ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

ΣΧΟΛΗ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ

Πρόβλεψη αποχώρησης πελατών με χρήση αλγορίθμων Μηχανικής Μάθησης

Customer churn prediction using Machine Learning algorithms

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Χαϊντούτης Βάιος

Επιβλέπων καθηγητής: Τσαφαράκης Στέλιος

Μέλη εξεταστικής επιτροπής: Δούμπος Μιχαήλ, Κρασαδάκη Ευαγγελία

Χανιά, 2022

Πίνακας Περιεχομένων

Πρόλογος

Με αφορμή την ολοκλήρωση των προπτυχιακών μου σπουδών, θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή της διπλωματικής μου εργασίας κ. Τσαφαράκη Στέλιο, για την εμπιστοσύνη και τη βοήθεια που μου παρείχε. Ακόμα, θα ήθελα να εκφράσω την ευγνωμοσύνη και την εκτίμησή μου στον διδακτορικό φοιτητή Κυριακίδη Αναστάσιο, για τη συνεισφορά του σε όλα τα στάδια εκπόνησης της εργασίας. Τέλος, οφείλω ένα μεγάλο ευχαριστώ στους φίλους μου, στην οικογένειά μου και ειδικά στη μητέρα μου, για τη στήριξή τους όλα αυτά τα χρόνια. Η εργασία αυτή είναι αφιερωμένη στη μνήμη του φίλου και συμφοιτητή Βόκα Δημήτρη.

Περίληψη

Οι αλγόριθμοι Μηχανικής Μάθησης (Machine Learning) εφαρμόζονται όλο και περισσότερο σε διάφορους τομείς, όπως σε αυτούς της βιομηχανίας, της ιατρικής και της αστρονομίας για την εξόρυξη δεδομένων (data mining) από σύνθετες βάσεις (datasets). Στην παρούσα εργασία θα εφαρμοστούν μερικοί από αυτούς τους αλγορίθμους σε μία βάση δεδομένων των πελατών μιας εταιρείας.

Το dataset που θα χρησιμοποιηθεί, προέρχεται από τη βιβλιοθήκη Kaggle, περιλαμβάνει 64.000 καταχωρήσεις πελατών. Οι μεταβλητές του dataset αφορούν τις αγοραστικές συνήθειες κάθε πελάτη, τις ενδεχόμενες παροχές και προσφορές που έχουν λάβει από την εταιρεία αλλά και το αν συνεχίζουν να ανήκουν στο πελατολόγιο της ή όχι (conversion rate).

Το πρόγραμμα που θα χρησιμοποιηθεί για τις αναλύσεις είναι το Weka (έκδοση 3.8.5) που αναπτύχθηκε στο πανεπιστήμιο του Waikato της Νέας Ζηλανδίας. Είναι ένα ελεύθερο λογισμικό που περιλαμβάνει διάφορους αλγορίθμους Μηχανικής Μάθησης που εφαρμόζονται για την ανάλυση και εξόρυξη δεδομένων.

Αρχικά, θα εκπαιδευτεί ένας ταξινομητής (classifier) για την πρόβλεψη της συμπεριφοράς πελατών της εταιρείας ως προς την αφοσίωση και την εμπιστοσύνη που θα δείξουν στην εταιρεία για τις μελλοντικές τους συναλλαγές, ο οποίος θα επιλεχθεί μέσα από μια συγκριτική αξιολόγηση (benchmarking) αλγορίθμων επιτηρούμενης μάθησης ως προς την ακρίβεια και ποιότητα των αποτελεσμάτων, με σκοπό να βρεθεί ο αλγόριθμος με τον καλύτερο συνδυασμό παραμέτρων.

Στη συνέχεια, επιχειρείται ένας διαχωρισμός των πελατών σε δύο κατηγορίες, αυτούς που συνεχίζουν τις συναλλαγές τους με την εταιρεία και αυτούς που αποχώρησαν (churners). Σε κάθε μία από τις ομάδες θα εφαρμοστούν αλγόριθμοι ομαδοποίησης και εξαγωγής κανόνων συσχέτισης για να εντοπιστούν μοτίβα και κοινά χαρακτηριστικά μεταξύ των πελατών που να εξηγούν και να δικαιολογούν τη φυγή ή όχι από την εταιρεία.

Στην ομαδοποίηση (clustering), θα γίνει ο διαχωρισμός του δείγματος σε ομάδες με χρήση του αλγορίθμου k-means, όπου τα στοιχεία της μιας ομάδας είναι όσο το δυνατόν πιο όμοια μεταξύ τους και όσο το συνατόν πιο διαφορετικά από τις υπόλοιπες ομάδες για να υπάρξει μια ξεκάθαρη κατηγοριοποίηση.

Η εξαγωγή κανόνων συσχέτισης (association rules) είναι μία μέθοδος για την ανακάλυψη ενδιαφέρων σχέσεων μεταξύ μεταβλητών σε μεγάλες βάσεις

δεδομένων. Με αυτό τον τρόπο εντοπίζονται κοινές συνήθειες των πελατών που ανάλογα με τη συχνότητα που εμφανίζονται οδηγούν σε χρήσιμα συμπεράσματα.

Τελικός στόχος είναι να βρεθεί το προφίλ των πελατών που είναι πιο πιθανό να αποχωρήσουν από την εταιρεία, να αναλυθεί και να εφαρμοστούν μέθοδοι εστιασμένου μάρκετινγκ προκειμένου να καταφέρει η εταιρεία να κρατήσει τους πελάτες αυτούς κοντά της, αλλά και ενδεχομένως να ανταμείψει το προφίλ πελατών που επιδεικνύουν συνέπεια και αφοσίωση στην εταιρεία.

Κεφάλαιο 1: Εισαγωγή

1.1 Στόχοι και δομή της διπλωματικής εργασίας

1.1.1 Στόχοι της έρευνας και της διπλωματικής εργασίας

Η συνεχής ανάπτυξη και ο εκσυγχρονισμός της βιομηχανίας, σε συνδυασμό με τη ραγδαία πρόοδο της τεχνολογίας των υπολογιστών, έχει δημιουργήσει την ανάγκη αλλά και το υπόβαθρο για την εφαρμογή καλά μελετημένων, στοχευμένων και αποτελεσματικών μεθόδων Μάρκετινγκ σε διάφορους τομείς, που προκύπτουν από τη στατιστική ανάλυση όλο και μεγαλύτερων βάσεων δεδομένων σε σύντομο χρόνο με μεγάλη ακρίβεια.

Η επανάσταση των δεδομένων συνεχίζεται με αμείωτο ρυθμό από το 2010 και μετά. Οι βάσεις δεδομένων συνεχώς μεγαλώνουν, οι μεταβλητές αυξάνονται, γίνονται πιο σύνθετες και οι στατιστικές αναλύσεις πιο περίπλοκες και χρονοβόρες.

Η Μηχανική Μάθηση εφαρμόζεται για την κατανόηση και απλοποίηση των δεδομένων, ώστε να εντοπιστούν μοτίβα και πατέντες στα δεδομένα που θα δώσουν λύσεις στα ερωτήματα που ανακύπτουν και θα οδηγήσουν τις επιχειρήσεις και τους αναλυτές τους στη λήψη πιο σωστών αποφάσεων σε θέματα παραγωγής και ικανοποίησης πελατών, άρα και σε μεγιστοποίηση του κέρδους.

1.1.2 Δομή της διπλωματικής εργασίας

Αρχικά θα γίνει μια αναφορά στις σύγχρονες μεθόδους Μάρκετινγκ που χρησιμοποιούνται από τις επιχειρήσεις και θα αναλυθούν οι έννοιες της ικανοποίησης και της αφοσίωσης πελατών αλλά και η σημασία της πρόβλεψης αποχώρησης πελατών. Θα συζητηθούν επίσης αποτελέσματα προγενέστερων ερευνών ως προς την πρόβλεψη αποχώρησης πελατών.

Στη συνέχεια, θα γίνει μια εισαγωγή στον τομέα της Μηχανικής Μάθησης και της αναγνώρισης προτύπων. Θα παρουσιαστούν τα είδη της Μηχανικής Μάθησης και κάποιες βασικές έννοιες και διαδικασίες, καθώς και ο τρόπος υλοποίησής τους. Σημαντική αναφορά γίνεται στη διαδικασία προ επεξεργασίας των δεδομένων έτσι ώστε να γίνει επιλογή μόνο των δεδομένων που είναι απαραίτητα για την ανάλυση.

Έπειτα θα παρουσιαστούν οι αλγόριθμοι που θα χρησιμοποιηθούν για τη δημιουργία βασικών μοντέλων ταξινόμησης και ομαδοποίησης και θα αναλυθούν τα μέτρα απόδοσης και προσαρμογής τους.

Στη συνέχεια θα αναλυθεί το πειραματικό μέρος της έρευνας, ενώ τέλος θα παρουσιαστούν συγκριτικά τα αποτελέσματα για την εφαρμογή του κάθε μοντέλου

ταξινόμησης για την πρόβλεψη αποχώρησης των πελατών, τα αποτελέσματα της ομαδοποίησης και της αναγνώρισης μοτίβων αλλά και κοινών καταναλωτικών συνηθειών των πελατών, καθώς και τα συμπεράσματα που θα προκύψουν από την έρευνα.

Κεφάλαιο 2: Στοχευμένο Μάρκετινγκ - Customer Satisfaction – Customer Churn

2.1 Στοχευμένο Μάρκετινγκ

Πιθανότατα τη μεγαλύτερη εξέλιξη στην ιστορία του Μάρκετινγκ να αποτελεί η συνειδητοποίηση εκ μέρους των επιχειρήσεων ότι οι ευρύτερες αγορές αποτελούνται από σημαντικά διαφορετικές μεταξύ τους ομάδες καταναλωτών και υποομάδες. Κατά συνέπεια, μια επιχείρηση θα πετύχει βέλτιστα αποτελέσματα εφόσον μελετήσει τη συνολική αγορά, εντοπίσει τις σχετικά ομοιογενείς υποομάδες της συνολικής αγοράς και στη συνέχεια αποφασίσει να διαμορφώσει το προϊόν της έτσι ώστε να ικανοποιεί στον μέγιστο δυνατό βαθμό τις ανάγκες μιας συγκεκριμένης υποομάδας. Η πρακτική αυτή, δηλαδή η εστίαση των προσπαθειών της επιχείρησης προς την εξυπηρέτηση μίας ή περισσότερων ομάδων ανθρώπων που έχουν παρόμοιες ανάγκες και παρόμοια χαρακτηριστικά είναι γνωστή ως στοχευμένο μάρκετινγκ ή ως μάρκετινγκ στόχου (Kotler & Armstrong, 2013). [1]

Αυτό δε θα μπορούσε να πραγματοποιηθεί χωρίς την χρήση στατιστικών μεθόδων και αλγορίθμων Μηχανικής Μάθησης. Τα τελευταία χρόνια, οι εταιρείες είναι σε θέση να γνωρίζουν πολλά περισσότερα πράγματα σε ότι έχει να κάνει με τους πελάτες τους και την αποδοχή που έχουν τα προϊόντα και οι υπηρεσίες που προσφέρουν σε αυτούς, μέσα από την εκ βάθους ανάλυση των βάσεων δεδομένων που δημιουργούν. Μελετούν τα στοιχεία που έχουν στη διάθεση τους και υιοθετούν την προσέγγιση του στοχευμένου μάρκετινγκ. Έτσι, εντοπίζουν νέες ευκαιρίες που εμφανίζονται στο εξωτερικό τους περιβάλλον, αναπτύσσουν το καταλληλότερο προϊόν για κάθε αγορά που στοχεύουν και εφαρμόζουν πιο αποτελεσματική διαφήμιση και επικοινωνία μάρκετινγκ.

2.1.1 Διαδικασία στοχευμένου μάρκετινγκ

Η πρακτική εφαρμογή της φιλοσοφίας του στοχευμένου μάρκετινγκ βασίζεται σε μια αυστηρά προσδιορισμένη διαδικασία, η οποία περιλαμβάνει τρία διαδοχικά και αλληλοεξαρτώμενα στάδια και είναι γνωστή ως STP, δηλαδή:

- Τμηματοποίηση της αγοράς (Segmentation)
- Στόχευση στην αγορά (Targeting)
- Χωροθέτηση, τοποθέτηση προϊόντος στην αγορά (Positioning)

Η διαδικασία STP είναι μια σημαντική έννοια στη μελέτη και την εφαρμογή του μάρκετινγκ. Το μάρκετινγκ μιάς επιχείρησης χτίζεται γύρω από το STP, καθώς

καταδεικνύει τους δεσμούς μεταξύ μιας συνολικής αγοράς και πώς μια επιχείρηση επιλέγει να ανταγωνιστεί στην αγορά αυτή. Συνήθως, η τμηματοποίηση διεξάγεται πρώτα, μετά η επιλογή ενός ή περισσότερων αγορών-στόχων και τελικά η χωροθέτηση.

2.1.2 Τμηματοποίηση

Στο στάδιο της τμηματοποίησης, το σύνολο της αγοράς διαχωρίζεται σε διακριτές ομάδες καταναλωτών, οι οποίες προσβλέπουν σε διαφορετικές ωφέλειες προϊόντος και αντιδρούν με διαφορετικό τρόπο η καθεμιά στο προσφερόμενο μείγμα μάρκετινγκ. Η καθεμία από τις ομάδες αυτές ονομάζεται τμήμα της αγοράς και θα πρέπει να έχει διακριτά χαρακτηριστικά από τις υπόλοιπες και εσωτερική ομοιογένεια για κάποια χαρακτηριστικά τα οποία προκύπτουν από τις μεταβλητές που χρησιμοποιήθηκαν για να τμηματοποιηθεί η αγορά. Ο διαχωρισμός των πελατών πρέπει να γίνει με ιδιαίτερη προσοχή, επομένως οι μεταβλητές που θα χρησιμοποιηθούν για την τμηματοποίηση παίζουν σημαντικό ρόλο, διότι βάση αυτών θα διαμορφωθούν οι ιδιαιτερότητες του κάθε τμήματος της αγοράς. Αυτό θα οδηγήσει σε καλύτερη αντιστοίχιση των αναγκών των πελατών, ενισχυμένα κέρδη για την επιχείρηση, καλύτερες ευκαιρίες για ανάπτυξη και διατήρηση περισσότερων πελατών.

Υπάρχουν διάφοροι τρόποι που μπορεί να τμηματοποιηθεί μια καταναλωτική αγορά:

- Με βάση Δημογραφικά στοιχεία, όπως ηλικιακή ομάδα, μόρφωση, φύλο, εισόδημα
- Με βάση Ψυχογραφικά στοιχεία, που αφορούν τον τρόπο ζωής των πελατών, την προσωπικότητα, την κοινωνική τάξη και τις αξίες του ατόμου
- Με βάση γεωγραφικά χαρακτηριστικά, μια προσέγγιση διαίρεσης της αγοράς βάση γεωγραφικών μονάδων, όπως κράτη, περιφέρειες, νομοί, πόλεις, γειτονιές.
- Με βάση τη συμπεριφορά των πελατών, δηλαδή τον τρόπο χρήσης του προϊόντος, την εμπιστοσύνη στη μάρκα ή στην επιχείρηση, την ευαισθησία τους στην τιμή, την προηγούμενη αγοραστική συμπεριφορά τους ή τους τρόπους αγοράς και χρήσης του προϊόντος

Οι δημογραφικές μεταβλητές αποτελούν τις συνηθέστερες βάσεις τμηματοποίησης της αγοράς, γιατί οι ανάγκες και οι επιθυμίες των καταναλωτών καθώς και ο τρόπος που χρησιμοποιούν ένα προϊόν συνήθως επηρεάζονται από τα δημογραφικά τους χαρακτηριστικά. Επίσης, οι δημογραφικές μεταβλητές είναι εύκολα και αξιόπιστα μετρήσιμες.

Εν κατακλείδι, κατά την τμηματοποίηση αναλύονται τα χαρακτηριστικά του κάθε τμήματος στην αγορά και αναπτύσσεται το αντιπροσωπευτικό προφίλ του μέσου ατόμου κάθε τμήματος. Πραγματοποιείται ανάλυση συστάδων από τη οποία προκύπτει ένας αριθμός από διακριτές και διαφορετικές μεταξύ τους ομάδες καταναλωτών. Η κάθε ομάδα αποτελεί ξεχωριστό τμήμα της αγοράς και παρουσιάζει υψηλή εσωτερική ομοιογένεια.

Για να είναι η τμηματοποίηση αποτελεσματική θα πρέπει τα επιμέρους τμήματα που εντοπίστηκαν να πληρούν κάποια κριτήρια. Θα πρέπει λοιπόν να είναι

1. Μετρήσιμα: το συνολικό μέγεθος του κάθε τμήματος να μετράται σε νούμερα όπως και η αγοραστική του δύναμη, το προφίλ του, τα δημογραφικά του χαρακτηριστικά
2. Ουσιαστικά: να έχουν προκύψει επαρκώς μεγάλα τμήματα, τα οποία να έχουν οικονομική σημασία και επαρκή ομοιογένεια
3. Προσβάσιμα: να υπάρχει η πρακτική δυνατότητα εκ μέρους της επιχείρησης να προσεγγίσει τα άτομα των τμημάτων της αγοράς που εντοπίστηκαν
4. Ενεργήσιμα: να είναι εφικτό να προσεγγιστούν και να εξυπηρετηθούν τα άτομα των τμημάτων της αγοράς που εντοπίστηκαν

2.1.3 Στόχευση της αγοράς

Η Στόχευση της αγοράς γίνεται όταν έχουν καθοριστεί τα διακριτά τμήματα της αγοράς. Είναι το δεύτερο στάδιο της STP. Πραγματοποιείται αξιολόγηση των τμημάτων που έχουν προκύψει και επιλέγονται τα πιο ελκυστικά για να τοποθετηθεί το προϊόν και να εφαρμοστεί το μείγμα μάρκετινγκ (προϊόν, τιμή, τόπος, προβολή, άνθρωποι, διαδικασίες, φυσική μαρτυρία).

Υπάρχουν πολλοί παράγοντες που χρειάζεται να εξεταστούν προκειμένου να επιλεγεί το κατάλληλο τμήμα ή τμήματα. Αρχικά το μέγεθος του τμήματος θα πρέπει να είναι αρκετά μεγάλο για να έχει νόημα η στόχευση σε αυτό και να έχει αρκετές προοπτικής μεγέθυνσης και ανάπτυξης στο μέλλον. Θα πρέπει ακόμα να είναι ελκυστικό από πλευράς κερδοφορίας, αλλά και να κριθεί αν η επένδυση στο συγκεκριμένο τμήμα συβαδίζει με τους στόχους και την πολιτική της επιχείρησης, καθώς και αν υπάρχουν οι πόροι, το ανθρώπινο δυναμικό και η τεχνογνωσία ώστε να αναπτυχθεί το κατάλληλο προϊόν για να καλυφθεί επαρκώς το επιλεγμένο τμήμα.

2.1.4 Χωροθέτηση

Μετά την τμηματοποίηση και την στόχευση ακολουθεί η χωροθέτηση, που αποτελεί το τελευταίο στάδιο της STP. Στο στάδιο αυτό έχουμε την πραγματοποίηση της προσφοράς, δηλαδή του προϊόντος, της εικόνας του και της

αξίας του, έτσι ώστε τα άτομα του στοχευμένου τμήματος να αντιλαμβάνονται, να κατανοούν και να εκτιμούν τη σημασία και την αξία του προϊόντος σε σχέση με τα ανταγωνιστικά προϊόντα. Η χωροθέτηση είναι η επικοινωνία της συνολικής πρότασης αξίας της επιχείρησης, αυτή που δημιουργεί και διατηρεί τους πελάτες.

Η χωροθέτηση στην αγορά περιλαμβάνει τρία βήματα:

1. Εντοπισμός πιθανών ανταγωνιστικών πλεονεκτημάτων του προϊόντος τα οποία αξίζει να αξιοποιηθούν
2. Επιλογή των καταλληλότερων ανταγωνιστικών πλεονεκτημάτων
3. Προβολή των ανταγωνιστικών πλεονεκτημάτων στα άτομα του στοχευμένου τμήματος

Η διαδικασία του STP (Τμηματοποίηση-Στόχευση-Χωροθέτηση) λοιπόν επιτρέπει σε μια επιχείρηση να διαμορφώσει μια στρατηγική μάρκετινγκ για τη δημιουργία και τοποθέτηση του κατάλληλου προϊόντος στην αγορά που θα αποφέρει το μεγαλύτερο κέρδος, στοχεύοντας αποτελεσματικά και με κατάλληλες μεθόδους σε συγκεκριμένη μερίδα της αγοράς. [2]

2.2 Διαχείριση πελατειακών σχέσεων (CRM)

Η Διαχείριση πελατειακών σχέσεων (Customer Relationship Management) είναι μια διαδικασία ή μεθοδολογία που χρησιμοποιείται από τις εταιρείες για να μάθουν περισσότερα για την καταναλωτική συμπεριφορά και τις ανάγκες των πελατών τους , έτσι ώστε να αναπτυχθεί ισχυρή σχέση εμπιστοσύνης μεταξύ πελατών και επιχείρησης. Η μέθοδος CRM συνδυάζει τα δεδομένα παλαιότερων και νέων πελατών για να δημιουργήσει ένα πλαίσιο ανταπόκρισης στις ανάγκες τους, ενώ βασίζεται σε έναν συνδυασμό της επιχειρηματικής διαδικασίας και της Τεχνολογίας Πληροφοριών (Information Technology), για να απαντήσει σε ερωτήματα όπως, ποιοί είναι οι πελάτες μου, γιατί είναι πελάτες μου. Για το λόγο αυτό, η αποδοτική διαχείριση της πληροφορίας και της γνώσης είναι υψίστης σημασίας για την επιτυχία της μεθόδου CRM.

Κάποια από τα κυριότερα οφέλη μιας αποδοτικής εφαρμογής της μεθόδου διαχείρισης πελατειακών σχέσεων για τις επιχειρήσεις είναι:

- οι βελτιωμένες και πιο φιλικές προς τον πελάτη ιστοσελίδες των επιχειρήσεων, καθώς η εταιρεία αναγνωρίζει καλύτερα τις ανάγκες των πελατών της και προσαρμόζει αντίστοιχα τις υπηρεσίες της
- σχηματίζεται μια ενοποιημένη και πλήρης εικόνα του πελάτη από την επιχείρηση
- υπολογίζεται ο κύκλος ζωής του πελάτη (customer lifetime value)
- σχεδιάζονται και αναπτύσσονται εξατομικευμένες συναλλαγές
- δυνατότητα πολύπλευρης επικοινωνίας με τον πελάτη

- ευκαιρία για διασταυρούμενες πωλήσεις προς τους πελάτες (cross selling)

Η CRM απαντάται σε τρία επίπεδα.

Στρατηγικό CRM: επικεντρώνεται στην ανάπτυξη μιας πελατοκεντρικής κουλτούρας. Το προϊόν, η παραγωγή και η πώληση είναι οι τρείς βασικοί προσανατολισμοί. [3]

Αναλυτικό CRM: στηρίζεται στην πληροφορία του πελάτη. Τα δεδομένα πελατών μπορεί να βρεθούν στα αποθετήρια των επιχειρήσεων, σαν δεδομένα πωλήσεων, οικονομικά δεδομένα, δεδομένα μάρκετινγκ, υπηρεσιακά δεδομένα. Με την εφαρμογή της Εξόρυξης δεδομένων (Data Mining), η επιχείρηση μπορεί να αξιολογήσει τα δεδομένα και να οδηγηθεί σε σημαντικά συμπεράσματα, όπως ποιοί αποτελούν τους πιο πιστούς πελάτες, ποιοί έχουν μεγαλύτερη πιθανότητα να αποχωρήσουν για κάποιον ανταγωνιστή, ή ποιοί πελάτες έχουν τη μεγαλύτερη αξία.

Συνεργατικό CRM: τα μέλη του προσωπικού διαφορετικών τμημάτων ανταλλάσουν πληροφορίες που συλλέγονται κατά την αλληλεπίδραση με τον πελάτη.

Το πρόβλημα πρόβλεψης αποχώρησης πελατών (churn prediction problem) είναι εφαρμογή του αναλυτικού CRM, και μπορεί να προσφέρει πολύ σημαντικές πληροφορίες σε ότι αφορά τη χάραξη αποτελεσματικής στρατηγικής για την απαλλαγή από τους ασύμφορους πελάτες και την επικερδή διατήρηση των πελατών με τη μεγαλύτερη αξία. Ωστόσο, θα πρέπει ανά πάσα στιγμή να εξασφαλίζεται ότι η επιχείρηση λειτουργεί σεβόμενη τον νέο κανονισμό προστασίας δεδομένων (General Data Protection Regulation, GDPR) της Ευρωπαϊκής Ένωσης. [4]

2.3 Customer Satisfaction

Η ικανοποίηση του πελάτη (customer satisfaction), είναι βασική προτεραιότητα των επιχειρήσεων και των υπηρεσιών γιατί είναι συνηφασμένη με το κέρδος και την εν γένη καλή και εύρυθμη λειτουργία της εκάστοτε επιχείρησης ή υπηρεσίας. Επομένως, μια εταιρεία προσπαθεί να προσφέρει ποιοτικές υπηρεσίες, να προτυποποιεί τις υπηρεσίες της και να εξυπηρετεί τους πελάτες της με κάθε τρόπο. Οι Giese & Cote (2000), αναφέρουν ότι η ικανοποίηση των καταναλωτών βασίζεται σε τρία χαρακτηριστικά:

- Η ικανοποίηση των καταναλωτών είναι μία συναισθηματική ή γνωσιακή αντίδραση
- Η αντίδραση αυτή βασίζεται στις προσδοκίες, στις επιθυμίες και στις εμπειρίες του καταναλωτή
- Η αντίδραση αυτή συμβαίνει μια συγκεκριμένη χρονική στιγμή, όπως για παράδειγμα μετά την αγορά του προϊόντος

Μια εταιρεία παροχής υπηρεσιών για να μπορέσει να κρατήσει το πελατολόγιο της και να κερδίσει νέους, πρέπει να καταφέρει να τους ικανοποιήσει. Για να το πετύχει αυτό πρέπει ο πελάτης να βρίσκεται στο επίκεντρο και όλα να καθορίζονται με βάση αυτόν. Στόχος λοιπόν αρχικά, είναι η δημιουργία υπηρεσιών που διαθέτουν κάποια βασικά χαρακτηριστικά που είναι δεδομένο ότι υπάρχουν. Ακόμα, είναι σημαντικό να υπάρχουν επιπλέον χαρακτηριστικά που ξεπερνούν τις προσδοκίες του καταναλωτή και προκαλούν την ικανοποίησή του, όπως η ύπαρξη ενός πακέτου προσφοράς. Τέλος, μπορούν να υπάρχουν χαρακτηριστικά που προσφέρουν απόλαυση στον καταναλωτή και κάνουν ιδιαίτερη την εμπερία του με την εταιρεία, όπως για παράδειγμα μια δωροκάρτα ή ένα εκπτωτικό κουπόνι για επόμενες αγορές. [5]

2.4 Customer Loyalty

Η αφοσίωση του πελάτη αναφέρεται στην πιθανότητα ένας πελάτης να πραγματοποιήσει επανειλλημένες επιχειρηματικές δραστηριότητες ή συναλλαγές με μια εταιρεία ή επωνυμία. Είναι το αποτέλεσμα της ικανοποίησης πελατών, της θετικής εντύπωσης που έχει δημιουργηθεί στον πελάτη αλλά και της συνολικής αξίας των αγαθών ή των υπηρεσιών που απολαμβάνει ο πελάτης από την επιχείρηση.

Οι εταιρείες επενδύουν πολλά στο χτίσιμο ισχυρής βάσης πελατών καθώς αποτελεί βασικό συστατικό για την ανάπτυξη και εδραίωση τους. Κάποιοι από τους κύριους λόγους που καθιστούν την αφοσίωση πελατών ιδιαίτερα σημαντική αναφέρονται παρακάτω:

- Οι σταθεροί πελάτες επενδύουν περισσότερα χρήματα από κάποιον που πραγματοποιεί τις πρώτες του αγορές στην εταιρεία
- Οι πιστοί πελάτες (loyal customers) παράγουν υψηλότερο ποσοστό μετατροπής (conversion rate). Το ποσοστό μετατροπής είναι το ποσοστό των πελατών που εκτελούν την επιθυμητή ενέργεια, όπως την πραγματοποίηση μιας αγοράς σε ένα κατάστημα.
- Υπάρχει μεγιστοποίηση κέρδους
- Η διατήρηση ενός ήδη υπάρχοντος πελάτη είναι γενικά φθηνότερη από την απόκτηση ενός νέου
- Η αφοσίωση πελατών βοηθά στον αποδοτικό σχεδιασμό της επιχείρησης
- Πραγματοποιούνται συχνότερα αγορές από πιστούς πελάτες
- Οι πιστοί πελάτες ξοδεύουν περισσότερα χρήματα σε επετείους και περιόδους εκπτώσεων

Στόχος των επιχειρήσεων είναι να αποκτήσουν και να διατηρήσουν τους πιστούς πελάτες, αλλά και να απαλλαγούν από εκείνους που είναι οικονομικά

ασύμφοροι. Η δημιουργία βάσης αφοσιωμένων πελατών απαιτεί συντονισμένη και συνεχή προσπάθεια από την επιχείρηση.

Σημαντική είναι η κατάτμηση των πελατών (customer segmentation). Μέσα από διάφορες στατιστικές μεθόδους και με την εφαρμογή αλγορίθμων μηχανικής μάθησης προκύπτει ομαδοποίηση πελατών με βάση κάποια κοινά χαρακτηριστικά ή κάποια μοτίβα συμπεριφοράς. Ακόμα, η επιχείρηση μπορεί να πάρει πληροφορίες και χρήσιμες συμβουλές απευθείας από τους πελάτες μέσα από την προσωπική επικοινωνία (customer feedback). Μέσα από τέτοιες διαδικασίες, οι επιχειρήσεις προχωρούν σε αναλύσεις των διαφορετικών ομάδων πελατών και μετρούν το βαθμό αφοσίωσης τους, ενώ μπορούν να εφαρμόσουν πιο αποτελεσματικά μεθόδους μάρκετινγκ στους πελάτες που τους ενδιαφέρει να κρατήσουν. [6]

2.5 Customer Churn

Στα πλαίσια της διαχείρισης πελατειακών σχέσεων (Customer Relationship Management), γίνεται προσπάθεια αποτροπής του φαινομένου αποχώρησης των πελατών (Customer churn prevention). Η αφοσίωση του πελάτη μετριέται συναρτήσει της καταναλωτικής συμπεριφοράς του. Το ποσοστό αποχώρησης πελατών (churn rate), είναι ένας δείκτης που αφορά πελάτες που ακύρωσαν μια συνδρομή ή έπαψαν να αγοράζουν από μια συγκεκριμένη επιχείρηση. Είναι το ποσοστό των πελατών που αποχώρησαν ως προς τους πελάτες που υπήρχαν στην αρχή μια συγκεκριμένης χρονικής περιόδου (μηνιαία, ετήσια). Ο υπολογισμός αυτού του δείκτη είναι ζωτικής σημασίας για μια επιχείρηση γιατί επιτρέπει τον εντοπισμό οικονομικών απωλειών με τη μορφή διαφυγόντων κερδών (revenue churn) ή την πτώση της επιχειρησιακής ανταγωνιστικότητας στην αγορά.

Το ποσοστό διαφυγής πρέπει να υπολογιστεί και να αναλυθεί, ώστε να γίνουν κατανοητοί οι λόγοι που προκαλούν τη διαφυγή πελατών ή κέρδους και να αντιμετωπιστούν.

Παραδείγματα διαφυγής πελατών είναι η ακύρωση μιας συνδρομής, το κλείσιμο ενός λογαριασμού, η μή ανανέωση ενός συμβολαίου ή μιας συμφωνίας, η απόφαση του καταναλωτή να αγοράσει ένα προϊόν από ανταγωνιστική επιχείρηση ή να χρησιμοποιήσει μία διαφορετική υπηρεσία από άλλο πάροχο.

Η μέτρηση του ποσοστού διαφυγής είναι ιδιαίτερα χρήσιμη για τις εταιρείες τηλεπικοινωνιών, καθώς ο ανταγωνισμός είναι μεγάλος και οι αλλαγές εκεί συμβαίνουν πολύ γρήγορα, οπότε το ποσοστό αυτό βοηθάει τις εταιρείες να δούν πού βρίσκονται σε σχέση με τους ανταγωνιστές τους. [7]

Στις μέρες μας, παρατηρείται ιδιαίτερα το φαινόμενο οι πελάτες κάποιας εταιρείας πορoχής υπηρεσιών να την εγκαταλείπουν και να απευθύνονται σε

κάποια άλλη για την ίδια ή παρόμοια υπηρεσία. Αυτό συμβαίνει γιατί οι πελάτες δεν είναι πλέον ικανοποιημένοι από το επίπεδο εξυπηρέτησης που τους προσφέρει η εταιρεία. Ένας από τους βασικότερους λόγους που συμβαίνει αυτό είναι η ελλιπής εξυπηρέτηση από τους υπαλλήλους της εταιρείας. Για αυτό το λόγο ο ρόλος των υπαλλήλων είναι πολύ σημαντικός για την εξυπηρέτηση πελατών.

Είναι σημαντική η δημιουργία και υιοθέτηση μοντέλων πρόβλεψης αποχώρησης πελατών (churn models), καθώς έτσι επιτυγχάνεται έγκαιρος εντοπισμός πελατών που έχουν αυξημένη πιθανότητα να αποχωρήσουν. Ακόμα, η χρήση τέτοιων μοντέλων μπορεί να έχει ως αποτέλεσμα την πιο αποδοτική ανάγνωση της αγοράς από τις εταιρείες, κάτι που θα οδηγήσει στη δημιουργία νέων προϊόντων, βελτίωση του τρόπου λειτουργίας της εταιρείας, χάραξη επιθετικών ανταγωνιστικών στρατηγικών και δημιουργία ανταγωνιστικού πλεονεκτήματος.

Τα τελευταία χρόνια υπάρχει αυξημένο ενδιαφέρον για τέτοιες μελέτες στον τομέα των τηλεπικοινωνιών, στις τράπεζες, στις ασφαλιστικές εταιρείες και σε πολλές ακόμα βιομηχανίες. Για την υλοποίηση μελετών πρόβλεψης αποχώρησης πελατών χρησιμοποιούνται αλγόριθμοι και τεχνικές μηχανικής μάθησης, όπως η ανάλυση δέντρων αποφάσεων (Decision trees learning), η λογιστική παλινδρόμηση (Logistic Regression), η ανάλυση παλινδρόμησης (Regression Analysis), ο αλγόριθμος Support Vector Machines, ο Naive Bayes, η ανάλυση καλαθιού καταναλωτή Market Basket Analysis), η προσέγγιση με ασαφή σύνολα (Rough Set Approach).

2.6 Προγενέστερες έρευνες

Η ανάλυση μεγάλων και σύνθετων βάσεων δεδομένων για τη δημιουργία μοντέλων πρόβλεψης συμπεριφοράς καταναλωτών είναι ένα ζήτημα που απασχολεί όλο και περισσότερο τις εταιρείες ανά τον κόσμο. Μέσα από την μελέτη τέτοιων βάσεων εξάγονται σημαντικά συμπεράσματα για την πορεία μιας εταιρείας, ενώ το μάρκετινγκ γίνεται όλο και πιο αποδοτικό και στοχευμένο. Η πρόβλεψη του ποσοστού αποχώρησης πελατών (***customer churn***) έχει εξελιχθεί σε πρωταρχικό στόχο των επιχειρήσεων και αποτελεί αντικείμενο μελέτης πολλών ερευνών. Η πλειοψηφία των διαθέσιμων βάσεων δεδομένων πελατών ωστόσο παρουσιάζει μεγάλη ανισοροπία στην κατανομή των κλάσεων. Είναι πολύ σύνηθες το ποσοστό πελατών που τελικά αποχωρεί από την εταιρεία, δηλαδή το αντικείμενο μελέτης των ερευνητών, να είναι πολύ μικρό σε σχέση με όσους παραμένουν, με αποτέλεσμα την αδυναμία των μοντέλων πρόβλεψης μηχανικής μάθησης να εξετάσουν αποδοτικά αυτό το ποσοστό πελατών και να οδηγηθούν σε χρήσιμα συμπεράσματα και αποτελέσματα (***class imbalance problem***). Οι έρευνες επικεντρώνονται σε τρόπους βελτίωσης της ακρίβειας και ποιότητας των

αποτελεσμάτων όσο και στην πιο ουσιαστική εξαγωγή γνώσης από τέτοιες βάσεις δεδομένων με χρήση αλγορίθμων μηχανικής μαθησης και επεξεργασίας δεδομένων.

Οι *Διαμαντάρας κ.α. (2015),* παρουσιάζουν μια συγκριτική μελέτη εφαρμογής αλγορίθμων μηχανικής μάθησης για το πρόβλημα πρόβλεψης αποχώρησης πελατών στον τομέα των τηλεπικοινωνιών. Τα αποτελέσματα δείχνουν μια ξεκάθαρη υπεροχή των ενισχυμένων (ensembled) αλγορίθμων σε σχέση με τις απλές μορφές τους, με τον αλγόριθμο Support Vector Machine και χρήση του Adaboost να πετυχαίνει μεγάλες τιμές ακρίβειας και F-measure, ενώ οι αλγόριθμοι Naïve Bayes και Logistic Regression αδυνατούν να πετύχουν υψηλές αποδόσεις πρόβλεψης. [8]

Οι *Huang et al. (2013),* αναπτύσσουν ένα υβριδικό μοντέλο πρόβλεψης αποχώρησης πελατών για το ανταγωνιστικό περιβάλλον του τομέα των τηλεπικοινωνιών, το οποίο συνδυάζει την ομαδοποίηση πελατών με χρήση μιας παραλλαγής του αλγορίθμου k-means (weighted k-means) και μια μέθοδο εξαγωγής κανόνων (FOIL) με τους κλασικούς αλγορίθμους ταξινόμησης. Έχοντας ως κύριο χαρακτηριστικό μέτρησης την τιμή AUC, το υβριδικό μοντέλο αποδεικνύεται ιδιαίτερα αξιόπιστο στην πρόβλεψη αποχώρησης των πελατών. [9]

Οι *Farquad et al. (2014)*, εφαρμόζουν σε μια υψηλού ποσοστού ανισορροπίας κλάσεων βάση δεδομένων πελατών τράπεζας ένα υβριδικό μοντέλο πρόβλεψης αποχώρησης πελατών. Αρχικά εφαρμόζονται μέθοδοι εξισορρόπησης των κλάσεων (SMOTE, Undersampling, Oversampling), με τη μέθοδο SMOTE να αποδεικνύεται η καλύτερη, και στη συνέχεια εφαρμόζεται ο αλγόριθμος Support Vector Machine για μείωση των μεταβλητών και την εξαγωγή κανόνων και ο Naïve Bayes Tree για την ταξινόμηση των πελατών, δημιουργώντας ένα αξιόπιστο μοντέλο εξαγωγής κανόνων και πρόβλεψης, που μπορούν να χρησιμοποιιηθούν για τον εντοπισμό πελατών με αυξημένη πιθανότητα αποχώρησης. [10]

Οι *Glady et al. (2006),* χρησιμοποιήσαν τη αξία διάρκειας ζωής πελάτη (customer lifetime value) ως οικονομικό δείκτης για την μέτρηση της απόδοσης των μοντέλων ταξινόμησης. Οι ταξινομητές του μοντέλου λαμβάνουν υπόψην τους το κόστος λανθασμένης πρόβλεψης (cost-sensitive classifiers). Η αξία των πελατών που είναι πιθανότερο να αποχωρήσουν μειώνεται σταδιακά με το χρόνο. Το μοντέλο επιτυγχάνει υψηλά επίπεδα απόδοσης καθώς η ανάληψη κόστους για την λανθασμένη ταξινόμηση βοηθάει στη σωστή ταξινόμηση των πελατών και ο δείκτης της αξίας του κάθε πελάτη στο χρόνο αποτελεί ικανοποιητικό κριτήριο για την ταξινόμηση του κάθε πελάτη. [11]

Οι *Gordini et al. (2015)*, αναπτύσσεται ένα μοντέλο πρόβλεψης αποχώρησης πελατών προσαρμοσμένο στη βιομηχανία του ηλεκτρονικού εμπορίου. Μετράται η ικανότητα πρόβλεψης με χρήση του αλγορίθμου Support Vector Machine, δίνοντας

ιδιαίτερη βάση στην τιμή AUC. Η απόδοση του αλγορίθμου συγκρίνεται με άλλες μεθόδους πρόβλεψης, όπως της λογιστικής παλινδρόμησης και των νευρωνικών δικτύων. Το αποτέλεσμα της μελέτης δείχνει ό,τι η βελτιστοποίηση παραμέτρων με την τεχνική SVMauc (support vector machine based on the AUC parameter selection technique) παίζει πολύ σπουδαίο ρόλο στη βελτίωση της απόδοσης του μοντέλου, όταν αυτό εφαρμόζεται σε θορυβώδη, μη ισορροπημένα δεδομένα μάρκετινγκ. Επομένως, η ανάπτυξη στρατηγικής διατήρησης πελατών μέσα από την πρότερη προσεκτική μελέτη των δεδομένων μάρκετινγκ αποδεινκύεται πολύ ικανοποιητική στην περίπτωση του ηλεκτρονικού εμπορίου. [12]

Οι *Amin (2017),* παρουσιάζουν μια βασισμένη σε κανόνες τεχνική λήψης αποφάσεων, που στηρίζεται στη θεωρία ακατέργαστων συνόλων (Rough Sets Theory), για την εξαγωγή σημαντικών κανόνων αποφάσεων που σχετίζονται με την αποχώρηση πελατών (customer churn). Η προτεινόμενη μέθοδος έχει ως αποτέλσμα την επιτυχή ταξινόμηση των πελατών σε αποχωρήσαντες και μη. Η υλοποίηση της μεθόδου γίνεται με χρήση τεσσάρων μηχανισμών δημιουργίας κανόνων, Exhaustive Algorithm (EA), Genetic Algorithm (GA), Covering Algorithm (CA), LEM2 Algorithm (LA), με το Genetic Algorithm να πετυχαίνει την καλύτερη απόδοση με μηδενικό σχεδόν ποσοστό λανθασμένων καταχωρήσεων. [13]

Οι *Burez et al. (2016),* παρουσιάζουν μια συγκρτιτική μελέτη των μεθόδων αντιμετώπισης μη ισορροπημένων βάσεων δεδομένων για προβλήματα πρόβλεψης αποχώρησης πελατών, με την εφαρμογή των μεθόδων σε πραγματικά δεδομένα. Τα αποτελέσματα δείχνουν ότι η μέθοδος ελάττωσης του δείγματος (Undersampling), μπορεί να οδηγήσει σε βελτιωμένα αποτελέσματα, ειδικά αν χρησιμοποιείται ως βασικό μέτρο απόδοσης του μοντέλου η τιμή AUC. Επιβεβαιώνεται επίσης ότι η μέθοδος αυτή είναι δοκιμή και λάθους, δεν υπάρχει δηλαδή μια συγκεκριμένη τεχνική, αλλά εφαρμόζεται διαφορετικά για κάθε περίπτωση. [14]

Κεφάλαιο 3: Machine Learning

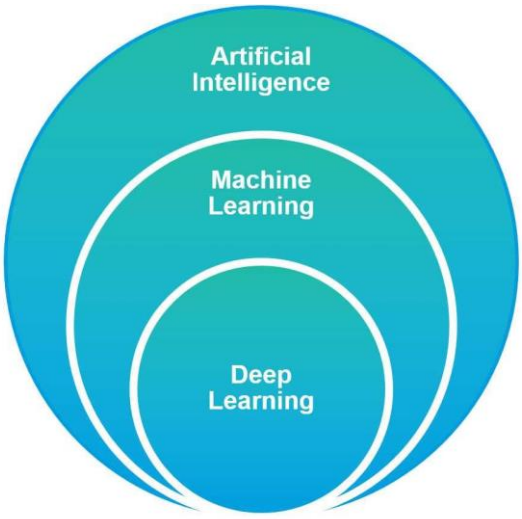
3.1 Γενικά για τη Μηχανική Μάθηση

Machine Learning (Μηχανική Μάθηση) είναι ένα υποσύνολο της Τεχνητής Νοημοσύνης. Εστιάζει κυρίως στο σχεδιασμό των συστημάτων, ώστε να τους επιτρέπει να μαθαίνουν και να κάνουν προβλέψεις βασισμένες σε κάποιας μορφής εμπειρία, δηλαδή τα δεδομένα για την περίπτωση των μηχανών.

Σύμφωνα με τον Tom M. Mitchel (1997), ένα πρόγραμμα υπολογιστή λέγεται ότι μαθαίνει από εμπειρία E(experience) ως προς μια κλάση εργασιών T(tasks) και ένα μέτρο επίδοσης P (performance measure), αν η επίδοσή του σε εργασίες της κλάσης T, όπως αποτιμάται από το μέτρο P, βελτιώνεται με την εμπειρία E.

Artificial Intelligence (Τεχνητή Νοημοσύνη) είναι ο τομέας της επιστήμης των υπολογιστών που ασχολείται με τη σχεδίαση ευφυών (νοημόνων) υπολογιστικών συστημάτων, δηλαδή συστημάτων που επιδεικνύουν χαρακτηριστικά που σχετίζονται με τη νοημοσύνη στην ανθρώπινη συμπεριφορά.

Deep Learning (Βαθιά Μάθηση) είναι ένα υποσύνολο της Μηχανικής Μάθησης το οποίο κάνει εφικτούς τους υπολογισμούς πολυεπίπεδων νευρωνικών δικτύων, εμπνευσμένα από βασικές διεργασίες του ανθρώπινου εγκεφάλου και μαθαίνουν να αναγνωρίζουν μοτίβα σε δεδομένα για τη λήψη αποφάσεων. [15]



Εικόνα 1
Τεχνητή Νοημοσύνη - Μηχανική Μάθηση – Βαθιά Μάθηση

3.1.1 Εξόρυξη δεδομένων (Data mining)

Για να λυθεί ένα πρόβλημα στον υπολογιστή, χρειάζεται ένας αλγόριθμος επίλυσης. Ένας αλγόριθμος, είναι μια αλληλουχία οδηγιών που πρέπει να πραγματοποιηθεί για να επιστρέψουν τα δεδομένα εισόδου ένα αποτέλεσμα.

Για κάθε πρόβλημα είναι πιθανό να υπάρχουν και περισσότεροι από ενας αλγόριθμοι. Ωστόσω υπάρχουν προβλήματα για τα οποία δεν υπάρχει αντίστοιχος αλγόριθμος. Εκεί εισέρχεται η Μηχανική Μάθηση (ML). Μέσα από βάσεις δεδομένων, το σύστημα εκπαιδεύεται και αποκτά γνώση ώστε να είναι σε θέση να αναγνωρίζει και να εντοπίζει συγκεκριμένα πρότυπα ή μοτίβα. Αυτά τα μοτίβα μπορεί να βοηθήσουν στην καταννόηση μιας διαδικασίας, ή μπορούν να χρησιμοποιηθούν για τη διεξαγωγή προβλέψεων. Υποθέτοντας ότι το κοντινό μέλλον δε θα διαφέρει σημαντικά από τη χρονική περίοδο που συλλέχθηκαν τα δεδομένα, οι προβλέψεις θα είναι ικανοποιητικές.

Η εφαρμογή των μεθόδων Μηχανικής Μάθησης σε μεγάλες βάσεις δεδομένων καλείται data mining (εξόρυξη δεδομένων). Ένας μεγάλος όγκος δεδομένων επεξεργάζεται για να κατασκευαστεί ένα απλό μοντέλο με πολύτιμη χρησιμότητα, έχοντας για παράδειγμα, μεγάλη προγνωστική ακρίβεια. [16]

3.1.2 Εφαρμογές Μηχανικής Μάθησης

Η Μηχανική Μάθηση βρίσκει εφαρμογή σε πολλούς τομείς. Στην οικονομία, οι τράπεζες αναλύουν παρελθοντικά δεδομένα για να χτίσουν μοντέλα που θα χτησιμοποιηθούν σε πιστωτικές εφαρμογές, εντοπισμό απάτης, καθώς και στο χρηματιστήριο. Στη βιομηχανία, μοντέλα μάθησης υιοθετούνται για βελτιστοποίηση, έλεγχο και αντιμετώπιση προβλημάτων. Στην ιατρική, προγράμματα μάθησης χρησιμοποιούνται για διαγνώσεις. Στις τηλεπικοινωνίες, μοτίβα κλήσεων αναλύονται για βελτιστοποίηση του δικτύου και μεγιστοποίηση της ποιότητας παροχής υπηρεσιών. Ακόμα, γιγάντιες ποσότητες δεδομένων στη φυσική, την ανστρονομία, τη βιολογία μπορούν να αναλυθούν μόνο με τη χρήση τέτοιων μεθόδων. Το διαδίκτυο είναι τεράστιο και συνεχώς μεγαλώνει και η αναζήτηση αξιόπιστης και χρήσιμης πληροφορίας δε μπορεί να γίνει χειροκίνητα.

Αλλά η Μηχανική Μάθηση δεν είναι απλά ένα πρόβλημα βάσης δεδομένων, αποτελεί και υποσύνολο της Τεχνητής Νοημοσύνης. Για να είναι λοιπόν ένα σύστημα έξυπνο, σε ένα διαρκώς μεταβαλλόμενο περιβάλλον, θα πρέπει να έχει την ικανότητα να μαθαίνει μόνο του.

Η Μηχανική Μάθηση δίνει λύση σε πληθώρα προβλημάτων όπως στην αναγνώριση ομιλίας και στη ρομποτική. Για παράδειγμα αναλύοντας εικόνες ενός προσώπου, ένα πρόγραμμα μάθησης εντοπίζει το μοτίβο αυτού του προσώπου και

μπορεί να το αναγνωρίσει σε άλλες εικόνες. Αυτό είναι ένα παράδειγμα αναγνώρισης μοτίβου (pattern recognition). [17]

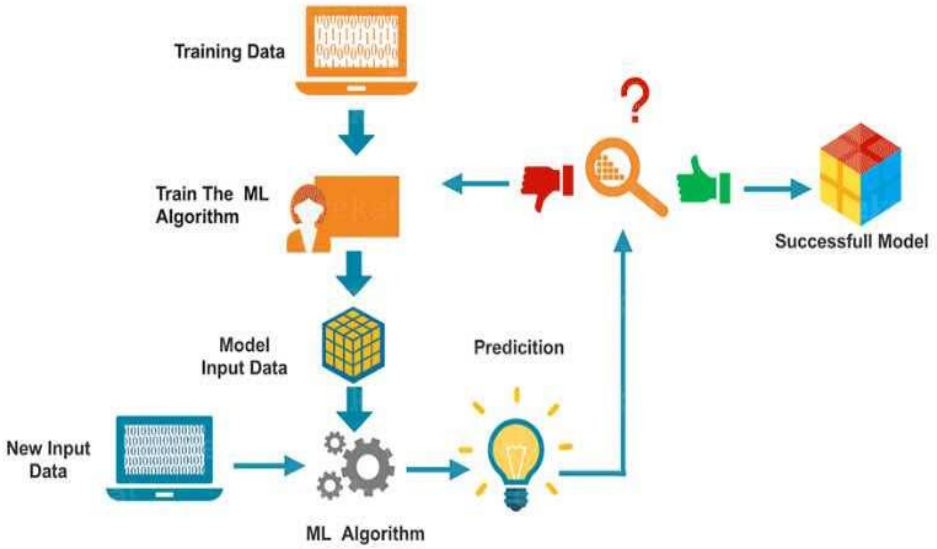
Η Μηχανική Μάθηση έχει μεγάλο αντίκτυπο στην οικονομία και στη ζωή μας γενικότερα. Ολόκληρες διαδικασίες και βιομηχανίες αυτοματοποιούνται και η αγορά εργασίας αλλάζει ριζικά. Τα μοντέλα που δημιουργούνται μπορούν να αναλύουν όλο και περισσότερα, πιο σύνθετα δεδομένα και να αποδίδουν γρηγορότερα πιο αξιόπιστα αποτελέσματα. Τέτοια μοντέλα βοηθούν στον γρήγορο εντοπισμό κερδοφόρων ευκαιριών ή στην αποφυγή ενδεχόμενων κινδύνων για τις επιχειρήσεις.

3.1.3 Υλοποίηση της Μηχανικής Μάθησης

Η Μηχανική Μάθηση προγραμματίζει ένα σύστημα ώστε να βελτιστοποιεί ένα κριτήριο απόδοσης χρησιμοποιώντας παραδειγματικά δεδομένα ή προηγούμενη εμπειρία. Το μοντέλο μπορεί να είναι προγνωστικό (predictive) για να κάνει προβλέψεις, περιγραφικό (descriptive) για την απόκτηση γνώσης, ή ένας συνδυασμός των δύο. Χρησιμοποιεί τη θεωρία της στατιστικής ανάλυσης για να χτίσει μαθηματικά μοντέλα, επειδή ο κύριος σκοπός του είναι να εξάγει συμπεράσματα από ένα δείγμα. [18]

Η Μηχανική Μάθηση στην πράξη αποτελείται από πέντε βασικά βήματα

- Προετοιμασία των δεδομένων για ανάλυση μέσω scrubbing, δηλαδή η επεξεργασία των δεδομένων με τέτοιο τρόπο ώστε αυτά να είναι διαχειρίσιμα, ακριβή και αξιόπιστα.
- Διαχωρισμός των δεδομένων σε training και σε test data. Το μοντέλο έχει επαρκή ακρίβεια όταν το ποσοστό σφάλματος (error rate) ανάμεσα σε training και test data είναι χαμηλό. Αυτό σημαίνει ότι το μοντέλο έχει μάθει τα υποκείμενα μοτίβα και τάσεις που βρίσκονται στα δεδομένα.
- Χρήση κατάλληλων αλγορίθμων για ανάλυση των δεδομένων.
- Διαμόρφωση και τροποποίηση των υποπαραμέτρων του αλγορίθμου (hyperparameters tuning).
- Ανάπτυξη μοντέλου που προβλέπει με ακρίβεια τα δεδομένα training και test.



Εικόνα 2
Υλοποίηση Μηχανικής Μάθησης

3.2 Είδη Μηχανικής Μάθησης

Έχουν αναπτυχθεί πολλές τεχνικές μηχανικής μάθησης που χρησιμοποιούνται ανάλογα με τη φύση του προβλήματος και εμπίπτουν σε δύο είδη: Μάθηση με επίβλεψη και Μάθηση χωρίς επίβλεψη.

3.2.1 Μάθηση με επίβλεψη (Supervised Learning)

Το σύστημα τροφοδοτείται από ένα σύνολο δεδομένων-καταχωρήσεων, κάθε μία από τις οποίες έχει ήδη αντιστοιχηθεί σε μία κατηγορία-κλάση (labeled data). Σκοπός της Μάθησης με επίβλεψη είναι η ταξινόμηση των καταχωρήσεων αυτών σε κλάσεις. Καθώς το σύστημα εκπαιδεύεται, προβλέπει την κλάση για τις καταχωρήσεις που δε γνωρίζει και οι προβλέψεις αυτές συγκρίνονται, με τις κλάσεις που πραγματικά ανήκουν οι καταχωρήσεις και υπολογίζεται η ακρίβεια πρόβλεψης του συστήματος. Μέσα από αυτή τη διαδικασία δοκιμής-λάθους το σύστημα αποκτά τελικά την εμπειρία έτσι ώστε να μπορεί να ταξινομήσει τις καταχωρήσεις στις υπάρχουσες κλάσεις. Η εφαρμογή της Μάθησης με επίβλεψη γίνεται με μοντέλα Ταξινόμησης (Classification) και μοντέλα Παλινδρόμησης (Regression). Παραδείγματα εφαρμογής της μάθησης με επίβλεψη είναι:

- Λογισμικό αναγνώρισης ομιλίας (speech recognition)
- Βιομετρική παρακολούθηση (αποτυπώματα)
- Πρόβλεψη πιστοληπτικής ικανότητας κατόχων πιστωτικών καρτών
- Δείκτες επανεισδοχής πελατών σε νοσοκομείο
- Αναγνώρηση ανεπιθύμητης αλληλογραφίας (spam recognition)

3.2.1.1 Ταξινόμηση (Classification)

Είναι η διαδικασία διαχωρισμού των καταχωρήσεων της βάσης δεδομένων σε διαφορετικές κλάσεις ή κατηγορίες. Χρησιμοποιείται για διάκριση αντικειμένων διαφορετικών κλάσεων (descriptive modeling) αλλά και ως εργαλείο πρόβλεψης κλάσεων των νέων εγγραφών (predictive modeling). Στη Στατιστική, καλείται Ανάλυση Διακρίσεων (Discriminant Analysis), ενώ στη Μηχανική, Αναγνώριση Μοτίβων (Pattern Recognition). Οι κλάσεις στην ουσία αποτελούν κατηγορίες αντικειμένων τα οποία παρουσιάζουν όμοια χαρακτηριστικά. Ένα αντικείμενο μπορεί να ανήκει μόνο σε μία κλάση (binary classification) είτε μπορεί να ανήκει σε περισσότερες από μία κλάσεις (multi-label classification). Επίσης, σε ένα πρόβλημα ταξινόμησης μπορεί να έχουμε παραπάνω από δύο κλάσεις (multi-class classification).

3.2.1.2 Παλινδρόμηση (Regression)

Είναι τεχνική πρόβλεψης μοντέλου που ερευνά τη σχέση ανάμεσα στη μεταβλητή απόκρισης (response variable) ή μία εξαρτημένη μεταβλητή (dependent variable) με ένα σύνολο επεξηγηματικών μεταβλητών (explanatory variables) ή ανεξάρτητων μεταβλητών (independent variables). Η παλινδρόμηση χρησιμοποιείται για συνεχείς μεταβλητές και επιτρέπει την πραγματοποίηση προβλέψεων, μαθαίνοντας τη σχέση μεταξύ των χαρακτηριστικών των δεδομένων, κάτι που χρησιμοποιείται σε πληθώρα εφαρμογών όπως την πρόβλεψη του καιρού.

3.2.2 Μάθηση χωρίς επίβλεψη (Unsupervised Learning)

Το εκπαιδευόμενο σύστημα τροφοδοτείται με δεδομένα, το σύνολο των οποίων δεν έχει αντιστοιχηθεί σε κάποια κλάση (unlabeled data). Στόχος είναι η αναγνώριση και η ομαδοποίηση αυτών των δεδομένων σε ομάδες (clusters) που δεν είναι γνωστές εκ των προτέρων, μέσα από την αναζήτηση ομοιοτήτων και συγκεκριμένων μοτίβων που εμφανίζονται στα δεδομένα εισόδου. Στη Στατιστική, αυτό ονομάζεται εκτίμηση πυκνότητας (density estimation). Οι κύριες μέθοδοι εφαρμογής της μάθησης χωρίς επίβλεψη είναι η Συσταδοποίηση (Clustering) και η εξαγωγή κανόνων συσχέτισης (Association rules). Παραδείγματα αυτής της κατηγορίας Μηχανικής Μάθησης είναι:

- Ομαδοποίηση ανθρώπων με βάση κάποιο χαρακτηριστικό
- Κατάτμηση πελατών με βάση την αγοραστική τους συμπεριφορά
- Προτάσεις προϊόντων σε πελάτες με βάση προηγούμενες αγορές τους

3.2.2.1 Ομαδοποίηση (Clustering)

Ο στόχος αυτής της μεθόδου είναι να βρεθούν και να διαχωριστούν συστάδες (clusters). Ένα μοντέλο συσταδοποίησης κατατάσσει τα δεδομένα ή τις καταχωρήσεις μιας βάσης δεδομένων ανάλογα με τις ιδιότητές τους σε ομάδες, κάνοντας μια φυσική ομαδοποίηση με βάση μη ταξινομημένα δεδομένα. Το αποτέλεσμα είναι ο διαχωρισμός ενός συνόλου (συνήθως πολυδιάστατων) δεδομένων σε ομάδες, ώστε τα σημεία που ανήκουν στην ίδια ομάδα να μοιάζουν όσο το δυνατόν περισσότερο και τα σημεία που ανήκουν σε διαφορετκές ομάδες να διαφέρουν όσο το δυνατόν περισσότερο. Η ομαδοποίηση απαιτεί κάποιο μέτρο της ομοιότητας ή διαφοράς μεταξύ των δεδομένων (απόσταση). Τυπικά μέτρα απόστασης μεταξύ δύο δεδομένων είναι η απόσταση Μανχάταν και η Ευκλείδια απόσταση.

3.2.2.2 Διαχωριστική Ομαδοποίηση (Partitional) – k-means algorithm

Ένας διαμερισμός των αντικειμένων σε μη επικαλυπτόμενα υποσύνολα (ομάδες), τέτοιος ώστε κάθε αντικείμενο να ανήκει ακριβώς σε ένα υποσύνολο.

3.2.2.2.1 K-means Algorithm

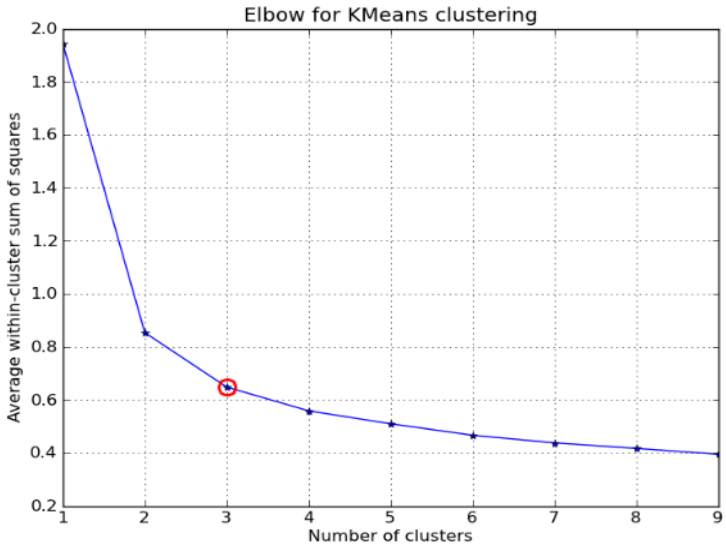
Ο πιο δημοφιλής αλγόριθμος διαχωριστικής συσταδοποίησης είναι ο αλγόριθμος **k-means**. Τα βασικά βήματα του αλγορίθμου είναι τα εξήςss:

- Δέχεται ως είσοδο τον αριθμό των συστάδων(clusters) που επιλέγεται τυχαία
- Κάθε ομάδα συσχετίζεται με ένα κεντρικό σημείο (centroid)
- Υπολογίζεται η Ευκλείδια απόσταση των σημείων από κάθε κεντρικό σημείο και κάθε σημείο ανατίθεται στην ομάδα με το κοντινότερο κέντρο
- Επανα-υπολογισμός κέντρου/κεντροειδούς κάθε συστάδας
- Επαναληπτικά, τα αντικείμενα μετακινούνται ανάμεσα στο σύνολο των συστάδων, έως ότου εντοπιστεί το επιθυμητό σύνολο. Αυτή η διαδικασία συνεχίζεται μέχρι να ελαχιστοποιηθεί η μέση τετραγωνική απόσταση των δεδομένων από τα πλησιέστερα κέντρα/κεντροειδή των συστάδων [19]

Ο k-means είναι απλός και κατανοητός αλγόριθμος, όπου τα αντικείμενα ανατίθενται αυτόματα σε ομάδες, έτσι δε μπορεί να χειριστεί τόσο αποτελεσματικά το θόρυβο και τα ακραία σημεία (outliers), ενώ ο αριθμός των ομάδων πρέπει να οριστεί από την αρχή. Επίσης, οι ομάδες που παράγονται μπορεί να διαφέρουν από το ένα τρέξιμο του αλγορίθμου στο άλλο. Το κέντρο μιας ομάδας είναι (συνήθως) το μέσο (mean) των σημείων της (το οποίο μπορεί να μην είναι ένα από τα δεδομένα εισόδου).

Για να επιλεγεί όσο το δυνατόν καλύτερος αρχικός αριθμός συστάδων μπορεί να χρησιμοποιηθεί η μέθοδος ελαχίστων διακυμάνσεων του Ward. Είναι μια διαδικασία συσταδοποίησης σε κάθε στάδιο της οποίας συνενώνονται οι συστάδες με το μικρότερο άθροισμα τετραγώνων των σφαλμάτων (least sum of the squared error-SSE), όπου το άθροισμα των τετραγώνων λειτουργεί ως κριτήριο της απώλειας.

Για τον προσδιορισμό του τελικού αριθμού των συστάδων σε ένα σύνολο δεδομένων εφαρμόζεται η μέθοδος του αγκώνα (elbow method). Η χρήση του αγκώνα ως σημείου αποκοπής είναι μια κοινή ευρετική μέθοδος στη μαθηματική βελτιστοποίηση για την επιλογή ενός σημείου όπου οι φθίνουσες αποδόσεις δεν αξίζουν πλέον το πρόσθετο κόστος. Στην ομαδοποίηση, αυτό σημαίνει ότι κάποιος πρέπει να επιλέξει έναν αριθμό συστάδων, έτσι ώστε η προσθήκη ενός άλλου συμπλέγματος να μην παρέχει πολύ καλύτερη μοντελοποίηση των δεδομένων.



Εικόνα 3
Elbow method για ομαδοποίηση με χρήση του k-means

Ο έλεγχος ποιότητας συσταδοποίησης μπορεί να γίνει και με τη χρήση του Συντελεστή Σκιαγράφησης (Silhouette coefficient). Ο μέσος s(i) του συνόλου των σημείων αποτελεί ένα μέτρο του πόσο καλά έχουν συσταδοποιηθεί τα δεδομένα. Στόχος είναι ο μέσος Silhouette να είναι όσο γίνεται πιο κοντά στη μονάδα. Εάν αυτό δε συμβαίνει, θα ξανατρέξουμε τον k-means με διαφορετικό αριθμό συστάδων και θα κρατήσουμε τη συσταδοποίηση με την υψηλότερη τιμή του συντελεστή σκιαγράφησης.

3.2.2.2 Ιεραρχική Ομαδοποίηση (Hierarchical)

Είναι μία εναλλακτική προσσέγγιση που προσπαθεί με ιεραρχικό τρόπο να ανακαλύψει τον αριθμό και τη δομή των ομάδων, χωρίς να χρειάζεται ο ακριβής ορισμός του αριθμού των ομάδων προκαταβολικά. Κάθε ομάδα έχει υποομάδες οργανωμέμες σε ένα ιεραρχικό δέντρο (εμφωλευμένες). Τα αντικείμενα σταδιακά σχηματίζουν ομάδες μεταξύ τους , ανάλογα με την ομοιότητά τους, μέχρι όλα τα αντικείμενα να έχουν αντιστοιχηθεί σε μια συστάδα.

3.2.2.3 Εξαγωγή κανόνων συσχέτισης (Association rules)

Είναι μία μέθοδος για την ανακάλυψη ενδιαφέρων σχέσεων μεταξύ μεταβλητών σε μεγάλες βάσεις δεδομένων χρησιμοποιώντας ορισμένα μέτρα ενδιαφέροντος. Βασική εφαρμογή των συγκεκριμένων αλγορίθμων είναι η ανάλυση της καταναλωτικής συμπεριφοράς των πελατών, βάσει της ταυτόχρονης αγοράς προϊόντων ή υπηρεσιών. Η ανάλυση καλαθιού αγοράς (market basket analysis), εξηγεί τους συνδυασμούς προϊόντων που εμφανίζονται συχνά μαζί στις συναλλαγές. [20]

Αυτό που μας ενδιαφέρει είναι μια δεσμευμένη πιθανότητα της μορφής P(X/Y), όπου Y είναι το προϊόν που θέλουμε να συσχετίσουμε με το X, το οποίο είναι το προϊόν που ξέρουμε ότι έχει ήδη αγοράσει ο πελάτης.

Για την εξαγωγή κανόνων συσχέτισης υπάρχουν τρία μεγέθη που υπολογίζονται συνήθως: support, confidence και το ποσοτικό μέτρο lift interest.

Confidence είναι η δεσμευμένη πιθανότητα P(X/Y), που συνήθως υπολογίζεται. Για να είναι ένας κανόνας αρκετά ισχυρός, θα πρέπει αυτή η πιθανότητα να προσεγγίζει τη μονάδα και να είναι σαφώς μεγαλύτερη από την P(Y), την πιθανότητα δηλαδή κάποιος να αγοράσει το προϊόν Y. Έχει σημασία επίσης, να μεγιστοποιηθεί η υποστήριξη (Support) του κανόνα, ακόμα και αν το μέτρο Confidence είναι πολύ υψηλό, καθώς αν οι πελάτες είναι λίγοι, ο κανόνας είναι ασήμαντος. Η υποστήριξη (Support) δείχνει τη στατιστική σημαντικότητα του κανόνα, ενώ η αυτοπεποίθηση (Confidence) δείχνει τη δύναμη του κανόνα. Ορίζεται η ελάχιστη Support και Confidence του κανόνα, και επιλέγονται οι κανόνες που πληρούν αυτά τα κριτήρια.

Αν X και Y είναι ανεξάρτητες μεταξύ τους, τότε το Lift θα έχει τιμή κοντά στη μονάδα. Αν το εύρος διαφέρει -αν P(X/Y) και P(Y) είναι διαφορετικά- περιμένουμε να υπάρχει μια εξάρτηση μεταξύ των δύο. Αν το Lift είναι μεγαλύτερο της μονάδας, η αγορά του X κάνει την αγορά του Y πιο πιθανή , ενώ αν είναι μικρότερο της μονάδας, η αγορά του X κάνει την αγορά του Y λιγότερο πιθανή.

Αυτές οι πρακτικές μπορούν εύκολα να γενικοποιηθούν και για περισσότερα από δύο αντικείμενα. Για παράδειγμα, [X,Y,Z] είναι ένα σύνολο τριών αντικειμένων, και μπορούμε να αναζητήσουμε ένα κανόνα X,Z→Y, P(Y/X,Z).

- Support=P (X, Y) (costumers who bought X and Y)/(costumers)
- Confidence=P (X, Y)/P(X)
- Lift: δείχνει τη δύναμη του κανόνα ως προς την τυχαία επανεμφάνιση του X και του Y. Lift=P(Y/X)/P(Y)

3.2.2.3.1 Apriori Algorithm

Ο **Apriori** είναι ένας αλγόριθμος που χρησιμοιείται ευρέως για την εξαγωγή κανόνων συσχέτισης. Προχωράει εντοπίζοντας τα συχνά μεμονωμένα στοιχεία στη βάση δεδομένων και επεκτείνοντάς τα σε όλο και μεγαλύτερα σύνολα στοιχείων, αρκεί αυτά τα σύνολα στοιχείων να εμφανίζονται αρκετά συχνά στη βάση δεδομένων.

Τα συχνά σύνολα (frequent itemsets) που αναζητά ο Apriori είναι σύνολα που η τιμή Support ξεπερνά ένα ορισμένο κατώφλι. Ο αλγόριθμος ξεκινά παίρνοντας ένα ζευγάρι μεταβλητών. Για δεδομένη τιμή Confidence, υπολογίζει το μέτρο support και αν είναι ανώτερο του κατωφλίου που έχει οριστεί, κρατάει τον κανόνα συσχέτισης του ζευγαριού. Στη συνέχεια δημιουργεί ζεύγη περσσότερων μεταβλητών προοδευτικά, κρατώντας μόνο τους κανόνες με την επιθυμητή τιμή support. Σταδιακά λοιπόν παράγονται όλο και πιο σύνθετοι κανόνες. Μέσα από αυτή τη διαδικασία γίνεται δυνατή η εξαγωγή γνώσης μέσω των κανόνων (knowledge extraction) και εντοπίζονται μοτίβα που μπορούν να χρησιμοποιηθούν για την κατανόηση ενός φαινομένου ή της συμπεριφοράς ορισμένης μερίδας πελατών με απώτερο σκοπό την καλύτερη και πιο στοχευμένη στρατηγική μάρκετινγκ και διαχείρισης.

3.3 Δεδομένα (Data)

Ως δεδομένα αναφέρονται τα γεγονότα και τα στατιστικά στοιχεία που συγκεντρώνονται για ανάλυση. Ένα οργανωμένο σύνολο δεδομένων αποτελεί μία βάση δεδομένων (dataset). Περιγραφική στατιστική (Descriptive Statistics), είναι η μέθοδος που χρησιοποιείται για την περιγραφή και κατανόηση των χαρακτηριστικών μιας συγκεκριμένης βάσης δεδομένων, δίνοντας σύντομες περιλήψεις για το δείγμα και τις μετρήσεις των δεδομένων.

Τα δεδομένα μπορεί να είναι ποιοτικά (qualitative) ή ποσοτικά (quantitative). Τα ποιοτικά δεδομένα έχουν να κάνουν με χαρακτηριστικά και περιγραφές που δε μπορούν να μετρηθούν εύκολα, αλλά μπορούν να παρατηρηθούν ξακάθαρα. Αυτά μπορεί να είναι ονομαστικά (nominal), όπου δεν εκφράζεται κάποια προτίμηση ή συγκεκριμένη σειρά κατάταξης, όπως το γένος, η καταγωγή ή ο τόπος κατοικίας, είτε δεδομένα διάταξης (ordinal), στα οποία έχει νόημα η διάταξη αλλά όχι η διαφορά μεταξύ των τιμών, όπως μια κλίμακα ικανοποίησης (1:καθόλου-10:πάρα πολύ), ή μια επιλογή προτίμησης (καλό, κακό, μέτριο).

Τα ποσοτικά δεδομένα μεταβάλλονται από άποψη ποσότητας και εκφράζονται με μία μονάδα μέτρησης. Χωρίζονται σε διακριτά (discrete), όπως είναι ο αριθμός πελατών, και συνεχή (continuous), όπως δηλαδή ο χρόνος προϋπηρεσίας ενός εργαζομένου ή ο μισθός του.

3.3.1 Προετοιμασία Δεδομένων (Data Preprocessing)

Είναι η διαδικασία μετατροπής ακατέργαστων δεδομένων σε μια μορφή που μπορούν να μοντελοποιηθούν με αλγορίθμους Μηχανικής Μάθησης. Η μάθηση μέσω της ανάλυσης δεδομένων είναι μία βασική εργασία της διαδικασίας δημιουργίας μοντέλων πρόβλεψης. Τα δεδομένα αφορούν παραδείγματα ή περιπτώσεις από τον τομέα που χαρακτηρίζει το πρόβλημα για το οποίο αναζητείται η λύση. Σε ένα μοντέλο πρόβλεψης, όπως ένα μοντέλο ταξινόμησης (classification) ή παλινδρόμησης (regression), τα δεδομένα δε μπορούν να χρησιμοποιηθούν απευθείας. Υπάρχουν τέσσερις βασικοί λόγοι που συμβαίνει αυτό:

- Τύπος δεδομένων: η υλοποίηση κάποιων αλγορίθμων απαιτεί τα δεδομένα να είναι αριθμητικά ή ονομαστικά αντίστοιχα
- Προϋποθέσεις δεδομένων: σε κάποιους αλγορίθμους τα δεδομένα υπόκεινται σε συγκεκριμένους περιορισμούς για να χρησιμοποιηθούν
- Σφάλματα δεδομένων: στατιστικός θόρυβος και σφάλματα στις καταχωρήσεις των δεδομένων χρειάζεται να διορθωθούν
- Πολυπλοκότητα δεδομένων: πολύπλοκες μη γραμμικές σχέσεις πρέπει να εξαιρεθούν από τη βάση δεδομένων

Οι πιο σημαντικές τεχνικές προετοιμασίας δεδομένων είναι:

- Καθαρισμός δεδομένων: αφαίρεση δεδομένων που προκαλούν ακραίο θόρυβο (outliers) ή καθαρισμών λανθασμένων καταχωρήσεων
- Feature selection: αναγνώριση και επιλογή των χαρακτηριστικών που είναι πιο σημαντικά για την επίλυση του ζητούμενου
- Μετατροπή δεδομένων: μετατροπή σε αντίστοιχη διαχειρίσιμη μορφή ανάλογα με τη φύση του προβλήματος και του εκάστοτε αλγορίθμου
- Feature engineering: η διαδικασία χρήσης της υπάρχουσας γνώσης ή πληροφορίας για εξαγωγή χαρακτηριστικών (ιδιότητες, επιλογή σημαντικών μεταβλητών) από μη επεξεργασμένα δεδομένα
- Dimensionality reduction: η διαστατική μείωση είναι μια εναλλακτική της επιλογής χαρακτηριστικών καθώς πραγματοποιεί πρόβλεψη για τις νέες καταχωρήσεις σε ένα μικρότερο διαστατικά χώρο. Μία χαρακτηριστική μέθοδος της διαστατικής μείωσης είναι η ανάλυση κυρίων συνιστωσών (Principal Component Analysis)

Στόχος είναι να βρεθεί ο τρόπος ώστε τα δεδομένα που θα εισαχθούν στον αλγόριθμο να είναι όσο το δυνατόν καλύτερα δομημένα για να αποδόσει το

μοντέλο πρόβλεψης. Η προετοιμασία των δεδομένων είναι ο δρόμος που θα οδηγήσει σε αυτό το αποτέλεσμα. [21]

3.3.2 Επιλογή Χαρακτηριστικών (Feature Selection)

Είναι η διαδικασία μείωσης του αριθμού των εισακτέων μεταβλητών κατά το χτίσιμο ενός μοντέλου πρόβλεψης. Είναι γενικά επιθυμητό ο αριθμός των μεταβλητών να μειωθεί για να μειωθεί το υπολογιστικό κόστος του μοντέλου αλλά και, σε κάποιες περιπτώσεις, να αυξηθεί η απόδοση του. Η επιλογή χαρακτηριστικών επικεντρώνεται στην αφαίρεση των μη-σημαντικών ή περιττών μεταβλητών από το μοντέλο. Με αυτή τη διαδικασία επίσης μπορεί να αποφευχθεί το ενδεχόμενο υπερπροσαρμογής (overfitting).

Η διαδικασία χωρίζεται σε επιτηρούμενη (supervised) και μη-επιτηρούμενη (unsupervised). Στην επιτηρούμενη επιλογή χαρακτηριστικών αφαιρούνται οι μεταβλητές που δε σχετίζονται με μία συγκεκριμένη μεταβλητή (target variable), ενώ στη μη-επιτηρούμενη, η αφαίρεση γίνεται με βάση τη συσχέτιση (correlation) που υπάρχει μεταξύ των μεταβλητών.

Οι μέθοδοι επιτηρούμενης επιλογής χαρακτηριστικών χωρίζονται σε wrapper και filter methods. Στη μέθοδο wrapper δημιουργούνται πολλά μοντέλα με διαφορετικό υποσύνολο μεταβλητών κάθε φορά και τελικά επιλέγονται οι μεταβλητές που δίνουν την καλύτερη απόδοση για ένα συγκεκριμένο μέτρο απόδοσης.

Οι μέθοδοι filter χρησιμοποιούν στατιστικές μεθόδους για να αξιολογήσουν τη σχέση κάθε μεταβλητής ως προς μια κεντρική μεταβλητή. Για το τελικό μοντέλο επιλέγονται οι μεταβλητές που έχουν τη μεγαλύτερη σχέση.

Ακόμα, υπάρχουν κάποιοι αλγόριθμοι μηχανικής μάθησης που εκτελούν αυτόματα επιλογή χαρακτηριστικών ως μέρος της διαδικασίας μάθησης του μοντέλου και αναφέρονται ως μέθοδοι εσωτερικής επιλογής χαρακτηριστικών (intrinsic). Αυτή η κατηγορία περιλαμβάνει αλγορίθμους όπως τα τιμωριτικά μοντέλα LASSO ή τα δέντρα αποφάσεων (Decision Trees), ενώ περιλαμβάνουν και τα συνδυαστικά (ensemble) δέντρα αποφάσεων όπως τον αλγόριθμο Random Forest.

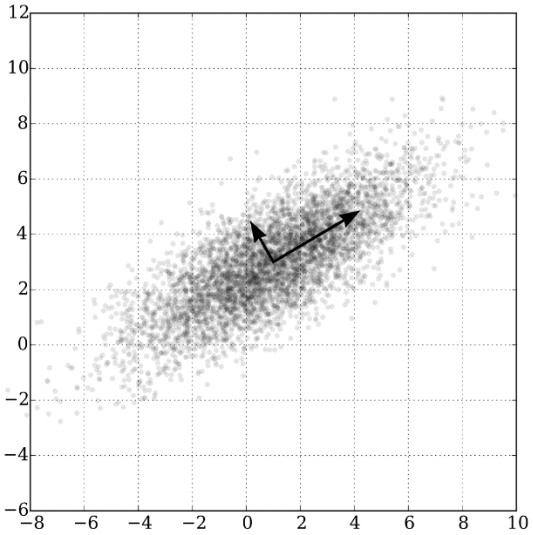
Η μη-επιτηρούμενη επιλογή χαρακτηριστικών βρίσκει εφαρμογή με τον πίνακα συσχέτισης (correlation matrix). Είναι ένας πίνακας που δείχνει τη συσχέτιση μεταξύ των μεταβλητών (-1 δεν υπάρχει συσχέτιση, 1 υπάρχει μεγάλη συσχέτιση). Κάθε κελί του πίνακα δείχνει τη συσχέτιση ενός ζεύγους μεταβλητών. Ο πίνακας συσχέτισης χρησιμοποιείται για να ομαδοποιήσει δεδομένα όταν στόχος είναι ο εντοπισμός μοτίβων, σαν αρχή για πιο εξειδικευμένες αναλύσεις, σαν αποκωδικοποιητής των μεταβλητών καθώς μπορούν να εντοπιστούν ενδιαφέρουσες συσχετίσεις. Μελετώντας τον πίνακα μπορεί να γίνει επιλογή μεταβλητών ανάλογα με τη φύση του προβλήματος. [22]

3.3.3 Dimensionality reduction - Principal Component Analysis

Η διαστατική μείωση (Dimensionality Reduction) είναι περισσότερο μία εναλλακτική την επιλογής χαρακτηριστικών παρά μία επιμέρους μέθοδός της. Οι βάσεις δεδομένων που χρησιμοποιούνται συνήθως για μελέτη αποτελούνται από πολλές μεταβλητές (features), οι οποίες δυσκολεύουν τη διαδικασία πρόβλεψης και αυξάνουν το υπολογιστικό κόστος. Για δεδομένο αριθμό καταχωρήσεων (rows), τα μοντέλα πρόβλεψης τείνουν να λειτουργούν πιο αποδοτικά με την ύπαρξη λιγότερων μεταβλητών (columns). Η μείωση του όγκου της πληροφορίας μπορεί να γίνει μέσω της ανάλυσης κυρίων συνιστωσών (Principal Components Analysis). [23]

Η PCA είναι ένας αλγόριθμος μη-επιτηρούμενης μάθησης που εντοπίζει τις σχέσεις μεταξύ των μεταβλητών μιας βάσης δεδομένων. Είναι επίσης ευρέως γνωστή ως ένα βήμα προεπεξεργασίας δεδομένων για αλγορίθμους επιτηρούμενης μάθησης. Το πλεονέκτημα της PCA είναι ό,τι το μεγαλύτερο κομμάτι της πληροφορίας που κρύβεται στα δεδομένα της βάσης διατηρείται στο ακέραιο χρησιμοποιώντας πολύ μικρότερο αριθμό μεταβλητών. Η μέθοδος PCA, αποτελεί μια γραμμική μέθοδο συμπίεσης δεδομένων, η οποία συνίσταται από τον επαναπροσδιορισμό των συντεταγμένων ενός συνόλου δεδομένων σε ένα άλλο σύστημα συντεταγμένων, το οποίο θα είναι καταλληλότερο στην επικείμενη ανάλυση δεδομένων. Αυτές οι νέες συντεταγμένες είναι το αποτέλεσμα ενός γραμμικού συνδυασμού που προέρχεται από τις αρχικές μεταβλητές και εκπροσωπούνται σε ορθογώνιο άξονα, ενώ τα επικείμενα σημεία διατηρούν μια φθίνουσα σειρά όσο αφορά την τιμή της διακύμανσης τους. Για το λόγο αυτό, η πρώτη κύρια συνιστώσα (principal component), διατηρεί περισσότερες πληροφορίες δεδομένων σε σύγκριση με τη δεύτερη, η οποία δε διατηρεί πληροφορίες που έχουν εισέλθει νωρίτερα στην πρώτη. Οι principal components δε συσχετίζονται.

Η συνολική ποσότητα των principal components είναι ίση με την ποσότητα των αρχικών μεταβλητών και παρουσιάζει τις ίδιες πληροφορίες στατιστικής. Εντούτοις, η συγκεκριμένη μέθοδος επιτρέπει τη μείωση του συνόλου των μεταβλητών, καθώς τα πρώτα συστατικά (principal components) διατηρούν περισσότερο από το 90% των στατιστικών δεδομένων από τα αρχικά δεδομένα. [24]



Εικόνα 4
Principal Component Analysis

3.4 Διαχωρισμός δεδομένων

3.4.1 Train/test split

Μπορούμε να μετρήσουμε την ικανότητα γενίκευσης μιας υπόθεσης, δηλαδή την απόδοση του αλγορίθμου, αν έχουμε πρόσβαση σε δεδομένα εκτός των εκπαιδευτικών. Είναι μια γρήγορη και εύκολη διαδικασία, τα αποτελέσματα της οποίας μας επιτρέπουν να συγκρίνουμε την απόδοση αλγορίθμων μηχανικής μάθησης για προβλήματα πρόβλεψης. Εφαρμόζεται σε προβλήματα ταξινόμησης ή παλινδρόμησης και για κάθε αλγόριθμο επιτηρούμενης μάθησης.

Η βάση δεδομένων χωρίζεται σε δύο μέρη. Το ένα μέρος περιλαμβάνει τα δεδομένα εκπαίδευσης που συνήθως αποτελούν περίπου το 80% της βάσης δεδομένων, και το υπόλοιπο ποσοστό καλείται σύνολο επιβεβαίωσης (validation set), και χρησιμοποιείται για τον έλεγχο της απόδοσης του μοντέλου. Μόλις το μοντέλο επιλεγεί και εκπαιδευτεί, εφαρμόζεται σε νέα, μη γνωστά δεδομένα, για την επίλυση του εκάστοτε προβλήματος πρόβλεψης. Η μέθοδος αυτή είναι αποδοτική για μεγάλες βάσεις δεδομένων.

3.4.2 k-fold cross validation

Όταν τα δεδομένα σε μια βάση δεν είναι επαρκή σε ποσότητα, χρησιμοποιείται η μέθοδος k-fold cross validation. Είναι μια στατιστική μέθοδο που δίνει προβλέψεις με γενικά χαμηλότερη μεροληψία συγκριτικά με άλλες μεθόδους. Η παράμετρος k αναφέρεται στον αριθμό των επιμέρους συνόλων στα οποία θα χωριστεί το δείγμα. Το πρώτο σύνολο συμπεριφέρεται ως validation set, ενώ τα υπόλοιπα k-1 σύνολα χρησιμοποιούνται ως training sets. Η μέθοδος εφαρμόζεται k

φορές και το τελικό αποτέλεσμα είναι ο αριθμητικός μέσος των k επαναλήψεων. [25]

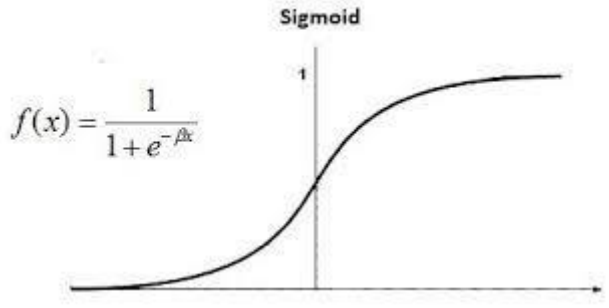
3.5 Αλγόριθμοι Επιτηρούμενης Μηχανικής Μάθησης (Supervised ML algorithms)

Οι αλγόριθμοι που χρησιμοποιήθηκαν για την ταξινόμηση των πελατών στις κλάσεις είναι οι Logistic Regression, k-Nearest Neighbors, Naïve Bayes, Random Forest και το θεωρητικό τους υπόβαθρο αναπτύσσεται παρακάτω.

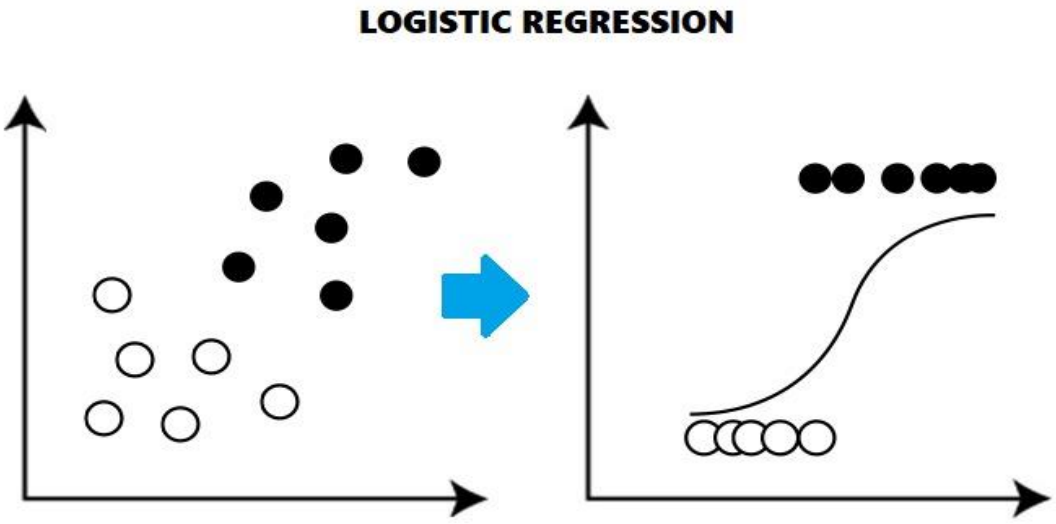
3.5.1 Logistic Regression

Ο αλγόριθμος λογιστικής παλινδρόμησης (Logistic Regression) μετράει τη σχέση της εξαρτημένης μεταβλητής (ονομαστική μεταβλητή κλάσης, αυτή που θέλουμε να προβλέψουμε) με τις υπόλοιπες ανεξάρτητες μεταβλητές (χαρακτηριστικά του dataset), υπολογίζοντας πιθανότητες χρησιμοποιώντας τη λογιστική συνάρτηση. Η εξαρτημένη μεταβλητή πρέπει να είναι διακριτή (discrete). Η λογιστική παλινδρόμηση δίνει διακριτά αποτελέσματα, ενώ η γραμμική παλινδρόμηση δίνει συνεχή αποτελέσματα. Η λογιστική παλινδρόμηση υπολογίζει την πιθανότητα ένα ενδεχόμενο να συμβεί (*p*), προς την πιθανότητα το ενδεχόμενο αυτό να μη συμβεί (*1-p*).

Οι πιθανότητες που υπολογίζονται μετατρέπονται στη συνέχεια σε δυαδικές τιμές έτσι ώστε να πραγματοποιηθεί η πρόβλεψη. Αυτή είναι η αποστολή της λογιστικής συνάρτησης, που καλείται αλλιώς σιγμοειδής (Sigmoid function). Οι τιμές που υπολογίζονται μετατρέπονται σε 0 ή 1 με χρήση ενός κατωφλίου ταξινόμησης (threshold classifier). Η τιμή κατωφλίου καθορίζει σε μεγάλο βαθμό την επιτυχία του μοντέλου. Η σιγμοειδής συνάρτηση, η οποία έχει τη χαρακτηριστική μορφή του γράμματος *S*, παίρνει οποιαδήποτε τιμή και την τοποθετεί μεταξύ του 0 και του 1, αλλά ποτέ ακριβώς στα όρια.



Εικόνα 5
Γραφική αναπαράσταση της λογιστικής συνάρτησης



Εικόνα 6
Υλοποίηση της λογιστικής παλινδρόμησης

3.5.2 k-Nearest Neighbors

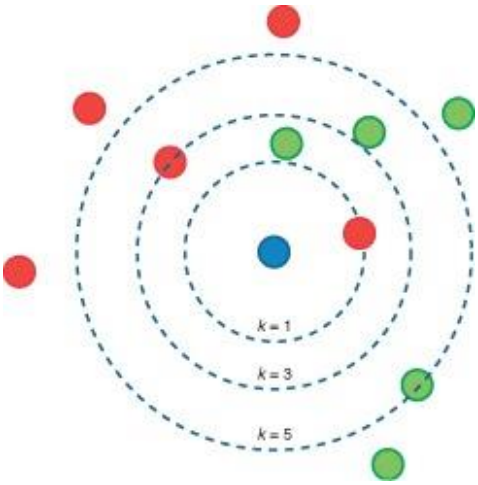
Ο αλγόριθμος του πλησιέστερου γείτονα (k-Nearest Neighbor - kNN) είναι ένας από τους πιο διαδεδομένους αλγορίθμους που βρίσκει εφαρμογή σε προβλήματα ταξινόμησης και παλινδρόμησης, καθώς η λειτουργία του είναι πολύ απλή και σε συγκεκριμένες περιπτώσεις παράγει πολύ ικανοποιητικά αποτελέσματα. Είναι μια μή παραμετρική μέθοδος (δεν κάνει κάποια υπόθεση για την υποκείμενη κατανομή των δεδομένων), βασισμένη στα παραδείγματα (ο αλγόριθμος δε μαθαίνει με σαφήνεια ένα μοντέλο αλλά απομνημονεύει τα εκπαιδευτικά παραδείγματα). Για αυτό το λόγο καλείται και τεμπέλης (lazy algorithm).

Όταν χρησιμοποιείται σε περιπτώσεις ταξινόμησης το αποτέλεσμα είναι η ένταξη κάθε αντικειμένου σε μια τάξη (προβλέπει μια τάξη – μια διακριτή τιμή). Υπάρχουν τρία βασικά στοιχεία για αυτή την προσέγγιση: ένα σύνολο επισημασμένων αντικειμένων (labeled objects), η απόσταση μεταξύ των αντικειμένων και η τιμή του k, του αριθμού δηλαδή των πλησιέστερων γειτόνων.

Για την ταξινόμηση ενός μη επισημασμένου αντικειμένου, υπολογίζεται η απόσταση από αυτό με τα επισημασμένα αντικείμενα, και η τάξη της πλειοψηφίας των πλησιέστερων γειτόνων χαρακτηρίζει τελικά το αντικείμενο, το αντικείμενο δηλαδή κατανέμεται στην αντίστοιχη τάξη. Η Ευκλείδια απόσταση (Euclidean distance) είναι η πιο κοινή μέθοδος μέτρησης της απόστασης μεταξύ των αντικειμένων και υπολογίζεται ως η τετραγωνική ρίζα των διαφορών των αντικειμένων. Άλλες δημοφιλείς μέθοδοι μέτρησης της απόστασης είναι η απόσταση Manhattan και η απόσταση Minkowski.

$$euclidean = d(x,y) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \qquad (1)$$

Ο προσδιορισμός του ιδανικού αριθμού πλησιέστερων γειτόνων k δεν είναι πάντα εύκολος. Ένας μικρός αριθμός γειτόνων σημαίνει ότι ο θόρυβος θα έχει μεγαλύτερη επιρροή στο αποτέλεσμα ενώ ένας μεγάλος αριθμός γειτόνων θα αυξήσει σημαντικά το υπολογιστικό κόστος. Κάθε πρόβλημα ταξινόμησης έχει τα δικά του χαρακτηριστικά και ιδιαιτερότητες οπότε ο ιδανικός αριθμός γειτόνων εξαρτάται από τη φύση του προβλήματος και προκύπτει από τη διαδικασία δοκιμής και λάθους.



Εικόνα 7
Γραφική αναπαράσταση k-Nearest Neighbor

3.5.3 Naïve Bayes

Ο Naïve Bayes είναι ένας πιθανολογικός αλγόριθμος μηχανικής μάθησης που βασίζεται στο θεώρημα Bayes, και χρησιμοποιείται σε πληθώρα προβλημάτων ταξινόμησης. Το θεώρημα Bayes χρησιμοποιείται για τον υπολογισμό πιθανοτήτων υπό όρους, δηλαδή της πιθανότητας να συμβεί ένα γεγονός δεδομένου ότι έχει συμβεί ένα άλλο γεγονός.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (2)$$

Η θεμελιώδης υπόθεση του αλγορίθμου είναι ότι κάθε χαρακτηριστικό έχει ανεξάρτητη και ίση συνεισφορά στη διαμόρφωση του αποτελέσματος. Οι μεταβλητές είναι ανεξάρτητες μεταξύ τους και κανένα χαρακτηριστικό δεν επηρεάζει περισσότερο ή λιγότερο το αποτέλεσμα της ταξινόμησης. Αυτές οι υποθέσεις δε μπορούν να ισχύουν σε πραγματικές συνθήκες καθώς είναι ιδανικές, για αυτό και ο αλγόριθμος λέγεται αφελής (Naïve). Το θεώρημα Bayes γράφεται ως εξής:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \qquad (3)$$

Η μεταβλητή *y* είναι η μεταβλητή κλάσης και η μεταβλητή *X* αντιπροσωπεύει τις παραμέτρους/χαρακτηριστικά του προβλήματος, που εισάγονται κάθε φορά στο πρόβλημα.

$$X = (x_1, x_2, x_n, \dots) \qquad (4)$$

Για την περίπτωση δυαδικής ταξινόμησης, η ταξινόμηση του κάθε αντικειμένου γίνεται με τον παρακάτω κανόνα:

$$Αν\ P(y_1|X) > P(y_2|X), το\ X\ ανήκει\ στην\ κλάση\ y_1 \qquad (5)$$

$$Αν\ P(y_1|X) < P(y_2|X), το\ X\ ανήκει\ στην\ κλάση\ y_2 \qquad (6)$$

Εκτός από τον κλασικό αλγόριθμο Naïve Bayes, υπάρχουν και παραλλαγές του, όπως ο Multinomial Naïve Bayes, ο Bernoulli Naïve Bayes και ο Gaussian Naïve Bayes.

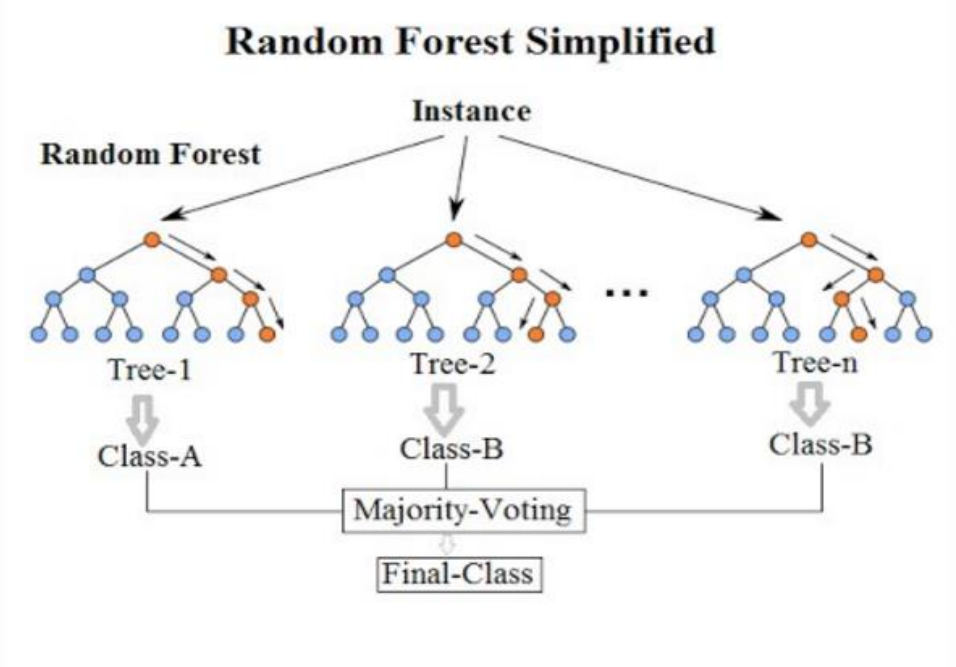
3.5.4 Random Forest

Τα δέντρα αποφάσεων (decision trees) είναι ένα είδος αλγορίθμου επιτηρούμενης μάθησης που χρησιμοποιείται σε προβλήματα ταξινόμησης και λειτουργεί τόσο για διακριτές όσο και για συνεχείς μεταβλητές. Κάθε κλαδί ενός δέντρου απόφασης αντιπροσωπεύει μια επιλογή μεταξύ πολλών εναλλακτικών και κάθε φύλλο αντιπροσωπεύει μια απόφαση.

Ο αλγόριθμος Random Forest αποτελελι μια ενισχυμένη (ensemble) μέθοδο μηχανικής μάθησης, που δημιουργεί ισχυρότερα μοντέλα και πετυχαίνει μεγαλύτερη ακρίβεια από τα δέντρα αποφάσεων. Ο αλγόριθμος χτίζει πολλά δέντρα αποφάσεων και τα ενώνει μεταξύ τους για να πάρει μια πιο ακριβή και αξιόπιστη πρόβλεψη. Ο αλγόριθμος λειτουργεί σε δύο βήματα. Πρώτα γίνεται η δημιουργία των δέντρων και μετά γίνεται η πρόβλεψη από τον ταξινομητή που έχει δημιουργηθεί. Η διαφορά του Random Forest από τα δέντρα αποφάσεων είναι ότι στον Random Forest η διαδικασία εύρεσης του ριζικού κόμβου και η διαδικασία διαχωρισμού των κόμβων γίνεται τυχαία.

Κατά τη δημιουργία των δέντρων αποφάσεων, αν ο αριθμός των περιπτώσεων στο σύνολο εκπαίδευσης είναι N, επιλέγονται N περιπτώσεις τυχαία και γίνεται αντικατάσταση από τα συνολικά δεδομένα. Αν υπάρχουν M μεταβλητές εισόδου, ένα υποσύνολο m μεταβλητών επιλέγεται τυχαία για να χωριστεί στη συνέχεια ο κόμβος.

Για τη διαδικασία της πρόβλεψης τα εκπαιδευτικά δεδομένα και οι κανόνες κάθε τυχαίου δέντρου απόφασης που έχει δημιουργηθεί συνδυάζονται για την πρόβλεψη και αποθήκευση του αποτελέσματος. Για κάθε δέντρο απόφασης προκύπτει μια πρόβλεψη. Η κλάση που έχουν προβλέψει τα περισσότερα δέντρα απόφασης αποτελεί την τελική πρόβλεψη του αλγορίθμου Random Forest. [26]



Εικόνα 8
Γραφική αναπαράσταση υλοποίησης αλγορίθμου Random Forest

3.6 Μέτρα απόδοσης αλγορίθμων Επιτηρούμενης Μηχανικής Μάθησης

Η απόδοση ενός αλγορίθμου, αποδίδεται στη ικανότητα του να αναγνωρίσει και να αντιστοιχίσει ένα δεδομένο ή μία νέα καταχώρηση στη σωστή κλάση. Η μέτρηση της απόδοσης γίνεται μέσα από την εκπαίδευση του μοντέλου στο training set και την εφαρμογή στο test set. Ακρίβεια (**Accuracy**) είναι το κλάσμα των ορθά θετικών καταχωρήσεων προς το σύνολο των καταχωρήσεων. Ωστόσο, η τιμή της ακρίβειας δεν είναι αρκετή για να μετρήσει αποτελεσματικά την απόδοση του αλγορίθμου. Παρακάτω αναλύονται διάφορα μέτρα απόδοσης που χρησιμοποιούνται συνδυαστικά με την ακρίβεια, καθώς και η σημασία τους.

$$Accuracy = \frac{N\_correctly\_classified\_cases}{N\_cases} \qquad (7)$$

Η ακρίβεια (**Precision**), είναι η θετική προγνωστική τιμή, δηλαδή τα δεδομένα που ορθώς ταξινομήθηκαν ως θετικά (True Positives), ως προς το άθροισμα των ορθά θετικών και λανθασμένα θετικών καταχωρήσεων.

$$Precision = \frac{TP}{TP+FP} \qquad (8)$$

Η ανάκληση (**Recall**) είναι το κλάσμα των ορθά θετικών καταχωρήσεων ως προς το άθροισμα των ορθά θετικών και των λανθασμένα αρνητικών καταχωρήσεων.

$$Recall = \frac{TP}{TP+FN} \qquad (9)$$

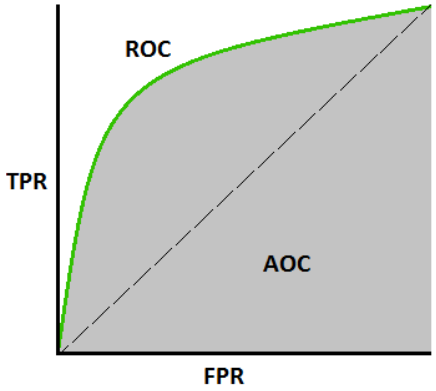
Συνήθως, η ακρίβεια και η ανάκληση δε χρησιμοποιούνται μεμονωμένα. Αντίθετα, οι τιμές για ένα μέτρο συγκρίνονται για ένα σταθερό επίπεδο στο άλλο μέτρο, ή και τα δύο συνδυάζονται σε ένα μόνο μέτρο. Χαρακτηριστικό παράδειγμα του συνδυασμού ακρίβειας και ανάκλησης είναι το **F-measure** ή **F1** (ο σταθμισμένος αρμονικός τους μέσος).

$$Fmeasure = 2\frac{Precision*Recall}{Precision+Recall} = \frac{2TP}{2TP+FP+FN} \qquad (10)$$

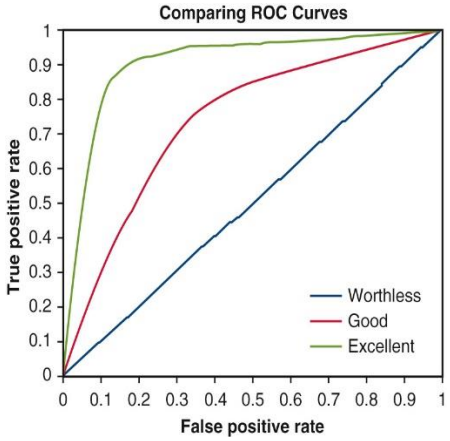
Άλλο παράδειγμα του συνδυασμού τους είναι ο συντελεστής συσχέτισης Matthews (**Matthews correlation coefficient**), που παίρνει τιμές στο [-1,1]. Η τιμή 1 σημαίνει ότι ότι οι πραγματικές κλάσεις είναι πλήρως συσχετισμένες με τις κλάσεις που προέκυψαν από την ταξινόμηση και το -1 ότι υπάρχει αρνητική συσχέτιση μεταξύ αυτών των δύο.

Ένα ακόμα μέτρο που χρησιμοποιείται για την απεικόνιση της διαγνωστικής ικανότητας ενός μοντέλου δυαδικού ταξινομητή είναι η χαρακτηριστική καμπύλη λειτουργίας, ή **καμπύλη ROC (Receiver Operating Characteristic curve**). Η καμπύλη ROC δημιουργείται σχεδιάζοντας τον πραγματικό θετικό ρυθμό (TPR), δηλαδή του μέτρου recall, έναντι του ψευδούς θετικού ρυθμού (FPR), που είναι η πιθανότητα ψευδούς συναγερμού, δηλαδή οι μή ορθές θετικές καταχωρήσεις, για διάφορες τιμές κατωφλίου. Η καμπύλη μπορεί ακόμα να θεωρηθεί ως δείκτης εκτίμησης του σφάλματος τύπου I που έχει μεγάλη σημασία για την επιτυχία του μοντέλου.

Από το γράφημα της καμπύλης ROC στην περίπτωση της δυαδικής ταξινόμησης υπολογίζεται το εμβαδόν κάτω από την καμπύλη (Area Under The Curve). Για τιμές ROC area κοντά στο 1 ο διαχωρισμός είναι επιτυχής και το μοντέλο λειτουργεί ιδανικά, ενώ για τιμές κοντά στο 0,5 το μοντέλο πραγματοποιεί ¨τυχαία¨ επιλογή.



Εικόνα 9
Αναπαράσταση καμπύλης ROC



Εικόνα 10
Σύγκριση καμπύλων ROC

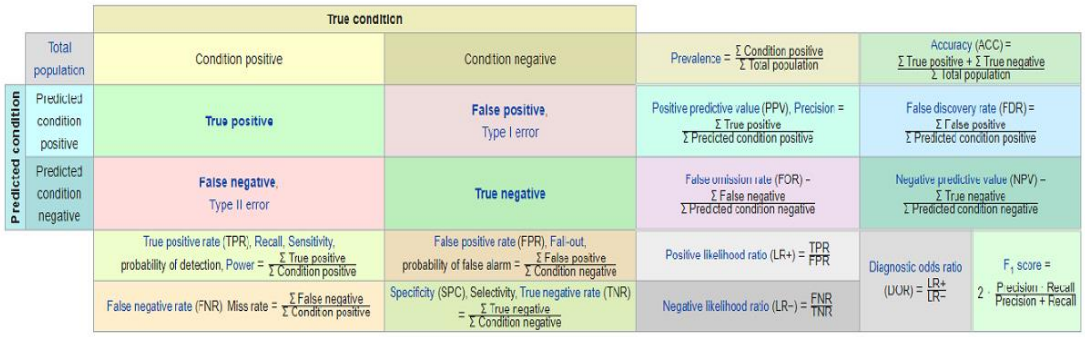
3.6.1 Confusion Matrix

Ο πίνακας σύγχησης ή τύπων στατιστικού σφάλματος (**Confusion Matrix**) περιγράφει την απόδοση ενός μοντέλου ταξινόμησης σε ένα σύνολο εκπαιδευτικών δεδομένων. Για την περίπτωση της δυαδικής ταξινόμησης, κάθε καταχώρηση αντιστοιχίζεται σε μία κλάση. Επειδή τα δεδομένα είναι ήδη γνωστά, ο πίνακας μας δείχνει τί μέρος των καταχωρήσεων έχει ταξινομηθεί σωστά στην αντίστοιχη κλάση. Αναλυτικότερα, για δύο κλάσεις, 0 η θετική και 1 η αρνητική κλάση, έχουμε:

- **True positives (TP)**: περιπτώσεις που σωστά αντιστοιχήθηκαν στη θετική κλάση
- **True negatives (TN)**: περιπτώσεις που σωστά αντιστοιχήθηκαν στην αρνητική κλάση
- **False positives (FP)**: περιπτώσεις που λανθασμένα αντιστοιχήθηκαν στη θετική κλάση
- **False negatives (FN)**: περιπτώσεις που λανθασμένα αντιστοιχήθηκαν στην αρνητική κλάση

Η μορφή του πίνακα εξυπηρετεί την πιο λεπτομερή ανάλυση της ταξινόμησης. Αν εξεταστεί απλά η ακρίβεια (Accuracy) του μοντέλου, υπάρχει ο κίνδυνος να οδηγηθούμε σε παραπλανητικά αποτελέσματα. Αν οι κλάσεις είναι σε ανισορροπία (class imbalance), δηλαδή η θετική κλάση αφορά το 95% των περιπτώσεων και μόλις το 5% την αρνητική, το μοντέλο πιθανότατα θα αγνοήσει τις αρνητικές περιπτώσεις και θα τις ταξινομήσει στη θετική κλάση. Αυτό έχει ως αποτέλεσμα πολύ υψηλή ακρίβεια και F-score για το training set, αλλά στην πραγματικότητα θα έχει αποτύχει να λύσει το πρόβλημα της ταξινόμησης. Επομένως ο Confusion Matrix αποτελεί ένα πολύ σημαντικό εργαλείο ελέγχου της απόδοσης του αλγορίθμου, καθώς ελέχγει τα ποσοστά σφάλματος I και II. Στο σφάλμα τύπου I συγκαταλέγονται οι False Positives περιπτώσεις. Είναι το πιο σοβαρό σφάλμα και πρέπει να ελέγχεται καθώς μπορεί να οδηγήσει σε σημαντικά

προβλήματα. Στο Μάρκετινγκ και στις περιπτώσεις ομαδοποίησης και ανάλυσης πελατών είναι πολύ σοβαρό λάθος η ταξινόμηση ενός ¨κακού¨ πελάτη στους ¨καλούς¨, ενώ παραδείγματος χάρη σε περιπτώσεις ανίχνευσης καρκίνου, η ταξινόμηση μιας θετικής σε καρκίνο περίπτωση στις ¨καθαρές¨ θα αγνοήσει ένα πρόβλημα υγείας που η έγκαιρη διάγνωση και θεραπεία είναι ζωτικής σημασίας.



Εικόνα 11
Confusion Matrix και μέτρα απόδοσης

3.7 Γενίκευση και θόρυβος (Generalization and Noise)

Σε ένα πρόβλημα ταξινόμησης, θέλουμε ένα μοντέλο να φτιάχνει μια περιγραφή που περιλαμβάνει όλα τα θετικά παραδείγματα, δηλαδή τα παραδείγματα που ανήκουν σε μια συγκεκριμένη κλάση, και κανένα από τα αρνητικά. Ένα σετ εκπαίδευσης (training set), αποτελεί ένα υποσύνολο της βάσης δεδομένων. Το εμπειρικό λάθος του μοντέλου, είναι το ποσοστό των εκπαιδευτικών παρατηρήσεων (training instances) που οι προβλέψεις για την κλάση δεν αντιστοιχούν στην πραγματική τους κλάση. Στην ταξινόμηση (classification), γίνεται προσπάθεια να διαχωριστούν επαρκώς τα όρια της μιας κλάσης από τις υπόλοιπες. Επομένως δημιουργείται ένα πρόβλημα Κ-κλάσεων, όπου Κ ο αριθμός των κλάσεων, άρα υπάρχουν Κ υποθέσεις που πρέπει το μοντέλο να μάθει.

Στην παλινδρόμηση (regression), όπου το αποτέλεσμα δεν είναι κλάση αλλά μια αριθμητική τιμή, θέλουμε να βρούμε μια συνάρτηση που ελαχιστοποιεί το εμπειρκό λάθος. Καθώς εξετάζονται όλο και περισσότερες παρατηρήσεις εκπαίδευσης, το μοντέλο μπορεί καλύτερα να διαχωρίζει τις αβέβαιες υποθέσεις και να τις αφαιρεί από τη διαδικασία μάθησης. Ένας συνήθης τρόπος να υπολογιστεί η απόδοση ενός παλινδρομικού μοντέλου είναι η μέθοδος της διαφοράς τετραγώνων (square of the difference).

Το ζητούμενο σε κάθε περίπτωση είναι η εξαγωγή του σωστού συμπεράσματος από ένα δεδομένου εισόδου εκτός του σετ εκπαίδευσης, η ισχυρή δηλαδή ικανότητα πρόβλεψης του αλγορίθμου. Το πόσο καλά έχει εκπαιδευτεί ένα μοντέλο στο σετ εκπαίδευσης για να πραγματοποιεί σωστή πρόβλεψη σε νέα δεδομένα ονομάζεται γενίκευση (generalization).

Ωστόσο, κατά την εκπαίδευση των αλγορίθμων συχνά παρουσιάζονται σφάλματα, τα οποία μπορεί να είναι τυχαία ή συστηματικά. Το τυχαίο σφάλμα που προκύπτει ονομάζεται θόρυβος (noise) και αφορά, είτε λανθασμένες καταχωρήσεις τιμών, είτε ακραίες τιμές στα δεδομένα. Η αντιμετώπηση του θορύβου είναι μια πολύ κρίσιμη διαδικασία για την απόδοση του μοντέλου. Καθώς εξετάζονται περισσότερα παραδείγματα, όλο και περισσότερες ασυνεπείς υποθέσεις αφαιρούνται από την υποθετική κλάση, ενώ παραμένουν οι πιο αξιόπιστες περιπτώσεις.

3.8 Bias-Variance

Κατά τη διαδικασία εκπαίδευσης των αλγορίθμων σε διαφορετικά δεδομένα προκύπτει ένα είδος σφάλματος που ονομάζεται μεροληψία (bias). Επειδή η διαδικασία μάθησης ειναι περίπλοκη και τα δεδομένα από μόνα τους δεν είναι τόσο επαρκή ώστε να βρεθεί η λύση με βεβαιότητα, γίνονται κάποιες υποθέσεις ως προς την τελική λύση. Το σύνολο των υποθέσεων που αποτελούν μέρος της διαδικασίας μάθησης καλείται επαγωγική μεροληψία (inductive bias). Μεροληψία του μοντέλου είναι η διαφορά ανάμεσα στην εκτιμώμενη πρόβλεψη από την πραγματική.

Για την καλύτερη γενίκευση του μοντέλου, είναι απαραίτητο η πολυπλοκότητα της υποθετικής τάξης να ταιριάζει με την πολυπλοκότητα της υποκείμενης συνάρτησης των δεδομένων. Η πολυπλοκότητα των δεδομένων είναι με άλλα λόγια το σφάλμα λόγω διασποράς (variance), που δείχνει πόσο απέχουν οι προβλέψεις ανάμεσα σε διαφορετικά σετ δεδομένων εκπαίδευσης για ένα συγκεκριμένο πραγματικό σημείο. Για δεδομένο σετ εκπαίδευσης, μπορούμε να βρούμε ένα υποσύνολο του σετ για το οποίο το εμπειρικό λάθος ελαχιστοποιείται. Ωστόσο, αν αυτό δεν επιλεγεί σωστά, η γενίκευση του μοντέλου δε θα είναι ικανοποιητική.

3.8.1 Bias-Variance trade off

Σε όλους τους αλγορίθμους μάθησης που εκπαιδεύονται από εκπαιδευτικά δεδομένα, υπάρχει μία ανταλλαγή-συμβιβασμός (trade-off) μεταξύ της διασποράς και της μεροληψίας. Όσο πιο απλό το μοντέλο, τόσο μεγαλύτερη είναι η μεροληψία, και όσο πιο σύνθετο είναι το μοντέλο, τόσο υψηλότερη είναι η διασπορά. Καθώς η ποσότητα λοιπόν των δεδομένων εκπαίδευσης αυξάνεται, το λάθος γενίκευσης μειώνεται. Καθώς η πολυπλοκότητα αυξάνεται, το λάθος γενίκευσης μειώνεται μέχρι ένα σημείο και μετά αυξάνεται. Αυτό μπορεί να ελεγχθεί, αυξάνοντας τον αριθμό των εκπαιδευτικών δεδομένων μέχρι ένα βαθμό.

3.9 Underfitting-Overfitting

Για να είναι ένα μοντέλο αξιόπιστο το training error πρέπει να είναι λίγο χαληλότερο από το test error. Βασικό πρόβλημα στους αλγορίθμους Μηχανικής Μάθησης που επηρεάζει την απόδοσή τους είναι οι περιπτώσεις της υπεραπλούστευσης (underfitting) και της υπερπροσαμογής (overfitting). Στην πρώτη περίπτωση εμφανίζεται υψηλό λάθος training error, σχεδόν ίδιο με το test error, που δίνει αναξιόπιστες προβλέψεις, ενώ υπάρχει και υψηλή μεροληψία, που δεν επιτρέπει την ικανοποιητική γενίκευση σε νέα δεδομένα.

Στην περίπτωση της υπερπροσαρμογής, εμφανίζεται υψηλή διασπορά και χαμηλό training error, πολύ πιο χαμηλό από το test error, καθώς το μοντέλο δε μπορεί να διαχωρίσει το θόρυβο, άρα οδηγείται σε αναξιόπιστες προβλέψεις.

Σε κάθε πρίπτωση τα φαινόμενα αυτά πρέπει να ελέγχονται και να ελαχιστοποιούνται όσο το δυνατόν περισσότερο για τη σωστή και ουσιαστική αξιοποίηση της πληροφορίας που μπορεί να αντληθεί από μια βάση δεδομένων.

Μία βασική μέθοδος εξάλειψης τέτοιων φαινομένων είναι η κανονικοποίηση (regularization). Σκοπός της είναι να αποφευγθεί η υπερπροσαρμογή σε περιπτώσεις με υψηλή διασπορά. Οι πιο συνήθεις τρόποι κανονικοποίησης είναι η LASSO (Least Absolute Shrinkage and Selection Operator), μια παλινδρομική μέθοδος που εφαρμόζει ποινή στους παλινδρομικούς συντελεστές συρρικνώνοντας τα βάρη κάποιων μεταβλητών στο μηδέν, η Ridge Regression που μειώνει επίσης τους συντελεστές και η Elastic Net. Συγκεκριμένα η LASSO πραγματοποιεί επιλογή χαρακτηριστικών (Feature Selection) και είναι ιδανική για περιπτώσεις πρόβλεψης δεδομένων.

Μία ακόμα μέθοδος αντιμετώπισης τέτοιων προβλημάτων είναι ο πολλαπλασιασμός ή ο υποπολλαπλασιασμός του δείγματος (Oversampling/Undersampling). Μέσω αυτών των μεθόδων είναι δυνατό οι κλάσεις να εξισορροπηθούν μεταξύ τους ώστε ο αλγόριθμος της ταξινόμησης να εφαρμοστεί σωστά, δίνοντας το επιθυμητό αποτέλεσμα. Ωστόσο, πρέπει να χρησιμοποιούνται με προσοχή, καθώς με τη μέθοδο Oversampling εγκιμωνεί ο κίνδυνος η βάση δεδομένων να γίνει πολύ μεγάλη αυξάνοντας σημαντικά το υπολογιστικό κόστος, ενώ αντίστοιχα στη μέθοδο Undersampling το δείγμα μπορεί να γίνει υπερβολικά μικρό, με αποτέλεσμα σημαντικό μέρος της πληροφορίας να χαθεί και να μη μπορεί να ανιχνευτεί πλέον κατά την ανάλυση. [27]

Κεφάλαιο 4: Πειραματικό μέρος της έρευνας

4.1 Εισαγωγή

Σε αυτό το κεφάλαιο παρουσιάζεται η διαδικασία που ακολουθήθηκε για την επεξεργασία των δεδομένων της βάσης πελατών που χρησιμοποιήθηκε για την παρούσα έρευνα, η επιλογή και εκπαίδευση των ταξινομητών για τη δημιουργία ενός μοντέλου πρόβλεψης αποχώρησης πελατών, καθώς επίσης το μεθοδολογικό πλαίσιο στο οποίο βασίστηκε η κατάτμηση του συνόλου των πελατών σε επιμέρους ομάδες με κοινά χαρακτηριστικά και η εξαγωγή κανόνων συσχέτισης, ενώ παρουσιάζονται αναλυτικά τα αποτελέσματα για κάθε περίπτωση.

4.2 Λογισμικό Weka

Το λογισμικό που χρησιμοποιήθηκε για τη δημιουργία του μοντέλου πρόβλεψης και ανάλυσης είναι το **Weka** (έκδοση 3.8.5). Το Weka είναι μια βιβλιοθήκη Μηχανικής Μάθησης που χρησιμοποιείται για την επίλυση πληθώρας προβλημάτων εξόρυξης δεδομένων. Επιτρέπει την υλοποίηση και εφαρμογή αλγορίθμων για εξαγωγή δεδομένων, ενώ δίνει τη δυνατότητα χρήσης αλγορίθμων από διάφορες εφαρμογές με χρήση της γλώσσσας Java. Το Weka αναπτύχθηκε στο University of Waikato της Νέας Ζηλανδίας και είναι ελεύθερο λογισμικό υπό την άδεια GNU General Public License. Περιλαμβάνει ένα σύνολο εργαλείων για προ επεξεργασία δεδομένων, ταξινόμηση, παλινδρόμηση, συσταδοποίηση, εξαγωγή κανόνων συσχέτισης, επιλογή χαρακτηριστικών και οπτικοποίησης δεδομένων.

4.3 Στατιστικά του dataset

Η βάση δεδομένων που θα χρησιμοποιηθεί για την έρευνα προέρχεται από την Kaggle, μια διαδικτυακή κοινότητα επιστημόνων δεδομένων και επαγγελματιών μηχανικής μάθησης που προσφέρει μια δημόσια πλατφόρμα δεδομένων.

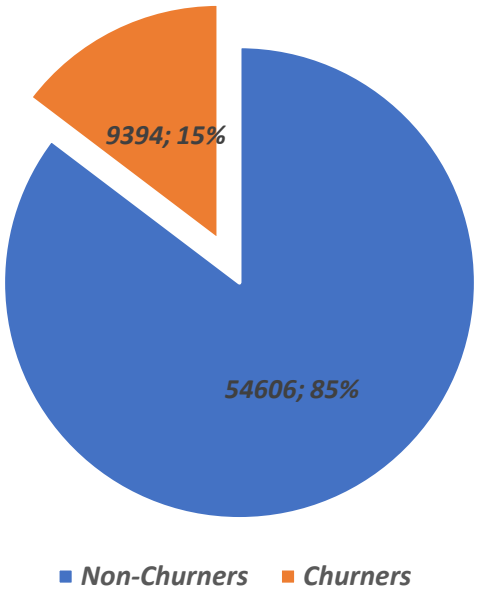
Η συγκεκριμένη βάση αποτελείται από υποθετικά δεδομένα διατήρησης πελατών μιας εταιρείας (δημογραφικά και οικονομικά) καθώς και από στοιχεία που αφορούν τις μεθόδους μάρκετινγκ που εφαρμόστηκαν από την εταιρεία προς τους πελάτες και τις αντίστοιχες επιλογές των πελατών σε αυτές τις προσφορές. Ακόμα, αποτυπώνεται η ικανοποίηση ή μη του κάθε πελάτη ως προς την αντιμετώπιση που είχε από την εταιρεία, με το αν συνεχίζει να ανήκει στο πελατολόγιο της ή όχι. Τα δεδομένα αυτά θα χρησιμοποιηθούν για την πρόβλεψη αποχώρησης πελατών της εταιρείας. (Marketing Promotion Campaign – Customer Retention data for Churn Prediction: https://www.kaggle.com/davinwijaya/customer-retention)

Τα χαρακτηριστικά της βάσης δεδομένων παρουσιάζονται παρακάτω:

- Recency: Το διάστημα (σε μήνες) που μεσολαβεί από την τελευταία συναλλαγή του πελάτη με την εταιρεία
- History: Η συνολική αξία σε U.S. dollars των παρελθοντικών συναλλαγών του πελάτη με την εταιρεία
- Used discount: Δείχνει αν ο πελάτης έχει χρησιμοποιήσει έκπτωση στο παρελθόν
- Used bogo: Δείχνει αν ο πελάτης έχει χρησιμοποιήσει μια προσφορά ένα συν ένα στο παρελθόν
- Zip code: Τόπος διαμονής του πελάτη. Διακρίνεται σε Suburban (προάστειο), Urban (αστική περιοχή), Rural (αγροτική περιοχή)
- Is referral: αναφέρει ποιοί πελάτες αποκτήθηκαν μέσω προωθητικών μηνυμάτων
- Channel: αναφέρεται στο είδος της υπηρεσίας που απολαμβάνει κάθε πελάτης για την επικοινωνία με την εταιρεία και μπορεί να είναι Web, Phone, ή Multichannel
- Offer: η προσφορά που έχει σταλεί σε κάθε πελάτη και μπορεί να είναι Discount, Buy One Get One (bogo) ή καθόλου προσφορά (No Offer)
- Conversion: διατήρηση πελάτη (αν συνεχίζει να πραγματοποιεί αγορές από την εταιρεία ή όχι)

Λάβαμε ένα σύνολο δεδομένων σε μορφή csv, το οποίο περιέχει 64.000 instances, 9 χαρακτηριστικών. Δεν υπάρχουν ελλιπή δεδομένα στο data set.

Η κλάση-στόχος είναι το ***conversion***, η οποία έχει μόνο δύο ετικέτες: 0 και 1. Πρόκειται λοιπόν για πρόβλημα δυαδικής ταξινόμησης (***binary classification problem***). Από τις 64.000 καταχωρήσεις πελατών, οι 54.606 παραμένουν πελάτες της εταιρείας, συνεχίζοντας τις αγορές τους, και μόνο 9.394 αποχωρούν από αυτή. Πρόκειται λοιπόν για ένα μή ισορροπημένο (***imbalanced***) σύνολο δεδομένων, καθώς το 85% των δεδομένων προέρχονται από την κλάση 0 (***non churners***), και μόλις το 15% από την κλάση 1 (***churners***), που είναι και η μειονοτική κλάση.



Γράφημα 1
Ποσοστό αποχώρησης πελατών από την εταιρεία

Η μεταβλητή ***recency*** παρουσιάζει το χρονικό διάστημα που μεσολαβεί από την τελευταία αγορά του πελάτη μέχρι σήμερα. 8.952 πελάτες (14%) πραγματοποίησαν αγορές τον τελευταίο μήνα και 22.393 (35%) το τελευταίο τρίμηνο, ενώ 2.332 πελάτες (3,6%) δεν πραγματοποιήσαν αγορά τους τελευταίους δώδεκα μήνες και 27.415 (43%) δεν πραγματοποίησαν αγορές το τελευταίο εξάμηνο. Ο μέσος όρος της μεταβλητής είναι 5 μήνες.

Η μεταβλητή ***history*** αναφέρεται στο συνολικό ποσό που έχει ξοδέψει κάθε πελάτης για τις αγορές του. Από το σύνολο των 64.000 πελατών, οι 50.173 έχουν ξοδέψει μέχρι 360$. Συγκεκριμένα, 36.741 πελάτες ξόδεψαν μέχρι 200$ και 13.432 από 200$ μέχρι 360$. Οι πελάτες αυτοί αποτελούν την πλειοψηφία του δείγματος καθώς ξεπερνούν το 78%, καθιστώντας τη μεταβλητή ανομοιόμορφα κατανεμημένη με ακραίες τιμές για τα υψηλότερα ποσά. Οι πελάτες που έχουν ξοδέψει σε αγορές πάνω από 1.000$ είναι 1.210, ενώ μόνο 42 έχουν ξοδέψει περισσότερα από 2.000$.

Η μεταβλητή ***used discount*** είναι μοιρασμένη καθώς το 45% των πελατών έχει λάβει έκπτωση στο παρελθόν και το 55% όχι. Ακόμα, το 45% των πελατών έχει λάβει προσφορά ένα συν ένα (***used bogo***).

Η προσφορά που έχει σταλεί σε κάθε πελάτη ως κίνητρο να συνεχίσει τις αγορές του στην εταιρεία περιγράφεται από τη μεταβλητή ***offer***. 33% των πελατών έχουν λάβει έκπτωση, 33% προσφορά ένα συν ένα, ενώ 33% δεν έλαβε καθόλου προσφορά.

Ο τρόπος επικοινωνίας της εταιρείας με τους πελάτες της περιγράφεται από τη μεταβλητή ***channel***. Η εταιρεία επικοινωνεί με το 44% των πελατών της μέσω

τηλεφώνου, και σε αντίστοιχο ποσοστό μέσω διαδικτύου, ενώ στο 12% των πελατών η επικοινωνία γίνεται και με τους δύο τρόπους.

Από τις 64.000 καταχωρήσεις πελατών, οι 31.856 (49,5%) έμαθαν για την εταιρεία και ξεκίνησαν να πραγματοποιούν αγορές σε αυτή μέσα από προωθητικά μηνύματα και διαφημίσεις, όπως περιγράφει η μεταβλητή ***referral***.

Οι πελάτες της εταιρείας κατοικούν σε αστικές περιοχές σε ποσοστό 40%, σε προάστεια σε ποσοστό 45% και σε αγροτικές περιοχές σε ποσοστό 15%. Τα ποσοστά αυτά περιγράφονται με τη μεταβλητή ***zip code***.

4.4 Διαδικασία υλοποίησης της έρευνας

Όπως αναλύθηκε προηγουμένως, η παρούσα έρευνα αφορά την αντιμετώπιση ενός ***Class Imbalanced Problem (CIP)*** για την πρόβλεψη αποχώρησης πελατών (***Customer Churn***). Οι σύγχρονες τεχνικές Μηχανικής Μάθησης έχουν χαμηλότερη απόδοση καθώς εφαρμόζονται σε datasets με μεγάλη απόκλιση στις τάξεις. Ωστόσο, μέσα από την λεπτομερή ανάλυση των βάσεων δεδομένων, μπορούν να αντληθούν σημαντικές γνώσεις για τις εταιρείες.

Τα προβλήματα άνισης κατανομής κλάσης προσεγγίζονται γενικά με δύο τρόπους:

1. Δίνοντας βάση στη σημαντικότητα της μειονοτικής κλάσης (***cost sensitive approach***)
2. Προσθέτοντας ή αφαιρώντας δεδομένα κατά την προεπεξεργασία όπου γίνεται ανακατανομή της τάξης ώστε να μειωθεί η επιρροή της άνισης κατανομής τάξης στο πρόβλημα ταξινόμησης (***Oversampling/Undersampling***)

Η εφαρμογή τέτοιων μεθόδων πρέπει να γίνεται με προσοχή, καθώς όταν αφαιρούνται δεδομένα της βασικής κλάσης ή αυξάνονται τα δεδομένα της μειονοτικής κλάσης μπορεί να χαθούν κρίσιμα δεδομένα, ενώ η μεγάλη αύξηση των δεδομένων κάνει τον αλγόριθμο να τρέχει πιο αργά.

Στην παρούσα έρευνα, αρχικά εφαρμόστηκαν οι αλγόριθμοι επιτηρούμενης μάθησης στο σύνολο της βάσης δεδομένων. Τα αποτελέσματα έδειξαν απόλυτη υπερπροσαρμογή των αλγορίθμων στην κύρια κλάση (***Overfitting***), αγνοώντας ολοκληρωτικά τη μειονοτική κλάση, συνεπώς καθιστώντας αποτυχημένη την όποια προσπάθεια πρόβλεψης του μοντέλου ως προς ενδεχόμενη αποχώρηση των πελατών.

Σε δεύτερο στάδιο, κατά την προ επεξεργασία των δεδομένων, χρησιμοποιήθηκε η τεχνική ανακατανομής κλάσεων του δείγματος (***resample***), με ταυτόχρονη μείωση του συνόλου των δεδομένων της βάσης (***Undersampling***). Αυτό είχε ως αποτέλεσμα τη μεγάλη βελτίωση τόσο της απόδοσης και της σωστής προσαρμογής, όσο και της ταχύτητας υλοποίησης των αλγορίθμων.

Ακόμα, εφαρμόστηκε η μέθοδος εξισορρόπησης κλάσης (***Class Balancer***), κατά την οποία αυξήθηκαν τα δεδομένα της μειονοτικής κλάσης με ταυτόχρονη μείωση των δεδομένων της κύριας κλάσης. Ωστόσο η εφαρμογή της συγκεκριμένης μεθόδου δεν είχε ικανοποιητικό αποτέλεσμα, συγκριτικά με τη μέθοδο resample.

Συνήθως, οι βάσεις που χρησιμοποιούνται για εξόρυξη δεδομένων είναι πολύ μεγάλες για απευθείας αναλύσεις χωρίς επεξεργασία. Η μείωση γίνεται εκτός από το συνολικό δείγμα και στις μεταβλητές – χαρακτηριστικά, στο στάδιο προ επεξεργασίας των δεδομένων, με τη διαδικασία επιλογής χαρακτηριστικών (***feature selection***).

Το λογισμικό Weka δίνει μια πληθώρα δυνατοτήτων για επιλογή χαρακτηριστικών. Ο συνηθέστερος τρόπος επιλογής χαρακτηριστικών έχει να κάνει με την κατάταξη (***ranking***) των μεταβλητών ως προς την συσχέτιση (***correlation***) που εμφανίζουν με τη μεταβλητή κλάσης (***filter-based feature selection method***). Επιλέγονται οι μεταβλητές που επηρεάζουν σε μεγαλύτερο ποσοστό το ζητούμενο αποτέλεσμα, στην προκειμένη περίπτωση το ποσοστό αποχώρησης πελατών (***churn rate***), γίνεται δηλαδή επιλογή χαρακτηριστικών που σχετίζονται καλύτερα με τη μεταβλητή κλάσης και μπορούν να μειώσουν τον υπολογιστικό φόρτο του αλγορίθμου πετυχαίνοντας καλύτερα αποτελέσματα. Από τις διαθέσιμες επιλογές του Weka για επιλογή χαρακτηριστικών, στο στάδιο της προ επεξεργασίας δεδομένων, πιο αποδοτική αποδείχθηκε η μέθοδος ***Correlation-based Feature Subset Selection.*** Άλλες μέθοδοι επιλογής χαρακτηριστικών που χρησιμοποιήθηκαν δίνοντας παρόμοια αποτελέσματα ήταν η ***Classifier Attribute Evaluation***, ***Correlation Attribute Evaluation*** και ***Information Gain Attribute Evaluation***, ενώ η μέθοδος διαστατικής μείωσης με την εφαρμογή ανάλυσης κυρίων συνιστωσών (***Principal Components Analysis***) δεν απέδωσε ικανοποιητικά στο συγκεκριμένο data set, καθώς τα μοντέλα πρόβλεψης που προέκυψαν δεν κατάφεραν να επιτύχουν υψηλές αποδόσεις πρόβλεψης και προσαρμογής στα δεδομένα.

Η συγκεκριμένη μεθοδολογία που εφαρμόστηκε για την γενικότερη επεξεργασία των δεδομένων καλείται ***external (data level) approach*** και έχει τα εξής χαρακτηριστικά:

- Δεν απαιτεί εκπαίδευση του ταξινομητή
- Χρησιμοποιείται στο στάδιο της προ επεξεργασίας δεδομένων, επομένως δεν επηρεάζει τον αλγόριθμο που θα χρησιμοποιηθεί

Το επόμενο στάδιο περιλαμβάνει την κατασκευή του μοντέλου ταξινόμησης με χρήση των αλγορίθμων επιτηρούμενης μάθησης που έχουν επιλεγεί. Για καλύτερο έλγχο της απόδοσης του μοντέλου και ελαχιστοποίησης της μεροληψίας, χρησιμοποιείται η στατιστική μέθοδος ***10-fold cross validation***. Το πιο ικανό μοντέλο πρόβλεψης αποχώρησης πελατών προκύπτει με χρήση του αλγορίθμου ***Random Forest***, καθώς πετυχαίνει ακρίβεια ***F-measure= 79,1%***, με ταυτόχρονη σωστή κατανομή των σωστών και λανθασμένων καταχωρήσεων (***True Positive Rate***

για την κύρια κλάση 0,838 και για την μειονοτική κλάση 0,691), ενώ και η καμπύλη ROC εμφανίζει πολύ ικανοποιητικό αποτέλεσμα (***ROC Area = 0,803***).

Με την ολοκλήρωση της ταξινόμησης και της επιτυχής κατασκευής και εκπαίδευσης του μοντέλου, περνάμε στον επόμενο στόχο της έρευνας, που είναι η κατάτμηση των πελατών (***clustering***) και η εξαγωγή κανόνων συσχέτισης (***Association Rules***). Τα αποτελέσματα αυτών των διαδικασιών και η μελέτη τους, θα οδηγήσουν στην λήψη αποφάσεων για την καλύτερη δυνατή αντιμετώπιση των πελατών από την εταιρεία μέσα από τη διαδικασία στοχευμένου μάρκετινγκ που πρόκειται να εφαρμοστεί.

Αρχικά χρησιμοποιήθηκε το σύνολο των δεδομένων για τη δημιουργία του μοντέλου ταξινόμησης πελατών. Στη συνέχεια γίνεται διαχωρισμός των πελατών σε δύο επιμέρους βάσεις δεδομένων. Η μία περιλαμβάνει τους πελάτες που αποχώρησαν (***churners***), ενώ η άλλη αυτούς που συνεχίζουν να ανήκουν στο πελατολόγιο της εταιρείας (***non churners***). Αυτό γίνεται γιατί στόχος είναι να βρεθούν συγκεκριμένες υπο ομάδες πελατών κάθε κατηγορίας τους οποίους θα στοχεύσει η εταιρεία, καθώς επίσης συγκεκριμένα μοτίβα και συνήθειες που θα δώσουν μια ξεκάθαρη εικόνα για τις μέχρι τώρα μεθόδους μάρκετινγκ που εφαρμόστηκαν αλλά και να ερευνηθούν νέες πιο αποδοτικές. Η ομαδοποίηση και η διαδικασία εξαγωγής κανόνων συσχέτισης εφαρμόζεται στις δύο επιμέρους βάσεις δεδομένων ξεχωριστά.

Η συσταδοποίηση (***clustering***) εφαρμόζεται σε μή ταξινομημένα δεδομένα (***unlabeled data***). Αγνοούμε λοιπόν το χαρακτηριστικό που προσδιορίζει την κλάση (conversion) και εφαρμόζουμε ομαδοποίηση με χρήση του αλορίθμου ***K-means***. Όσο μικρότερη είναι η τιμή του ***sum of squared error***, τόσο καλύτερη είναι η ομαδοποίηση που πραγματοποιεί ο αλγόριθμος (***elbow method***).

Αρχικά, επιλέγουμε δύο συστάδες για ομαδοποίηση (number of clusters = 2), και στη συνέχεια αυξάνουμε σταδιακά τον αριθμό συστάδων, μέχρι να βρεθεί ο ιδανικός αριθμός συστάδων που δίνει την καλύτερη και πιο ουσιαστική ομαδοποίηση. Το υψηλό σφάλμα ελαχίστων τετραγώνων που εμφανίζει ο αλγόριθμος αρχικά θα μπορούσε να οφείλεται σε κάποια ακραία σημεία (outliers) στα δεδομένα των clusters. Με χρήση κατάλληλων φίλτρων του λογισμικού Weka, καταφέρνουμε να μειώσουμε το σφάλμα με τον εντοπισμό και την αφαίρεση των ακραίων τιμών.

Παρατηρούμε ό,τι για μεγαλύτερο αριθμό συστάδων (k=20,k=30), η τιμή του σφάλματος μειώνεται. Ωστόσο, η τμηματοποίηση των πελατών σε τόσες ομάδες δεν έχει πρακτική σημασία καθώς δημιουργεί πολλές και ολιγομελείς ομάδες πελατών που δεν αξίζει να αναληθούν περαιτέρω στα πλαίσια του στοχευμένου μάρκετινγκ.

Η διαδικασία εξαγωγής κανόνων συσχέτισης (***Association Rules Extraction***) εφαρμόζεται επίσης σε μη ταξινομημένα δεδομένα. Αγνοώντας τη μεταβλητή κλάσης εφαρμόζουμε τον αλγόριθμο ***Apriori*** στις δύο νέες βάσεις δεδομένουν που έχουν προκύψει από το διαχωρισμό του αρχικού data set. Ο αλγόριθμος

διαχειρίζεται ονομαστικά δεδομένα (***nominal***). Στόχος μας είναι για μια δεδομένη ελάχιστη τιμή ***confidence***, να εντοπίσουμε κανόνες με τη μεγαλύτερη τιμή ***support***. Ιδανικά η τιμή ***confidence*** θα πρέπει να προσεγγίζει τη μονάδα. Τελικά θα κρατήσουμε τους πιο ισχυρούς κανόνες που θα προκύψουν. Για τιμές ***lift*** μεγαλύτερες της μονάδας, τα χαρακτηριστικά που αποτελούν τον κανόνα έχουν σημαντική συσχέτιση μεταξύ τους, ενώ για τιμές ***lift*** μικρότερες τις μονάδας, τα χαρακτηριστικά είναι μάλλον ανεξάρτητα μεταξύ τους.

4.5 Αποτελέσματα

4.5.1 Αποτελέσματα Classification

Στην αρχή οι αλγόριθμοι επιτηρούμενης μάθησης εφαρμόστηκαν χωρίς να υπάρξει οποιαδήποτε επεξεργασία και προετοιμασία των δεδομένων. Αυτό είχε ως αποτέλεσμα την υπερπροσαρμογή των αλγορίθμων στα δεδομένα, αδυνατώντας να προχωρήσουν σε αξιόπιστες προβλέψεις. Το κύριο πρόβλημα ήταν η αδυναμία των μοντέλων να εντοπίσουν τη μειονοτική τάξη και να κατανήμουν σωστά τους πελάτες στην αντίστοιχη κλάση. Είχαμε λοιπόν πολύ υψηλή τιμή ***σφάλματος τύπου I*** (καταχωρήσεις πελατών λανθασμένα ως θετικές – false positive). Αυτό αποτελεί το πιο σύνηθες πρόβλημα για μη ισορροπημένες βάσεις δεδομένων. Η υψηλή ακρίβεια που εμφανίζει το μοντέλο είναι παραπλανητική, καθώς η πλειοψηφία των περιπτώσεων ταξινομείται λάνθασμένα στην κύρια κλάση, καθιστώντας αναξιόπιστο το μοντέλο πρόβλεψης.

0	*1*	*Classified as*
True Positive (TP)	*False Positive (FP)* *Type I Error*	*0 = non churners*
False Negative (FN) *Type II Error*	*True Negative (TN)*	*1 = churners*

Πίνακας 1
Confusion Matrix

Churners	*9.394*	*100%*
Naïve Bayes churners	*208*	*2%*
kNN churners	*829*	*8%*
Logistic Regression churners	*9*	*-*
Random Forest churners	*1325*	*14%*

Πίνακας 2
Ποσοστό ταξινόμησης πελατών στη μειονοτική κλάση (TN)

	TP Rate (0)	*TN Rate (1)*	*F-Measure*	*ROC Area*
Naïve Bayes	*0,988*	*0,022*	*0,788*	*0,635*
kNN	*0,942*	*0,088*	*0,784*	*0,547*
Logistic Regression	*1*	*0*	*0,786*	*0,647*
Random Forest	*0,894*	*0,141*	*0,771*	*0,566*

Πίνακας 3
Απόδοση αλγορίθμων ταξινόμησης σε μη επεξεργασμένα δεδομένα

Η αδυναμία αναγνώρισης της μειονοτικής κλάσης και ταξινόμησης των πελατών σε αυτή αντιμετωπίστηκε με κατάλληλη επεξεργασία των δεδομένων της βάσης. Αρχικά έγινε επιλογή χαρακτηριστικών (***feature selection***), σε μια προσπάθεια να ξεχωρίσουν οι μεταβλητές που έχουν μεγαλύτερη συσχέτιση με τη μεταβλητή κλάσης. Το αποτέλεσμα της επιλογής χαρακτηριστικών ήταν η αφαίρεση των μεταβλητών ***Used discount, Used bogo*** και ***Channel***. Στη συνέχεια, εφαρμόστηκαν μέθοδοι ισορρόπησης των κλάσεων με ανακατανομή του δείγματος (***Resample***) και ταυτόχρονη μείωση του συνολικού αριθμού καταχωρήσεων (***Undersampling***), βελτιώνοντας την απόδοση και μειώνοντας το υπολογιστικό κόστος των αλγορίθμων ταξινόμησης.

Χαρακτηριστικά
1. Recency
2. History
3.Zip code
4.Referral
5.Offer
6.Conversion (Μεταβλητή κλάσης)

Πίνακας 4
Χαρακτηριστικά/Μεταβλητές που επελέγησαν

	TP Rate (0)	*TN Rate (1)*	*F-Measure*	*ROC Area*
Naïve Bayes	*0,990*	*0,017*	*0,917*	*0,632*
kNN	*0,920*	***0,439***	*0,846*	*0,696*
Logistic Regression	*1*	*0*	*0,786*	*0,635*
Random Forest	*0,926*	***0,434***	*0,849*	*0,726*

Πίνακας 5
Απόδοση αλγορίθμων ταξινόμησης για επεξεργασμένα δεδομένα στο 50% του δείγματος

Παρατηρούμε πως μετά την επιλογή χαρακτηριστικών και τη μείωση του δείγματος, οι αλγόριθμοι ***kNN*** για k=3 και ***Random Forest*** αποδίδουν πολύ καλύτερα, αναγνωρίζοντας σε μεγαλύτερο ποσοστό τη μειονοτική κλάση και ταξινομώντας σωστά περισσότερους πελάτες σε αυτή. Τα ποσοστά ***TN Rate*** της μειονοτικής κλάσης για τους ***kNN (0,439)*** και ***Random Forest (0,434)*** εμφανίζονται εμφανώς βελτιωμένα σε σχέση με τα ποσοστά του *πίνακα 3*.

Στη συνέχεια γίνεται περαιτέρω μείωση του δείγματος (κατά 75%), με ταυτόχρονη μερική ανακατανομή των κλάσεων. Γίνεται μερική εξισορρόπηση κλάσεων καθώς αυξάνεται η μεροληψία (bias) ως προς την κλάση 1 (churners), εκπαιδεύοντας τον αλγόριθμο στο να αναγνωρίζει πιο αποδοτικά τη μειονοτική κλάση στο δείγμα. Η προσέγγιση αυτή έχει σαν αποτέλεσμα την περαιτέρω βελτίωση της απόδοσης των αλγορίθμων ταξινόμησης, ιδιαίτερα των ***kNN*** και ***Random Forest***, καθώς το ***TN Rate*** της μειονοτικής κλάσης για τους αλγορίθμους αυτούς συνεχίζει να παρουσιάζει βελτίωση (***0,563*** και ***0,564*** αντίστοιχα).

	TP Rate (0)	*TN Rate (1)*	*F-Measure*	*ROC Area*
Naïve Bayes	*0,955*	*0,078*	*0,583*	*0,623*
kNN	*0,783*	***0,563***	*0,712*	*0,691*
Logistic Regression	*0,966*	*0,066*	*0,580*	*0,625*
Random Forest	*0,784*	***0,564***	*0,714*	*0,714*

Πίνακας 6

Απόδοση αλγορίθμων ταξινόμησης για επεξεργασμένα δεδομένα στο 25% του δείγματος με μερική ανακατανομή κλάσεων

Σε συνέχεια της παραπάνω μεθοδολογίας και πραγματοποιώντας διαφορετικούς συνδυασμούς χαρακτηριστικών, μεγέθους δείγματος και ποσοστού ανακατανομής τάξεων και πάντα με την εφαρμογή της μεθόδου ***10-fold cross validation***, ο ***πίνακας 7*** παρουσιάζει τα βέλτιστα δυνατά αποτελέσματα για την ταξινόμηση των πελατών. Παρατηρούμε αρχικά πως παρά τις όποιες τροποποιήσεις που έγιναν, δε βελτιώθηκε η απόδοση και η αξιοπιστία των αλγορίθμων ***Naïve Bayes*** και ***Logistic Regression***, καθώς δεν κατέστη δυνατό να αναγνωρίσουν τη μειονοτική κλάση σε βαθμό που να επιτρέπει την αποτελεσματική πρόβλεψη πελατών που θα αποχωρήσουν από την εταιρεία (churners).

Οι αλγόριθμοι ***kNN*** και ***Random Forest*** ωστόσο, δημιουργούν ένα πολύ ανταγωνιστικό και αποδοτικό μοντέλο πρόβλεψης, εμφανίζοντας υψηλή απόδοση (***F-Measure***), πολύ καλή προσαρμογή στα δεδομένα, με ικανότητα επιτυχούς αναγνώρισης των κλάσεων και ταξινόμησης των πελατών σε αυτές, ***ελαχιστοποιώντας το σφάλμα τύπου I***, καθώς και υψηλά ποσοστά ***ROC Area***, της τάξης του 80%, που είναι μια ιδιαίτερα ικανοποιητική τιμή για περιπτώσεις ταξινόμησης.

	TP Rate (0)	*TN Rate (1)*	*F-Measure*	*ROC Area*
Naïve Bayes	*0,945*	*0,105*	*0,594*	*0,635*
kNN	***0,837***	***0,690***	***0,790***	***0,787***
Logistic Regression	*0,954*	*0,046*	*0,589*	*0,639*
Random Forest	***0,838***	***0,691***	***0,791***	***0,803***

Πίνακας 7

Απόδοση αλγορίθμων ταξινόμησης για επεξεργασμένα δεδομένα στο 50% του δείγματος με μερική ανακατανομή κλάσεων

Ο αλγόριθμος ***Random Forest*** με χρήση της μεθόδου ***10-fold cross validation*** αποδίδει καλύτερα από τους άλλους αλγορίθμους ταξινόμησης που εφαρμόστηκαν στο συγκεκριμένο data set. Στους ***πίνακες 8*** και ***9*** παρουσιάζονται τα αποτελέσματα εφαρμογής του ***Random Forest*** για διαφορετικές προσεγγίσεις ανακατανομής των κλάσεων και για διαφορετικές μεθόδους επιλογής χαρακτηριστικών για το 50% του συνολικού δείγματος. Παρατηρούμε ότι η εφαρμογή του ***Random Forest*** για ανακατανομή κλάσεων κατά 50% και με επιλογή χαρακτηριστικών μέσω της

μεθόδου Correlation-based Feature Subset Selection (CfsSubsetEval) αποδίδει καλύτερα σε σχέση με τις άλλες μεθόδους επεξεργασίας δεδομένων.

Random Forest	*TP Rate (0)*	*TN Rate (1)*	*F-Measure*	*ROC Area*
Χωρίς ανακατανομή	*0,926*	*0,434*	*0,849*	*0,726*
Ανακατανομή κατά 50%	***0,838***	***0,691***	***0,791***	***0,803***
Class balancer (Weka)	*0,569*	*0,606*	*0,587*	*0,626*
Spread subsample (Weka)	0.866	0,242	0,629	0,619

Πίνακας 8

Εφαρμογή Random Forest στο dataset για διαφορετικές προσεγγίσεις ανακατανομής κλάσεων στο 50% του συνολικού δείγματος

Random Forest	*TP Rate (0)*	*TN Rate (1)*	*F-Measure*	*ROC Area*
Cfs subset eval.	***0,838***	***0,691***	***0,791***	***0,803***
Correlation attr. Eval.	0,837	0,654	0,780	0,780
Gain ratio attr. Eval.	*0,839*	*0,650*	*0,778*	*0,776*

Πίνακας 9

Αποτελέσματα της εφαρμογής του αλγορίθμου Random Forest για διαφορετικές μεθόδους επιλογής χαρακτηριστικών στο 50% του δείγματος

Ο ***πίνακας 10*** παρουσιάζει τα αποτελέσματα εφαρμογής του αλγορίθμου ***Random Forest*** για τις μεταβλητές που προέκυψαν από την ανάλυση κυρίων συνιστωσών (***Principal Components Analysis***). Αρχικά ο αλγόριθμος εφαρμόζεται στο σύνολο του δείγματος, ενώ στη δεύτερη και τρίτη περίπτωση για μερική ανακατομή κλάσεων στο 50% του δείγματος. Στις πρώτες δύο περιπτώσεις χρησιμοποιήθηκαν οι 6 μεταβλητές με την υψηλότερη βαθμολογία από αυτές που προέκυψαν από τη διαδικασία διαστατικής μείωσης, ενώ στην τρίτη περίπτωση χρησιμοποιήθηκαν οι 4 καλύτερες. Παρατηρούμε πως τα αποτελέσματα του μοντέλου πρόβλεψης, παρά την βελτίωση των αποτελεσμάτων με την ανακατανομή του δείγματος αλλά και των κλάσεων, δεν καθιστούν το μοντέλο ικανό και αξιόπιστο, καθώς αδυνατούν να εντοπίσουν και να ταξινομήσουν σωστά τη μειονοτική κλάση. Αυτό επιβεβαιώνεται και από τη μελέτη του πίνακα συσχέτισης των μεταβλητών. Δεν εντοπίζεται κάποια σημαντική συσχέτιση μεταξύ των μεταβλητών, επομένως οι μεταβλητές που προκύπτουν από την ανάλυση κυρίων συνιστωσών δε βελτιώνουν την ικανότητα πρόβλεψης, παρά μόνο το υπολογιστικό κόστος του αλγορίθμου.

Random Forest-PCA	*TP Rate (0)*	*TN Rate (1)*	*F-Measure*	*ROC Area*
No filter-6 attributes	*0,992*	*0,018*	*0,918*	*0,609*
Filter 50%-6 attributes	0,906	0,225	0,640	0,664
Filter 50%-4 attributes	*0,908*	*0,226*	*0,642*	*0,665*

Πίνακας 10

Αποτελέσματα της εφαρμογής του αλγορίθμου Random Forest με χρήση της μεθόδου ανάλυσης κυρίων συνιστωσών

1													
-0.25	*1*												
-0.03	*0.11*	*1*											
-0.03	*0.11*	***-0.82***	*1*										
0.01	*-0.01*	*0*	*-0.38*	*1*									
0	*0*	*0*	***-0.74***	*-0.34*	*1*								
0.01	*0*	*0*	*0*	*0*	*0*	*1*							
0.05	*0*	*0*	*0*	*0*	*0*	*0*	*1*						
0.03	***0.22***	*0.02*	*0.01*	*0*	*-0.01*	*0*	*-0.02*	*1*					
0.04	*-0.14*	*-0.02*	*0*	*0*	*0.01*	*-0.01*	*-0.03*	***-0.78***	*1*				
-0.11	*-0.13*	*0.06*	*0*	*0*	*0*	*0.02*	*0.07*	*-0.33*	*-0.33*	*1*			
0	***0.4***	*0*	*0*	*0*	*0*	*0*	*0*	*0.01*	*-0.01*	*0*	*1*		
0	*0*	*0*	*0.01*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*-0.5*	*1*	
0	*0*	*0*	*0*	*0.01*	*0*	*0*	*0*	*-0.01*	*0.01*	*0*	*-0.5*	*-0,5*	*1*

Πίνακας 11

Correlation Matrix

1.Recency	*8.Is referral*
2.History	*9.Channel=Phone*
3.Used discount	*10.Channel=Web*
4.Used bogo	*11.Channel=Multichannel*
5.Zip=Suburban	*12.Offer=Buy One Get One*
6.Zip=Rural	*13.Offer=No offer*
7.Zip=Urban	*14.Offer=Discount*

Πίνακας 12

Μεταβλητές πίνακα συσχέτισης

4.5.2 Αποτελέσματα Clustering

Το σύνολο των δεδομένων χωρίζεται σε δύο επιμέρους βάσεις. Η πρώτη βάση αποτελείται από εκείνους που αποχώρησαν (***churners***) από την εταιρεία, ενώ η δεύτερη βάση αποτελείται από πελάτες που παραμένουν στην εταιρεία (***non churners***). Ο αλγόριθμος ***k-means*** εφαρμόζεται και στις δύο βάσεις δεδομένων με σκοπό, στην πρώτη περίπτωση, την αποτύπωση του προφίλ των πελατών που αποχωρούν με βεβαιότητα από την εταιρεία, ενώ στη δεύτερη περίπτωση, την αποτύπωση όσο πιο ικανοποιητικά γίνεται του προφίλ των πελατών που συνεχίζουν στην εταιρεία και είναι ικανοποιημένοι από το μείγμα μάρκετινγκ που εφαρμόζεται.

Οι πελάτες που αποχώρησαν από την εταιρεία ανέρχονται σε ***9.394***. Ο αλγόριθμος *k-means* εφαρμόζεται σε αυτά τα δεδομένα, όπως ακριβώς και στη βάση δεδομένων των πελατών που παραμένουν στην εταιρεία. Αρχικά, ο *k-means* εφαρμόζεται σε μη επεξεργασμένα δεδομένα. Στη συνέχεια, εφαρμόζεται επιλογή χαρακτηριστικών ***feature selection*** με κατάταξη (*ranking*) ως προς τη συσχέτιση (*correlation*) μεταξύ των μεταβλητών και όχι ως προς τη μεταβλητή κλάσης (αφαίρεση ***Channel*** και ***Used discount***) , ενώ τέλος γίνεται περαιτέρω επεξεργασία στα δεδομένα με αφαίρεση των ακραίων τιμών από το δείγμα για αύξηση της ποιότητας συσταδοποίησης και ελαχιστοποίηση του σφάλματος ελαχίστων τετραγώνων (***sum of squared error***).

Ο ***πίνακας 13*** παρουσιάζει ενδεικτικά το αποτέλεσμα της συσταδοποίησης για αριθμό συστάδων ***k=5*** και ***k=10***, για την αντίστοιχη μέθοδο επεξεργασίας των δεδομένων που εφαρμόζεται σε κάθε περίπτωση. Παρατηρούμε ότι η τιμή του ***sum of squared error*** μειώνεται καθώς αυξάνεται ο αριθμός των επιλεγμένων συστάδων. Επίσης,η τιμή του σφάλματος μειώνεται όταν έχει προηγηθεί το στάδιο της επιλογής χαρακτηριστικών, αλλά και στην περίπτωση που πραγματοποιείται αφαίρεση ακραίων τιμών του δείγματος. Όπως παρουσιάζεται στον *πίνακα 13*, για αριθμό συστάδων *k=5*, η τιμή του σφάλματος αρχικά είναι *23.366*. Μετά το στάδιο της επιλογής χαρακτηριστικών το σφάλμα μειώνεται στις *17.808*, ενώ μετά την αφαίρεση των ακραίων τιμών το σφάλμα μειώνεται στις *17.159*. Αντίστοιχη μείωση του σφάλματος παρατηρείται και για την περίπτωση που επιλέγεται αριθμός συστάδων *k=10*, καθώς αρχικά η τιμή είναι *20.097*, στη συνέχεια μειώνεται στις *14.289*, ενώ τέλος μειώνεται στις *13.997*. Παρατηρούμε λοιπόν ό,τι για συγκεκριμένο αριθμό συστάδων, η τιμή του ***sum of squared error*** είναι μικρότερη για την περίπτωση όπου στα δεδομένα έχει εφαρμοστεί επιλογή χαρακτηριστικών αλλά και απομάκρυνση των ακραίων τιμών του δείγματος.

No of clusters	*Data preparation*	*Cluster sum of squared error*
5	*No preparation*	*23.366*
10		*20.097*
5	*Attribute selection*	*17.808*
10		*14.289*
5	*Outliers removed*	*17.159*
10		*13.997*

Πίνακας 13
Συσταδοποίηση για k=5, k=10 με αντίστοιχη επεξεργασία δεδομένων (churners)

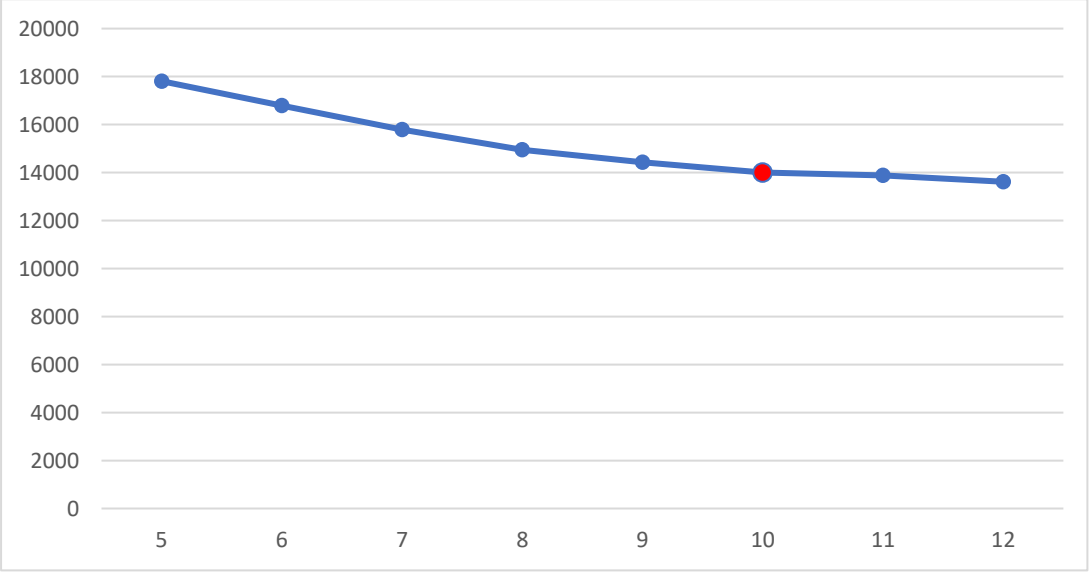
Στη συνέχεια, ο αλγόριθμος *k-means* εφαρμόζεται επαναληπτικά για αριθμό συστάδων *k=2,3,...,11,12* στα δεδομένα της βάσης δεδομένων πελατών που

αποχώρησαν, για τα οποία έχει προηγηθεί τόσο η επιλογή χαρακτηριστικών, όσο και η αφαίρεση των ακραίων τιμών. Για μικρό αριθμό συστάδων (*k=2,3,4*), η τιμή του σφάλματος παραμένει σε υψηλές τιμές, επομένως ο αλγόριθμος επιστρέφει κακής ποιότητας συσταδοποίησης. Παρατηρούμε ό,τι για μεγαλύτερο αριθμό συστάδων (*k=20,k=30*), η τιμή του σφάλματος μειώνεται. Ωστόσο, η τμηματοποίηση των πελατών σε τόσες ομάδες δεν έχει πρακτική σημασία καθώς δημιουργεί πολλές και ολιγομελείς ομάδες πελατών που δεν αξίζει να αναληθούν περαιτέρω στα πλαίσια του στοχευμένου μάρκετινγκ.

Με την εφαρμογή του ***elbow method***, βρίσκουμε ότι ο βέλτιστος αριθμός συστάδων για τις οποίες εχει πρακτική σημασία η συσταδοποίηση με ταυτόχρονα καλή ποιότητα αποτελέσματος είναι ***k=10***. Συνεπώς αυτή η περίπτωση επιστρέφει την καλύτερης ποιότητας συσταδοποίηση.

No of clusters	*Cluster sum of squared error*
5	*17.808*
6	*16.792*
7	*15.783*
8	*14.945*
9	*14.427*
10	*13.997*
11	*13.886*
12	*13.618*

Πίνακας 14
Elbow method



Γράφημα 2
Elbow method για εύρεση ιδανικού αριθμού συστάδων

Η πολυπληθέστερη ομάδα που προκύπτει από την εφαρμογή του k-means για αριθμό συστάδων ***k=10*** στη βάση δεδομένων των πελατών που αποχώρησαν από την εταιρεία (***churners***) περιλαμβάνει ***1505*** πελάτες (δηλαδή περίπου ***16%*** του συνολικού δείγματος). Τα χαρακτηριστικά της ομάδας αυτής είναι:

- Το μεγάλο χρονικό διάστημα που μεσολαβεί από την τελευταία φορά που αυτοί πραγματοποίησαν αγορά στην εταιρεία (9 μήνες), συγκριτικά με το μέσο όρο (5 μήνες)
- Το μικρό χρηματικό ποσό που διατέθηκε κατά μέσο όρο για τις αγορές του κάθε πελάτη (140$), σε σχέση με το μέσο όρο των χρημάτων που δανάνησε ο κάθε πελάτης και αφορά το σύνολο του δείγματος (242$)
- Οι πελάτες κατοικούν σε προαστιακές περιοχές
- Οι πελάτες δεν έχουν χρησιμοποιήσει στο παρελθόν προσφορά ένα συν ένα (buy one get one – bogo)
- Η εταιρεία δεν είχε προσεγγίσει τους συγκεκριμένους πελάτες μέσω προωθητικών ενεργειών στο παρελθόν
- Η προσφορά που έγινε στους πελάτες στα πλαίσια της τελευταίας προωθητικής καμπάνιας ήταν έπτωση στις αγορές τους (discount)

Χαρακτηριστικά	***1.505 πελάτες (16%)***
Recency	*9*
History	*140$*
Used bogo	*No*
Zip code	*Suburban*
Is referral	*No*
offer	*Discount*

Πίνακας 15
Κύρια ομάδα πελατών που αποχώρησαν (churners)

Τα χαρακτηριστικά της συγκεκριμένης ομάδας σκιαγραφούν με μεγάλη ακρίβεια το προφίλ πελατών που αποχωρούν από την εταιρεία. Πρόκειται για πελάτες που πραγματοποιούν σπάνια αγορές σε αυτή, καθώς δεν έχουν ξοδέψει αρκετά χρήματα για αγορές, ενώ μεσολαβεί μεγάλο χρονικό διάστημα χωρίς να έχει πραγματοποιηθεί αγορά. Ακόμα, η εταιρεία δεν είχε προσεγγίσει τους συγκεκριμένους πελάτες με προωθητικά μέσα για να τους εντάξει στο πελατολόγιό της και δεν τους είχε προσφερθεί στο παρελθόν κάποιο κίνητρο ώστε να επιστρέψουν για αγορές στην εταιρεία, όπως η προσφορά buy one get one. Πρόκειται για ένα προφίλ πελατών με το οποίο η εταιρεία δεν είχε ασχοληθεί στο παρελθόν, σε συνέπεια αυτοί να πραγματοποιήσουν σποραδικές αγορές και τελικά να αποχωρήσουν, καθώς δεν αναπτύχθηκε αίσθημα εμπιστοσύνης και αφοσίωσης στην εταιρεία, παρά την τελική εκπτωτική προσφορά που έγινε σε αυτούς.

Στη συνέχεια παρουσιάζονται τα χαρακτηριστικά των δύο επόμενων πολυπληθέστερων ομάδων πελατών που αποχώρησαν από την εταιρεία, όπως προκύπτει από τη συσταδοποίηση.

Χαρακτηριστικά	**1.074 πελάτες (15%)**	**1057 πελάτες (11%)**
Recency	4	6
History	170$	201$
Used bogo	Yes	No
Zip code	Rural	Urban
Is referral	no	Yes
offer	No offer	bogo

Πίνακας 16
Ομάδες πελατών που αποχώρησαν (churners)

Η δεύτερη πολυπληθέστερη ομάδα περιλαμβάνει πελάτες που κατοικούν σε αγροτικές περιοχές. Πρόκειται για πελάτες που δεν πραγματοποιούν αγορές συχνά (τελευταία αγορά πριν από τέσσερις μήνες) και δεν ξοδεύουν αρκετά χρήματα (170$). Είναι πελάτες που έχουν χρησιμοποιήσει πακέτο προσφοράς από την εταιρεία στο παρελθόν (bogo), χωρίς ωστόσο να τους έχει προσεγγίσει η εταιρεία μέσω προωθητικής καμπάνιας, ούτε με κάποια νέα προσφορά για να τους κρατήσει κοντά της. Πρόκειται λοιπόν για πελάτες που η εταιρεία δε στόχευσε ποτέ συστηματικά. Φαίνεται να είναι ένα προφίλ πελάτη, όπως και της πρώτης ομάδας, που δεν ενδιαφέρει την εταιρεία συγκριτικά με άλλες ομάδες πελατών, καθώς δεν ξοδεύουν αρκετά χρήματα για τις αγορές τους, δεν πραγματοποιούν αγορές σε σταθερή βάση και είναι κάτοικοι αγροτικών περιοχών, κάτι που ενδεχομένως δεν επιτρέπει τη δυνατότητα εφαρμογής μεθόδων μάρκετινγκ και στρατολόγησης πελατών στο βαθμό που αυτό είναι εφικτό σε αστικές ή ημιαστικές περιοχές.

Η τρίτη πολυπληθέστερη ομάδα αφορά πελάτες αστικών περιοχών που ξοδεύουν κατά μέσο όρο περισσότερα χρήματα (210$) σε σχέση με τις άλλες δύο ομάδες, χωρίς ωστόσο να πραγματοποιούν συχνά αγορές (6 μήνες από την τελευταία αγορά), ούτε να ξεπερνούν το ποσό που δαπανά κάθε πελάτης κατά μέσο όρο (242$). Είναι πελάτες που απέκτησε η εταιρεία μέσω μεθόδων μάρκετινγκ στο παρελθόν και που τους έχει γίνει νέα προσφορά για τις επόμενες αγορές τους (bogo), χωρίς ωστόσο να καταφέρει να τους κρατήσει κοντά της.

Κοινά χαρακτηριστικά και στις τρεις πολυπληθέστερες ομάδες που προέκυψαν είναι το χαμηλό ποσό που ξοδεύει κάθε πελάτης για τις αγορές του αλλά και τα μεγάλα διαστήματα που μεσολαβούν μεταξύ των αγορών. Φαίνεται λοιπόν, πως παρά τις διαφορετικές προσφορές που έχουν γίνει σε κάθε ομάδα πελατών στο παρελθόν, παρά τον τρόπο με τον οποίο ο κάθε πελάτης ξεκίνησε να πραγματοποιεί αγορές στην εταιρεία και παρά τον τόπο διαμονής του, η συχνότητα με την οποία πραγματοποιεί αγορές αλλά και τα ποσά που ξοδεύει σε αυτές είναι

τελικά τα χαρακτηριστικά που καθορίζουν σε μεγαλύτερο βαθμό το αν ο πελάτης παραμένει πιστός ή όχι στην εταιρεία.

Με τον ίδιο τρόπο εφαρμόζεται ο *k-means* στη βάση δεδομένων των πελατών που παραμένουν στην εταιρεία (***non churners***), που είναι η πολυπληθέστερη βάση, καθώς αποτελείται από ***54.606*** καταχωρήσεις. Ο ***πίνακας 17*** παρουσιάζει ενδεικτικά το αποτέλεσμα της συσταδοποίησης για ***k=5, k=10*** και ***k=12*** , με την αντίστοιχη επεξεργασία δεδομένων για κάθε περίπτωση. Παρατηρούμε ό,τι για δεδομένο αριθμό συστάδων, η τιμή του σφάλματος ελαχίστων τετραγώνων μειώνεται μετά την επιλογή χαρακτηριστικών και την αφαίρεση των ακραίων τιμών, όπως συμβαίνει και στη βάση δεδομένων των πελατών που αποχώρησαν. Για παράδειγμα, για αριθμό συστάδων *k=10*, η αρχική τιμή του σφάλματος είναι *113.446*. Μετά την επεξεργασία δεδομένων, η τιμή του σφάλματος μειώνεται στις *92.802*, ενώ με την αφαίρεση και των ακραίων τιμών η τιμή μειώνεται περαιτέρω στις *90.052*.

No of clusters	*Data preparation*	*Cluster sum of squared error*
5	*No preparation*	*136.624*
10		*113.446*
12		*110.525*
5	*Attribute selection*	*106.545*
10		*92.802*
12		*88.947*
5	*Outliers removed*	*105.462*
10		*90.052*
12		*85.357*

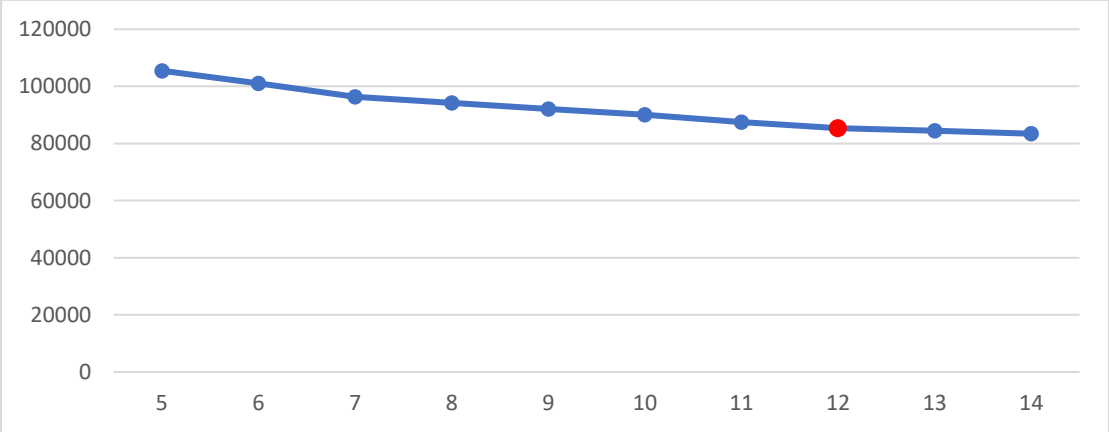
Πίνακας 17
Συσταδοποίηση για k=5, k=10 με αντίστοιχη επεξεργασία δεδομένων (non-churners)

Στη συνέχεια, ο αλγόριθμος *k-means* εφαρμόζεται επαναληπτικά για αριθμό συστάδων *k=2,3,...,13,14* στη βάση δεδομένων πελατών που παραμένουν στην εταιρεία. Η τμηματοποίηση των πελατών σε περισσότερες συστάδες δεν έχει πρακτική σημασία καθώς δημιουργούνται πολλές ομάδες με μικρό αριθμό πελατών, οι οποίες δεν αποτελούν πληθυσμιακά επαρκείς ομάδες για περαιτέρω έρευνα.

Με την εφαρμογή του ***elbow method***, βρίσκουμε ότι ο βέλτιστος αριθμός συστάδων για τις οποίες εχει πρακτική σημασία η συσταδοποίηση με ταυτόχρονα καλής ποιότητας αποτελέσματα είναι ***k=12***.

No of clusters	*Cluster sum of squared error*
5	*105.462*
6	*101.042*
7	*96.310*
8	*94.254*
9	*92.115*
10	*90.052*
11	*87.527*
12	*85.357*
13	*84.442*
14	*83.433*

Πίνακας 18
Elbow method



Γράφημα 3
Elbow method για εύρεση ιδανικού αριθμού συστάδων

Οι πολυπληθέστερες ομάδες που προκύπτουν από την εφαρμογή του k-means για αριθμό συστάδων ***k=12*** στη βάση δεδομένων των πελατών που δεν αποχώρησαν από την εταιρεία (***non churners***) είναι δύο. Ή πρώτη περιλαμβάνει ***8.185*** πελάτες (δηλαδή περίπου ***15%*** του συνολικού δείγματος) και η δεύτερη περιλαμβάνει ***6.074*** πελάτες (δηλαδή περίπου ***11%*** του συνολικού δείγματος).

Χαρακτηριστικά	**8.185 πελάτες (15%)**	**6.074 πελάτες (11%)**
Recency	1	3
History	319$	117$
Used bogo	Yes	No
Zip code	Suburban	Urban
Is referral	no	Yes
offer	discount	bogo

Πίνακας 19
Πολυπληθέστερες ομάδες πελατών που δεν αποχώρησαν (non churners)

Η πρώτη ομάδα αποτελείται από πελάτες που πραγματοποιούν πολύ συχνά αγορές (μηνιαία), ενώ ξοδεύουν και αρκετά χρήματα για αυτές (319$). Είναι κάτοικοι προαστιακών περιοχών που έχουν λάβει προσφορές στο παρελθόν αλλά και τώρα (bogo και discount αντίστοιχα).

Η δεύτερη ομάδα αποτελείται από πελάτες που ενώ δεν ξοδεύουν πολλά χρήματα στις αγορές τους (117$), πραγματοποιούν αγορές πολύ συχνά (τελευταία αγορά πριν τρεις μήνες). Είναι κάτοικοι αστικών περιοχών που στρατολογήθηκαν από την εταιρεία και τους έχει γίνει ξανά προσφορά για τις επόμενες αγορές τους (bogo).

Είναι εμφανές πως οι ομάδες πιστών πελατών απαρτίζονται από άτομα που πραγματοποιούν πολύ συχνά αγορές στην εταιρεία, είτε ξοδεύουν πολλά ή λιγότερα χρήματα. Πελάτες λοιπόν αστικών και προαστιακών περιοχών που πραγματοποιούν συχνά συναλλαγές με την εταιρεία αποτελούν τη βάση πιστών πελατών της εταιρείας.

4.5.3 Αποτελέσματα Association Rules

Ο αλγόριθμος ***Apriori*** εφαρμόζεται αρχικά στο σύνολο των δεδομένων, ενώ στη συνέχεια στις δύο επιμέρους βάσεις δεδομένων που έχουν δημιουργηθεί.

Η εφαρμογή του αλγορίθμου στο σύνολο των δεδομένων δεν καταφέρνει να αναγνωρίσει κάποιο συγκεκριμένο πλαίσιο ή μοτίβο μέσα στο οποίο να κινείται ο καταναλωτής και φανερώνει το αν συνεχίζει να πραγματοποιεί τις αγορές του στην επιχείρηση ή αν αποχωρεί από αυτή.

Η διαδικασία εξαγωγής κανόνων συσχέτισης λοιπόν, εφαρμόζεται επαναληπτικά για διαφορετικό σύνολο μεταβλητών κάθε φορά στη βάση δεδομένων πελατών που αποχώρησαν από την εταιρεία (***churners***).

Οι σημαντικότεροι κανόνες που προέκυψαν παρουσιάζονται παρακάτω:

- *History (0-348$), Used discount=yes, Used bogo=no → Conversion=1 (3262)*
- *History (0,348$), Used bogo=yes, Used discount=no → Conversion=1 (2824)*
- *Referral=yes, Channel=Web, History (0-348$), Used discount=yes, Used bogo=no → Conversion=1 (2717)*
- *Referral=yes, Channel=Web, History (0-348$), Used bogo=yes, Used discount=no → Conversion=1 (2315)*
- *Zip Code=Suburban, History (0-348$), Channel=Web, Used discount=yes, Offer=bogo → Conversion=1 (2061)*
- *Zip Code=Suburban, History (0-348$), Channel=Phone, Referral=no, Offer=discount → Conversion=1 (1953)*
- *Recency (9-12), History (0-348$), Zip Code=Urban → Conversion=1 (1323)*
- *Recency (9-12), History (0-348$), Zip Code=Rural, Referral=no → Conversion=1 (1029)*
- *Zip Code=Rural, Referral=no, Channel=Phone, Offer=discount → Conversion=1 (935)*

Πίνακας 20
Association Rules για churners

Παρατηρούμε ότι από τους 9.394 πελάτες που αποχώρησαν, ελάχιστοι έλαβαν από την εταιρεία έκπτωση για αγορές αλλά και προσφορά ένα συν ένα (bogo) ταυτόχρονα, κάτι που φαίνεται να επηρέασε την απόφαση τους να μη συνεχίσουν τις συναλλαγές τους με την εταιρεία. Ακόμα, οι πελάτες που έλαβαν κατά το παρελθόν εκπτωτικές προσφορές μέσω προωθητικών ενεργειών του τμήματος μάρκετινγκ της εταιρείας, εμφανίζονται δυσαρεστημένοι που αυτή η πολιτική δε συνεχίστηκε από την εταιρεία με κάποια επιπλέον προσφορά.

Παρατηρούμε ακόμα, πως αρκετά μεγάλος αριθμός πελατών (3262 και 2824) που δεν έλαβαν περισσότερες από μία εκπτωτικές προσφορές στο παρελθόν, δεν έχουν ξοδέψει αρκετά χρήματα για αγορές (μέχρι 348$), μέχρι την αποχώρησή τους από την επιχείρηση. Επίσης, κάτοικοι προαστιακών περιοχών, παρότι έχουν αποκτηθεί μέσω προωθητικών ενεργειών του τμήματος μάρκετινγκ και επικοινωνούν μέσω διαδικτύου με την επιχείρηση, δεν ξοδεύουν αρκετά χρήματα με αποτέλεσμα να αποχωρούν τελικά. Τέλος, είναι φανερό πως ανεξαρτήτου τόπου διαμονής, πελάτες που δεν πραγματοποιούν συχνά αγορές ξοδεύοντας ταυτόχρονα λίγα χρήματα αποχωρούν από την επιχείρηση.

Από την εφαρμογή του αλγορίθμου Apriori στη βάση δεδομένων των πελατών που αποχώρησαν από την εταιρεία (***churners***), είναι ξεκάθαρο πως οι πελάτες που αποχωρούν από την εταιρεία δεν ξοδεύουν πολλά χρήματα για τις αγορές τους συγκριτικά με τους υπόλοιπους, δεν πραγματοποιούν συχνά αγορές, ενώ η πλειοψηφία τους δεν έχει λάβει από την εταιρεία αρκετές προσφορές και

κίνητρα για να συνεχίσουν να πραγματοποιούν συναλλαγές με αυτή, ακόμη και αν αυτοί οι πελάτες γνώρισαν την εταιρεία μέσω μεθόδων μάρκετινγκ και διαφήμισης.

Από την εφαρμογή του ***Apriori*** στη βάση δεδομένων πελατών που παραμένουν στην εταιρεία (***non churners***), προκύπτουν οι παρακάτω κανόνες:

- *Zip Code=Suburban, Referral=yes, Channel=Web → Conversion=0 (10687)*
- *Zip Code=Suburban, Referral=yes, Channel=Phone → Conversion=0 (10841)*
- *Recency (1-4), Channel=Web, History (0-360$) → Conversion=0 (8285)*
- *Recency (1-4), Channel=Phone, History (0-360$) → Conversion=0 (8025)*
- *Recency (1-4), Referral=yes, History (0-360$) → Conversion=0 (7553)*
- *Zip Code=Urban, Used discount=yes, Offer=discount → Conversion=0 (7326)*
- *Zip Code=Urban, Used bogo=yes, Offer=bogo → Conversion=0 (7052)*
- *Recency (1-4), zip code=Urban, History (361-650) → Conversion=0 (7042)*

Πίνακας 21
Association Rules για non churners

Παρατηρούμε ότι οι πιστοί πελάτες της εταιρείας πραγματοποιούν συναλλαγές συχνά με αυτή. Οι πελάτες που πραγματοποίησαν αγορές το τελευταίο τετράμηνο και επικοινωνούν με την εταιρεία τόσο μέσω διαδικτύου όσο και μέσω τηλεφώνου, καθώς επίσης και αυτοί που έμαθαν αρχικά για την εταιρεία από τις μεθόδους μάρκετινγκ που εφαρμόστηκαν στο παρελθόν, είναι πιστοί στην εταιρεία παρότι δεν ξοδεύουν πολλά χρήματα για τις αγορές τους. Οι κάτοικοι προαστιακών περιοχών που έμαθαν για την επιχείρηση μέσω μάρκετινγκ, αποτελούν μια κατηγορία πιστών πελατών, ανεξάρτητα από τον τρόπο επικοινωνίας που χρησιμοποιούν. Επίσης, πιστοί πελάτες εμφανίζονται να είναι όσοι κάτοικοι αστικών περιοχών έχουν λάβει την ίδια προσφορά με αυτή που χρησιμοποίησαν στο παρελθόν.

Ακόμα, από τους 7.655 πελάτες που κατοικούν σε αστικές περιοχές και πραγματοποιούν αγορές συχνά (1-4 μήνες), 7.042 πελάτες έχουν ξοδέψει μεγαλύτερο ποσό χρημάτων για τις αγορές τους, συγκριτικά με τον μέσο όρο των χρημάτων που ξοδεύει ο κάθε πελάτης, κάτι που επιβεβαιώνει τη σημαντικότητα των πελατών αστικών περιοχών για την εταιρεία.

Κεφάλαιο 5: Συμπεράσματα έρευνας

5.1 Benchmarking αλγορίθμων

Στην παρούσα έρευνα εφαρμόστηκαν, με χρήση του λογισμικού ***Weka***, αλγόριθμοι επιτηρούμενης μηχανικής μάθησης για την ταξινόμηση πελατών μιας εταιρείας (***binary classification problem***) και την δημιουργία ενός μοντέλου πρόβλεψης αποχώρησης των πελατών, σε μία μη ισορροπημένη βάση δεδομένων (***class imbalance problem***). Η επιτυχία του μοντέλου κρίθηκε από την ικανότητα να αναγνωρίζει επαρκώς τις δύο κλάσεις και να κατανέμει όσο το δυνατόν καλύτερα τους πελάτες στην αντίστοιχη κλάση. Τα κύρια μέτρα απόδοσης του μοντέλου είναι η ακρίβεια ***F-measure***, οι τιμές ***TP Rate*** για την κύρια κλάση και ***TN Rate*** για τη μειονοτική κλάση, καθώς και η τιμή ***Roc Area***, η οποία μπορεί να θεωρηθεί και δείκτης εκτίμησης του σφάλματος τύπου I, που έχει μεγάλη σημασία για την επιτυχία του μοντέλου.

Ο αλγόριθμος που προσαρμόζεται καλύτερα στα δεδομένα δίνοντας ένα πολύ αποδοτικό και ικανό μοντέλο πρόβλεψης είναι ο ***Random Forest***, ενώ και ο ***kNN*** για k=3 δίνει πολύ ικανοποιητικά αποτελέσματα. Οι αλγόριθμοι ***Naïve Bayes*** ***Logistic Regression*** αντιθέτως, δεν αποδίδουν ικανοποιητικά στο συγκεκριμένο πρόβλημα.

Κατά τη δημιουργία των μοντέλων, εφαρμόστηκαν διάφορες μέθοδοι επιλογής χαρακτηριστικών και ανακατανομής των δύο κλάσεων, καθώς και ανάλυση κυρίων συνιστωσών. Παρατηρούμε ότι η εφαρμογή του ***Random Forest*** για ανακατανομή κλάσεων κατά 50% και με επιλογή χαρακτηριστικών μέσω της μεθόδου ***Correlation-based Feature Subset Selection (CfsSubsetEval)*** αποδίδει καλύτερα σε σχέση με τις άλλες μεθόδους επεξεργασίας δεδομένων, δίνοντας ένα πολύ αξιόπιστο και ικανό μοντέλο πρόβλεψης αποχώρησης πελατών.

	TP Rate (0)	*TN Rate (1)*	*F-Measure*	*ROC Area*
***1**. Random Forest*	*0,838*	*0,691*	*0,791*	*0,803*
***2**. kNN*	*0,837*	*0,690*	*0,790*	*0,787*
***3**. Logistic Regression*	*0,954*	*0,046*	*0,589*	*0,639*
***4**. Naïve Bayes*	*0,945*	*0,105*	*0,594*	*0,635*

Πίνακας 22
Benchmarking αλγορίθμων

5.2 Γενικά συμπεράσματα

Το σύνολο της βάσης πελατών χωρίστηκε σε δύο νέες βάσεις, αποτελούμενες από πελάτες που αποχώρησαν από την εταιρεία (***churners***) και πελάτες που παραμένουν σε αυτή (***non churners***). Στις βάσεις αυτές εφαρμόστηκαν

αλγόριθμοι ομαδοποίησης (***k-means***) και αλγόριθμοι εξαγωγής κανόνων συσχέτισης (***Apriori***).

Κοινά χαρακτηριστικά και στις τρεις πολυπληθέστερες ομάδες που προέκυψαν από την ομαδοποίηση των πελατών της βάσης δεδομένων που αποχώρησαν από την εταιρεία, είναι το χαμηλό ποσό που ξοδεύει κάθε πελάτης για τις αγορές του αλλά και τα μεγάλα διαστήματα που μεσολαβούν μεταξύ των αγορών. Είναι φανερό πως παρά τις διαφορετικές προσφορές που έχουν γίνει σε κάθε ομάδα πελατών στο παρελθόν, παρά τον τρόπο με τον οποίο ο κάθε πελάτης ξεκίνησε να πραγματοποιεί αγορές στην εταιρεία και παρά τον τόπο διαμονής του, η συχνότητα με την οποία πραγματοποιεί αγορές αλλά και τα ποσά που ξοδεύει σε αυτές είναι τελικά τα χαρακτηριστικά που καθορίζουν σε μεγαλύτερο βαθμό το αν ο πελάτης παραμένει πιστός ή όχι στην εταιρεία.

Από την εφαρμογή του k-means στη βάση δεδομένων των πιστών πελατών της εταιρείας, προκύπτει πως οι ομάδες πιστών πελατών απαρτίζονται από άτομα που πραγματοποιούν πολύ συχνά αγορές στην εταιρεία, είτε ξοδεύοντας πολλά χρήματα είτε όχι. Η εταιρεία είναι πιο εύκολο να προσεγγίσει πελάτες αστικών και προαστιακών περιοχών μέσω του μάρκετινγκ, σε σχέση με πελάτες αγροτικών πειροχών στους οποίους η εταιρεία δε δίνει ιδιαίτερη βάση για θέματα προώθησης προϊόντων και προσφορών. Οι πελάτες αστικών και προαστιακών περιοχών λοιπόν, που πραγματοποιούν συχνά συναλλαγές με την εταιρεία, αποτελούν τη βάση πιστών πελατών της εταιρείας.

Από την εφαρμογή του αλγορίθμου Apriori στη βάση δεδομένων των πελατών που αποχώρησαν από την εταιρεία, βλέπουμε ότι οι πελάτες που δεν ξοδεύουν αρκετά χρήματα για αγορές στην εταιρεία είναι πιο πιθανό να αποχωρήσουν. Οι πελάτες αυτοί φαίνεται να μην είναι ικανοποιημένοι με τις προσφορές που έχουν λάβει μέχρι τώρα, με αποτέλεσμα να μην πραγματοποιούν συχνά αγορές. Η εταιρεία δίνει ένα είδος προσφοράς σε κάθε πελάτη, είτε έκπτωση, είτε προσφορά ένα συν ένα. Για να μπορέσει να κρατήσει περισσότερους πελάτες κοντά της, θα μπορούσε να προσφέρει συνδυαστικά πακέτα προσφορών στους πελάτες της, ακόμα και σε αυτούς που έχουν ήδη λάβει κάποια προσφορά στο παρελθόν, σκεπτόμενοι πάντα το ποσό που έχουν δαπανήσει για αγορές.

Η εφαρμογή του αλγορίθμου Apriori στη βάση δεδομένων των πελατών που δεν αποχώρησαν από την εταιρεία, έδειξε ό,τι οι πιστοί πελάτες είναι κατά βάση αυτοί που πραγματοποιούν συχνά αγορές, ανεξάρτητα με το ποσό που ξοδεύουν.

Από το συνδυασμό των αποτελεσμάτων της ομαδοποίησης και της εξαγωγής γνώσης μέσω των κανόνων, προκύπτει ό,τι οι πελάτες που είναι πιθανότερο να αποχωρήσουν είναι κυρίως αυτοί που δεν πραγματοποιούν συχνά αγορές με την εταιρεία, ενώ αντίστοιχα οι πιστοί πελάτες είναι αυτοί που πραγματοποιούν αγορές αρκετά συχνά.

Συγκεκριμένα, κάτοικοι πραστιακών και αγροτικών περιοχών, οι οποίοι δεν έχουν προσεγγιστεί με μεθόδους μάρκετινγκ από την εταιρεία, δεν

πραγματοποιούν συχνά αγορές με αυτή και ξοδεύουν λίγα χρήματα, αποτελούν το προφίλ πελατών που είναι πιθανότερο να αποχωρήσουν. Στοχεύοντας στους συγκεκριμένους πελάτες, η εταιρεία μπορεί να βρεί τρόπους καλύτερης επικοινωνίας με αυτούς, να αυξήσει τους τρόπους διαφήμισης και προώθησης των προϊόντων της στις περιοχές όπου κατοικούν αλλά και να προσφέρει κίνητρα σε αυτούς τους πελάτες για πραγματοποίηση αγορών σε πιο σταθερή βάση.

Οι πελάτες που αποτελούν τη βάση πιστών πελατών της εταιρείας και επιδεικνύουν ιδιαίτερη αφοσίωση σε αυτή είναι πελάτες αστικών και προαστιακών περιοχών που πραγματοποιούν συχνά αγορές. Από αυτούς τους πελάτες, οι κάτοικοι αστικών περιοχών μαθαίνουν για την εταιρεία μέσω διαφήμισης και προώθησης συχνότερα από τους υπόλοιπους. Η εταιρεία θα πρέπει να συνεχίσει αυτή τη μέθοδο στις αστικές περιοχές και ταυτόχρονα να την επεκτείνει πιο αποδοτικά στις προαστιακές, όπου υπάρχουν πελάτες που ξοδεύουν αρκετά χρήματα και είναι σημαντικό να συνεχίσουν να ανήκουν στους πιστούς πελάτες.

Από την ανάλυση των κανόνων συσχέτισης φαίνεται ακόμα πως αρκετοί πελάτες αστικών περιοχών δαπανούν μεγάλα ποσά σε αγορές, κάτι που ισχύει και για αρκετούς κατοίκους προαστιακών περιοχών, όπως προέκυψε από τη διαδικασία ομαδοποίησης. Είναι χαρακτηριστικό πως από τους 20 πελάτες που έχουν ξοδέψει για αγορές τα περισσότερα χρήματα, οι 17 παραμένουν πιστοί στην εταιρεία. Κατά βάση οι πελάτες που ξοδεύουν τα μεγαλύτερα ποσά, είναι και οι πιο πιστοί, οπότε η εταιρεία οφείλει να προσφέρει τις καλύτερες δυνατές υπηρεσίες σε αυτή την κατηγορία πελατών για να τους κρατήσει αφοσιωμένους και ικανοποιημένους.

5.3 Αξιοποίηση της παρούσας εργασίας και περαιτέρω έρευνα

Στην παρούσα εργασία αναλύθηκαν οι έννοιες της ικανοποίησης και αφοσίωσης πελατών, καθώς επίσης και η σημασία πρόβλεψης αποχώρησης πελατών, που αποτελούν τη βάση για την ανάπτυξη και εφαρμογή σύγχρονων μεθόδων μάρκετινγκ από τις επιχειρήσεις. Στόχος της έρευνας ήταν η δημιουργία ενός απλού και αποδοτικού μοντέλου πρόβλεψης αποχώρησης πελατών μιας εταιρείας με χρήση αλγορίθμων μηχανικής μάθησης, όπως επίσης και η μελέτη της βάσης δεδομένων των πελατών της, για τη σκιαγράφηση του προφίλ πελατών που είναι πιο πιθανό να αποχωρήσουν, εκείνων που αποτελούν τη βάση πιστών πελατών, αλλά και τον εντοπισμό ενδιαφέρων σχέσεων μεταξύ των χαρακτηριστικών, που να δικαιολογούν τη συμπεριφορά των πελατών, με χρήση αλγορίθμων συσταδοποίησης και εξαγωγής κανόνων συσχέτισης.

Η συγκεκριμένη εργασία αποτελεί μια αρχική προσέγγιση για τη χρήση και εφαρμογή μεθόδων μηχανικής μάθησης σε θέματα μάρκετινγκ και διαχείρησης πελατών. Οι αλγόριθμοι και οι μέθοδοι επεξεργασίας δεδομένων εφαρμόστηκαν στο περιβάλλον του ελεύθερου λογισμικού Weka, μια βιβλιοθήκη αλγορίθμων μηχανικής μάθησης. Σε συνέχεια της εργασίας, οι αλγόριθμοι που εφαρμόστηκαν

μπορούν να αναπτυχθούν στο περιβάλον της γλωσσας ***Python*** ή της γλώσσας ***R***, για σύγκριση των αποτελεσμάτων σε κάθε περίπτωση. Επίσης, για την ταξινόμηση των δεδομένων μπορούν να χρησιμοποιηθούν τα ***Artificial Neural Networks*** (Τεχνητά Νευρωνικά δίκτυα). Η εξαγωγή κανόνων μπορεί να πραγματοποιηθεί με χρήση ***Rough Sets*** (θεωρία ακατέργαστων συνόλων). Ακόμα, για την παραγωγή κανόνων μπορεί να εφαρμοστεί η μέθοδος ***fs QCA*** (fuzzy-set Qualitative Comparative Analysis).

Βιβλιογραφία

[1] Kotler, P., & Armstrong, G. (2013), *Principles of Marketing,* North York: Pearson

[2] Παντουβάκης, Α., Σιώμκος, Γ., Χρήστου, Ε., *Μάρκετινγκ,* Αθήνα: Λιβάνης

[3] Kotler, P. (2004), *A three-part plan for upgrading your marketing department for new challenges,* Strategy & Leadership, Vol.32 No.5, pp 4-9

[4] Chalmeta, R. (2006), *Methodology for customer relationship management*, Journal of Systems and Software, Volume 79, Issue 7, pp 1015-1024

[5] Giese, J. & Cote, J. (2000), *Defining Customer Satisfaction,* Academy of Marketing Science Review, 4, pp 1-24

[6] *What is Customer Loyalty: Definition and Guide*, (n.d.). Διαθέσιμο στον δικτυακό τόπο: https://sendpulse.com/support/glossary/customer-loyalty (10/5/2021)

[7] Bhat, A. (n.d.), *Customer Churn: Definition, Rate, Analysis and Prediction*, Διαθέσιμο στον δικτυακό τόπο: https://www.questionpro.com/blog/customer-churn/ (15/5/2021)

[8] Vafeiadis, T., Diamantaras, K., et al. (2015), *A comparison of machine learning techniques for customer churn prediction,* Simulation Modelling Practice and Theory, 55, pp 1-9

[9] Huang, Y. & Kechadi, T. (2013), *An effective hybrid learning system for telecommunication churn prediction,* Expert Systems with Applications, 40, pp 5635-5647

[10] Farquad, M., Ravi, V. & Raju, B. (2014), *Churn prediction using comprehensive support vector machine: An analytical CRM application,* Applied Soft Computing, 19, pp 31-40

[11] Glady, N., Baesens, B. & Croux, C. (2009), *Modeling churn using customer lifetime value,* European Journal of Operational Research, 197, pp 402-411

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

[13] Amin, A. et al. (2017), *Customer churn prediction in the telecommunication sector using a rough set approach,* Neurocomputing, 237, pp 242-254

[14] Burez, J. & Van den Poel, D. (2009), *Handling Class Imbalance in Customer Churn Prediction,* Expert Systems with Applications, 36, pp 4626-4636

[15] Mitchell, T. (1997), *Machine Learning,* New York: McGraw-Hill Education

[16] Kononenko, I. et al. (2007), *Machine Learning and Data Mining,* Swanston, Cambridge: Woodhead Publishing

[17] Ethem, A. (2009), *Introduction to Machine Learning,* Boston: The MIT Press; 2nd edition

[18] Siebes, A., *A Gentle Introduction to Machine Learning,* Algorithmic Data Analysis Group, Department of Information and Computing Sciences, Universiteit Utrecht

[19] Τσαφαράκης, Σ. *Διαχωριστικοί αλγόριθμοι (k-means),* Εκπαιδευτικές σημειώσεις στο μάθημα Εισαγωγή στην Τεχνητή Νοημοσύνη, Χανιά: Πολυτεχνείο Κρήτης, Σχολή Μηχανικών Παραγωγής και Διοίκησης

[20] Ταβερναράκη Σ. (2020), *Αλγόριθμοι μηχανικής μάθησης για την προσέλκυση πελατών: Μια συγκριτική αξιολόγηση στο χώρο των τραπεζικών υπηρεσιών,* Διπλωματική εργασία, Πολυτεχνείο Κρήτης, Σχολή Μηχανικών Παραγωγής και Διοίκησης

[21] Theobald, O. (2017), *Machine Learning for Absolute Beginners, 2nd edition,* Scatterplot Press

[22] Brownlee, J. (2019), *How to choose a feature selection method for machine learning,* Διαθέσιμο στον δικτυακό τόπο: https://machinelearningmastery.org/ (18/5/2021)

[23] Σκιαδάς, Χ. (2006), *Εκπαιδευτικές σημειώσεις στο μάθημα Ανάλυση Δεδομένων,* Χανιά: Πολυτεχνείο Κρήτης, Σχολή Μηχανικών Παραγωγής και Διοίκησης

[24] Yildirim, S. (2020), *Principal Component Analysis-Explained,* Διαθέσιμο στον δικτυακό τόπο: https://towardsdatascience.com (21/5/2021)

[25] Brownlee, J. (2020), *Train-Test split for evaluating machine learning algorithms,* Διαθέσιμο στον δικτυακό τόπο: https://ht.com/train-test-split-for-evaluating-machine-learning-algorithms/

[26] Jain, A. *100 Days of ML Code,* Διαθέσιμο στον δικτυακό τόπο: *https://github.com/Avik-Jain/100-Days-Of-ML-Code*

[27] Hastie, T. Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning,* NY: Springer