS Frequency

Από τα ανωτέρω γραφήματα, μπορούμε να παρατηρήσουμε ότι οι Churned και οι At Risk πελάτες πραγματοποιούν ελάχιστες καταθέσεις στις πρώτες ημέρες από την εγγραφή τους, για αυτό είναι μικροί οι αριθμοί της συχνότητας. Επομένως, είναι πιο χρήσιμο να εστιάσουμε στους Potential και στους Loyal πελάτες. Όσον αφορά τους Potential, βλέπουμε, κατά κύριο λόγο, ότι η αξία μειώνεται, όσο αυξάνεται η συχνότητα των καταθέσεων, πράγμα το οποίο σημαίνει ότι η πλειοψηφία αυτών των πελατών επιλέγουν να κάνουν είτε λίγες αλλά μεγάλες, σε ποσό, καταθέσεις, είτε συχνότερες αλλά μικρότερων ποσών. Ένα παρόμοιο φαινόμενο παρατηρείται και στους Loyal πελάτες, αλλά σε αυτήν την περίπτωση αναφερόμαστε σε μεγαλύτερα ποσά καταθέσεων. Αυτό που παρατηρείται επομένως, είναι ότι υπάρχει περιθώριο βελτίωσης όσον αφορά την αναλογία συχνότητας και αξίας στους πελάτες «υψηλής αξίας».



Εικόνα 29 – Monetary Value VS Recency

Όσον αφορά τα παραπάνω γραφήματα, παρατηρούμε στο σύνολό τους ότι η σχέση κέρδους-εγγύτητας δεν κυμαίνεται σε αρκετά ικανοποιητικά επίπεδα. Το ζητούμενο είναι υψηλές τιμές κέρδους σε συνδυασμό με χαμηλές τιμές εγγύτητας, όμως στο σύνολο των γραφημάτων παρατηρούμε αρκετά υψηλές τιμές εγγύτητας.



Εικόνα 30 – Frequency VS Recency

Στα γραφήματα της συχνότητας σε συνάρτηση με την εγγύτητα, παρατηρείται μία παρόμοια λογική. Εστιάζοντας στους Potential και στους Loyal πελάτες, κατά κύριο λόγο, βλέπουμε αρκετά υψηλές τιμές εγγύτητας, με όχι τόσο αυξημένες τιμές συχνότητας. Ωστόσο, συγκεκριμένα στους Loyal πελάτες, παρατηρούμε ότι ένα μέρος τους έχει αυξημένη συχνότητα, με μικρότερη εγγύτητα, επομένως πρόκειται για τους πελάτες, οι οποίοι είναι και οι πιο σημαντικοί, από άποψη κέρδους για την επιχείρηση.

Γενικά, αξίζει να σημειωθεί ότι η εφαρμογή της μεθόδου RFM είναι πολύ αποτελεσματική στην ταξινόμηση των πελατών και της συναλλακτικής τους συμπεριφοράς. Αυτό συμβαίνει, καθώς με αυτόν τον τρόπο, η επιχείρηση μπορεί να προσαρμόσει τη στρατηγική της ανάλογα με τις μετρήσεις και τα μεγέθη που παρέχει η εν λόγω μέθοδος.

Κατά αυτόν τον τρόπο, μπορούν να δημιουργηθούν επιχειρηματικές αποφάσεις και μοντέλα δράσης της επιχείρησης, με σκοπό τη βέλτιστη διαχείριση των πελατών, τόσο συνολικά, όσο και ανά κατηγορία πελατών, με βάση την αξία που προσδίδουν στην επιχείρηση.

5.8 Μοντελοποίηση Customer Lifetime Value

Στην προηγούμενη παράγραφο, οι πελάτες κατηγοριοποιήθηκαν σε τέσσερις βασικές κλάσεις. Σε αυτό το σημείο, μπορεί να μετρηθεί η συμπεριφορά της εκάστοτε κλάσης, έτσι ώστε να ληφθούν οι κατάλληλες επιχειρησιακές αποφάσεις.

Ο ορισμός της αξίας του πελάτη μπορεί να διαφέρει από επιχείρηση σε επιχείρηση και από περίπτωση σε περίπτωση, επομένως ο προσδιορισμός της αποτελεί μια δυναμική διαδικασία που εξαρτάται από τις στρατηγικές που εφαρμόζονται. Σε αυτή τη διαδικασία, διαδραματίζουν σημαντικό ρόλο, οι μέθοδοι marketing, οι ενέργειες απόκτησης νέων πελατών, καθώς και διατήρησης των υφιστάμενων πελατών, αλλά και η βελτιστοποίηση του κέρδους από κάθε συγκεκριμένη κατηγορία πελατών.

Υπό αυτό το πλαίσιο, ο δείκτης CLV αποτελεί ένα χρήσιμο εργαλείο για την αναγνώριση της συμπεριφοράς των συναλλαγών των πελατών, αλλά και για την καταγραφή των χαρακτηριστικών των πελατών.

Απαραίτητη προϋπόθεση για την εφαρμογή του CLV είναι να καθοριστεί η χρονική περίοδος κατά την οποία μετράται. Αυτή η περίοδος μπορεί να κυμαίνεται από 3 μήνες έως 24 μήνες. Μια περίοδος μέτρησης περίπου 15 μηνών κρίθηκε κατάλληλη για τους σκοπούς αυτής της ανάλυσης, λαμβάνοντας υπόψη τη συνήθη εμπορική πρακτική.

Η μέτρηση του CLV βασίζεται στον υπολογισμό της μεθόδου RFM που εφαρμόστηκε στην προηγούμενη παράγραφο. Ως εκ τούτου, είναι απαραίτητο να χρησιμοποιηθεί η βαθμολογία που έχει εκχωρηθεί σε κάθε πελάτη (Customer ID) για τις τρεις διαστάσεις της μεθόδου (Recency – Frequency – Monetary Value).

	Customer ID	Recency	Frequency	MonetaryValue	RecencyCluster	FrequencyCluster	MonetaryCluster	RFM	RFM Score	Segment
0	10005	35	4	160.0	1	1	2	112	4	At Risk
1	10006	142	3	78.0	3	1	1	311	5	Potential
2	10007	1	5	225.0	0	1	2	012	3	At Risk
3	10010	1	17	1290.0	0	2	3	023	5	Potential
4	10023	104	4	95.0	2	1	1	211	4	At Risk
...
47089	168153	142	5	132.0	3	1	2	312	6	Potential
47090	168156	124	3	315.0	2	1	3	213	6	Potential
47091	168157	142	1	10.0	3	0	0	300	3	At Risk
47092	168161	121	9	404.0	2	2	3	223	7	Potential
47093	168165	34	1	50.0	1	0	1	101	2	At Risk

Εικόνα 31 – Πίνακας κατηγοριοποίησης

Με βάση την κατηγοριοποίηση των πελατών που πραγματοποιήθηκε στην προηγούμενη παράγραφο, δημιουργήθηκε ο παραπάνω πίνακας με τη βαθμολογία που προκύπτει από τη μέθοδο RFM για κάθε πελάτη.

Το ανωτέρω αποτελεί και το σύνολο δεδομένων, το οποίο χρησιμοποιείται για τον υπολογισμό του δείκτη Customer Lifetime Value για το χρονικό διάστημα που ορίστηκε.

Για τον υπολογισμό του CLV, μεταβλητή στόχο αποτελεί το μέγεθος Lifetime Value (LTV), το οποίο ορίζεται από τον τύπο:

$$LTV = \text{Total Gross Revenue} - \text{Total Cost}$$

Στη συγκεκριμένη ανάλυση, όμως, δε μελετάται η έννοια του κόστους. Αυτό συνεπάγεται ότι το μέγεθος LTV είναι ίσο με τη μεταβλητή κέρδους (Revenue), η οποία έχει προσδιοριστεί προηγουμένως από τη μέθοδο RFM και περιλαμβάνεται στο τελικό σύνολο δεδομένων.

Αξίζει να σημειωθεί ότι γενικά ο υπολογισμός του δείκτη CLV είναι πρόβλημα παλινδρόμησης. Ωστόσο, για τους σκοπούς της παρούσας μελέτης, το μοντέλο που θα δημιουργηθεί θα είναι μέρος της οικογένειας των μοντέλων κατηγοριοποίησης (Classification), αφού σε αυτή την περίπτωση είναι επιθυμητό να διαχωριστεί ο CLV σε διαφορετικές κλάσεις.

5.8.1 Εφαρμογή K-means

Επομένως, θα ακολουθηθεί παρόμοια λογική όπως στην παράγραφο 5.7. Αυτό σημαίνει ότι με τη βοήθεια του αλγόριθμου K-Means θα εντοπιστούν οι υφιστάμενες ομάδες LTV.

Ο απώτερος στόχος αυτής της διαδικασίας είναι η κατηγοριοποίηση των πελατών στις ακόλουθες 3 κατηγορίες για την πιο ευέλικτη διαχείριση από λειτουργική άποψη.

- Low LTV (Χαμηλό LTV)
- Med LTV (Μεσαίο LTV)
- High LTV (Υψηλό LTV)
- Highest LTV (Μέγιστο LTV)

	count	mean	std	min	25%	50%	75%	max
LTV Cluster								
Low LTV	12189.0	21.781284	9.114979	1.67864	15.0	23.286235	25.000	40.0
Mid LTV	11437.0	66.399707	28.277886	13.58100	50.0	52.000000	85.000	180.0
High LTV	11726.0	174.183481	107.493861	42.14168	100.0	150.000000	200.000	1500.0
Highest LTV	11742.0	1407.660718	3115.751567	115.00000	388.0	651.600000	1351.875	150820.0

Εικόνα 32 – K-Means LTV

Παρατηρώντας τα αποτελέσματα από την εφαρμογή του αλγορίθμου της συσταδοποίησης, βλέπουμε ότι το cluster που αφορά το μέγιστο LTV είναι αυτό που συγκεντρώνει τους πελάτες με τη μεγαλύτερη μέση τιμή LTV.

5.8.2 Feature Engineering

Μετά την ολοκλήρωση του αλγορίθμου της συσταδοποίησης στο τελικό σύνολο δεδομένων, ακολούθησε η εφαρμογή ορισμένων τεχνικών επεξεργασίας χαρακτηριστικών (feature engineering), ώστε να προετοιμάσουμε κατάλληλα το σύνολο δεδομένων για την είσοδό του στο μοντέλο μηχανικής μάθησης.

Τεχνικές επεξεργασίας χαρακτηριστικών που εφαρμόστηκαν:

- Η κατηγορική μεταβλητή “Segment” μετατράπηκε σε αριθμητική. Πιο συγκεκριμένα, για τις κατηγορίες “Churned”, “At Risk”, “Potential” και “Loyal”, χρησιμοποιήθηκαν οι αριθμοί 1, 2, 3 και 4 αντίστοιχα.
- Πραγματοποιήθηκε ο υπολογισμός της συσχέτισης (correlation) μεταξύ των μεταβλητών που αποτελούν το τελικό σύνολο δεδομένων. Στον παρακάτω πίνακα, είναι εμφανές ότι οι μεταβλητές οι οποίες σχετίζονται με το κέρδος είναι και αυτές που έχουν τη μεγαλύτερη συσχέτιση με την εξαρτημένη μεταβλητή του LTV.

LTV Cluster	1.000000
MonetaryCluster	0.928659
FrequencyCluster	0.843806
RFM Score	0.803227
Frequency	0.402288
MonetaryValue	0.287232
LTV	0.287232
CLV	0.090075
Customer ID	0.069365
Recency	-0.049548
RecencyCluster	-0.059937

Name: LTV Cluster, dtype: float64

Εικόνα 33 – Data Correlation

- Πραγματοποιήθηκε ο διαχωρισμός του τελικού συνόλου δεδομένων σε υποσύνολα εκπαίδευσης (training set) και ελέγχου (test set).

Ολοκληρώνοντας την υλοποίηση των παραπάνω σταδίων, είναι πλέον εφικτή η εφαρμογή διαφόρων μοντέλων μηχανικής μάθησης, μέσω των οποίων θα προκύψει το βέλτιστο δυνατό.

5.8.3 Μοντέλο Μηχανικής Μάθησης

Σημαντικότερο σκοπό της παρούσας ανάλυσης αποτελεί η δημιουργία ενός μοντέλου μηχανικής μάθησης το οποίο θα παρέχει τη δυνατότητα της κατηγοριοποίησης των πελατών της επιχείρησης με βάση τη βαθμολογία του δείκτη CLV.

Μετά από έρευνα της βιβλιογραφίας και λαμβάνοντας υπόψη τη φύση της ανάλυσης, έχει επιλεγεί ο αλγόριθμος κατηγοριοποίησης XGBOOST Classifier ως βασικός αλγόριθμος κατηγοριοποίησης.

Από την εφαρμογή του αλγορίθμου, τόσο στο σύνολο εκπαίδευσης όσο και στο σύνολο ελέγχου, προέκυψε το ακόλουθο classification report.

Training Accuracy: 1.0				
Test Accuracy: 0.9985788725722407				
	precision	recall	f1-score	
0	1.00	1.00	1.00	
1	0.98	1.00	0.99	
2	1.00	0.80	0.89	
3	0.50	1.00	0.67	
accuracy			1.00	
macro avg	0.87	0.95	0.89	
weighted avg	1.00	1.00	1.00	

Εικόνα 34 – Classification Report

Με βάση το Classification Report, μπορούμε να παρατηρήσουμε τα εξής:

- **Accuracy:**
 Η ακρίβεια στο σύνολο εκπαίδευσης είναι 100%, υποδεικνύοντας ότι το μοντέλο πραγματοποιεί τις βέλτιστες προβλέψεις για τα δεδομένα εκπαίδευσης. Η ακρίβεια στο σύνολο ελέγχου είναι 99,95%, γεγονός που υποδηλώνει ότι το μοντέλο αποδίδει πολύ καλά σε δεδομένα δοκιμής. Ωστόσο, είναι σημαντικό να σημειωθεί ότι η ακρίβεια ενδέχεται να μην είναι η πιο κατατοπιστική μετρική σε μη ισορροπημένα σύνολα δεδομένων.
- **Precision, Recall και F1-Score:**
 Τα συγκεκριμένα μεγέθη υπολογίζονται με βάση τους ακόλουθους τύπους.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\text{F1 - Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Οι τιμές της Ακρίβειας (Precision), της Ανάκλησης (Recall) και της Βαθμολογίας F1 (F1-Score) είναι εξαιρετικές σε κάποιες κατηγορίες και αρκετά καλές για τις υπόλοιπες. Αυτό υποδηλώνει ότι το μοντέλο έχει πολύ καλή απόδοση τόσο όσον αφορά την ακρίβεια (την ικανότητα να αναγνωρίζει σωστά τα θετικά περιστατικά) όσο και την ανάκληση (την ικανότητα να αναγνωρίζει σωστά όλες τις θετικές περιπτώσεις).

5.8.4 Σύγκριση Μοντέλων Κατηγοριοποίησης

Στα πλαίσια της παρούσας εργασίας, κρίθηκε σκόπιμο να εφαρμοστούν και άλλοι διαφορετικοί αλγόριθμοι κατηγοριοποίησης, ώστε να επιλεγθεί το τελικό μοντέλο, που αποτελεί το βέλτιστο.

Συγκεκριμένα, εφαρμόστηκαν 5 μοντέλα μηχανικής μάθησης και πραγματοποιήθηκε η αξιολόγησή τους με βάση τα στοιχεία των μετρικών τους.

Model: SVM
Accuracy: 0.8810289389067524
Precision: 0.8870295638338469
Recall: 0.8810289389067524
RMSE: 0.3449218188129704

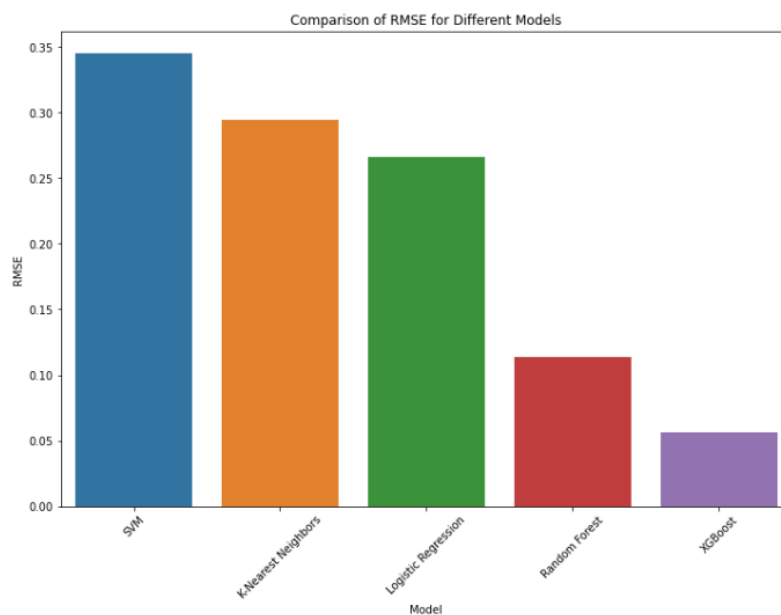
Model: K-Nearest Neighbors
Accuracy: 0.9131832797427653
Precision: 0.9139857205237699
Recall: 0.9131832797427653
RMSE: 0.29464677201224304

Model: Logistic Regression
Accuracy: 0.9485530546623794
Precision: 0.9594141137979678
Recall: 0.9485530546623794
RMSE: 0.2659690768477198

Model: Random Forest
Accuracy: 0.9871382636655949
Precision: 0.9877098963915685
Recall: 0.9871382636655949
RMSE: 0.11340959542474854

Model: XGBoost
Accuracy: 0.9967845659163987
Precision: 0.9968215249288539
Recall: 0.9967845659163987
RMSE: 0.05670479771237427

Εικόνα 35 – Σύγκριση Αλγορίθμων



Εικόνα 36 – Σύγκριση RMSE αλγορίθμων

Κεφάλαιο 6: Συμπεράσματα και Μελλοντικές Επεκτάσεις

6.1 Συμπεράσματα

Στα πλαίσια της παρούσας διπλωματικής εργασίας, προτείνεται μία μέθοδος για τον υπολογισμό της αξίας ζωής του πελάτη από την οπτική γωνία της μηχανικής μάθησης. Αυτή η συγκεκριμένη μέτρηση είναι μια σημαντική μέτρηση για όλους μεγάλους οργανισμούς και τις επιχειρήσεις.

Ωστόσο, η εύρεση και η διαχείριση δεδομένων που αφορούν δραστηριότητες εταιρειών iGaming είναι ένα από τα βασικά ζητήματα που ανακύπτουν κατά την πειραματική φάση. Είναι αντικειμενικά δύσκολο να παραμετροποιηθεί ένα μοντέλο που θα μπορούσε να χρησιμοποιηθεί στην παραγωγή υπό ρεαλιστικές συνθήκες όταν οι μεταβλητές είναι κωδικοποιημένες και τα σύνολα δεδομένων είναι ελλιπή.

Η προσέγγιση RFM φαίνεται να είναι αρκετά προσαρμόσιμη όσον αφορά το εύρος των εφαρμογών της. Παρόλο που η συγκεκριμένη μέθοδος χτίστηκε πάνω στη λογική του λιανικού εμπορίου και στα αντίστοιχα προφίλ των καταναλωτών, στην παρούσα εργασία αποδεικνύεται σε μεγάλο βαθμό προσαρμόσιμη και σε άλλου είδους δεδομένα.

Επιπλέον, ο αλγόριθμος XGBoost Classifier αποδεικνύεται αποτελεσματικός ακόμα κι αν το αντικείμενο της ανάλυσης είναι η κατηγοριοποίηση σε πολλαπλές κλάσεις. Ωστόσο, λαμβάνοντας υπόψη την πειραματική εφαρμογή αυτής της εργασίας στον συγκεκριμένο αλγόριθμο αλλά και στους υπόλοιπους αλγορίθμους κατηγοριοποίησης, διαπιστώθηκε ότι το πεδίο βελτιστοποίησης της συγκεκριμένης μεθόδου είναι περιορισμένο. Μια ιδιαιτερότητα είναι ότι, στο μεγαλύτερο μέρος της βιβλιογραφίας, η πιο δημοφιλής μέθοδος υπολογισμού του CLV ακολουθεί τη μέθοδο της παλινδρόμησης.

6.2 Μελλοντικές Επεκτάσεις

Ο συνδυασμός της υψηλής απόδοσης και της αλγοριθμικής δομής του XGBoost του επιτρέπει να λειτουργεί αποτελεσματικά σε υπολογιστικά περιβάλλοντα παραλληλισμένα (Parallelization) και καταναμημένα (Distributed), προσφέροντας προοπτικές για μελλοντική έρευνα για την επέκταση των τρεχουσών μεθόδων σε αρχιτεκτονικές επιπέδου συστήματος και μεγάλων δεδομένων.

Υπό αυτό το πλαίσιο, η πειραματική έρευνα και η ανάπτυξη των τρεχουσών υλοποιήσεων σε βάσεις δεδομένων NoSQL θα ήταν αρκετά ενδιαφέρουσα. Μια τέτοια παραμετροποίηση μπορεί να δημιουργήσει τις κατάλληλες συνθήκες για τη δημιουργία ενός μοντέλου, ικανού να σταθεί σε ένα εργασιακό περιβάλλον που χρησιμοποιεί πρωτίστως αρχιτεκτονικά μεγάλα δεδομένα.

Μία επιπλέον μελλοντική επέκταση αυτής της εργασίας θα μπορούσε να είναι η δημιουργία ενός περιβάλλοντος front-end που, χρησιμοποιώντας πλαίσια εφαρμογών όπως το Flask για την Python, θα μπορεί να χρησιμοποιεί τα παραπάνω μοντέλα μηχανικής μάθησης για τον

υπολογισμό του CLV μέσω μίας εφαρμογής τελικού χρήστη. Μία τέτοια εφαρμογή θα μπορούσε να αποτελέσει πολύ χρήσιμο εργαλείο σε ένα εργασιακό περιβάλλον.

Ένα μοντέλο το οποίο κατηγοριοποιεί τους πελάτες με βάση την αξία που προσθέτουν στην επιχείρηση θα μπορούσε να εξελιχθεί σε ένα νέο συνδυασμένο μοντέλο που θα «λειτουργεί παράλληλα» με το υπάρχον μοντέλο και θα μπορεί να παρέχει συστάσεις πωλήσεων και παροχής υπηρεσιών. Αυτό επιτρέπει εξατομικευμένες προτάσεις για προϊόντα και υπηρεσίες ανά κατηγορία πελατών.

Τέλος, πρέπει να σημειωθεί ότι στην παρούσα διπλωματική εργασία, το πρόβλημα του υπολογισμού του δείκτη Customer Lifetime Value υπολογίζεται με τη λογική του προβλήματος κατηγοριοποίησης. Η μηχανική μάθηση μέσω ενός μεγάλου αριθμού μοντέλων και στατιστικών μεθόδων μπορεί να προσφέρει μία εναλλακτική λύση, όπως η χρήση νευρωνικών δικτύων για τη δημιουργία μοντέλων. Μια τέτοια λογική θα πήγαινε προς την κατεύθυνση της βελτιστοποίησης της αξιοπιστίας και της απόδοσης των αποτελεσμάτων.

Βιβλιογραφία - Αναφορές

- [1] Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners, J. Dean, 2014
- [2] Business Intelligence and Analytics: From Big Data to Big Impact, Hsinchun Chen, R. Chiang, V. Storey, 2012
- [3] Exploring the role of customer relationship management (CRM) systems in customer knowledge creation, F. Khodakarami, Yolande E. Chan, 2014
- [4] A Data Mining System for Enhancing Profit Growth based on RFM and CLV, Lahcen Abidar, Ikram El Asri, Dounia Zaidouni, Abdeslam Ennouaary, 2022
- [5] Investigating Two Customer Lifetime Value Models from Segmentation Perspective, Abdulkadir Hiziroglu, Serkan Sengul, 2012
- [6] Customer segmentation and strategy development based on customer lifetime value: A case study, Su-yeon Kim, Taesoo Jung, E. Suh, Hyunseok Hwang, 2006
- [7] Customer Lifetime Value Analysis to Profitability Management: A framework to strategic marketing, M. Safari, 2015
- [8] Customer Lifetime Value (CLV) Measurement based on RFM model, S. Babak, Amir Khanlari, 2007
- [9] Marketing Intelligence & Planning, Ngai, E.W.T., 2005
- [10] Customer Management as the Origin of Collaborative Customer Relationship Management, Alexander H. Kracklauer, D. Quinn Mills, Dirk Seifert, 2004
- [11] Customer lifetime value models for decision support, Saharon Rosset, Einat Neumann, Uri Eick, Nurit Vatnik, 2003