ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Π.Μ.Σ «ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ & ΑΝΑΛΥΤΙΚΗ»

Πρόβλεψη Απώλειας Πελατών για Εμπορικές Επωνυμίες Καταναλωτικών Αγαθών με τη χρήση Δεδομένων Ηλεκτρονικών Συναλλαγών

Εκπόνηση: Μαρία-Μαλεβή Παπαδάκη, ΜΕ2224
Επιβλέπων Καθηγητής: Μιχαήλ Φιλιππάκης

Πειραιάς
Φεβρουάριος 2024

Περίληψη

Η παρούσα διπλωματική εργασία ασχολείται με το πρόβλημα της Πρόβλεψης Αποχώρησης Πελατών, εστιάζοντας στη νέα προσέγγιση της πρόβλεψης πιθανής αποχώρησης πελατών σε επίπεδο εμπορικής επωνυμίας καταναλωτικών αγαθών στα πλαίσια του ηλεκτρονικού εμπορίου. Ο κύριος στόχος είναι να εντοπιστούν πελάτες που είναι πιθανό να απομακρυνθούν από μια συγκεκριμένη εμπορική επωνυμία, προσφέροντας πολύτιμες πληροφορίες για τους κατασκευαστές προϊόντων αναφορικά με τη διαμόρφωση μελλοντικών στρατηγικών και προωθητικών ενεργειών σε συνεργασία με τους εμπόρους λιανικής.

Βασιζόμενη σε πραγματικά δεδομένα συναλλαγών από ένα μεγάλο διαδικτυακό φαρμακείο του εξωτερικού, η μελέτη δεν περιλαμβάνει δημογραφικά, γεωγραφικά ή προσωπικά στοιχεία πελατών και χρησιμοποιεί αποκλειστικά δεδομένα συναλλαγών. Αξιοποιεί τη γλώσσα SQL για την απόκτηση δεδομένων και τη γλώσσα Python για την κατασκευή συνόλου δεδομένων, την εκπαίδευση μοντέλων μηχανικής μάθησης και την οπτικοποίηση των αποτελεσμάτων.

Η πειραματική μελέτη αποκαλύπτει ότι η πρόβλεψη της αποχώρησης πελατών σε επίπεδο εμπορικής επωνυμίας είναι δυνατή, ανοίγοντας νέους δρόμους για έρευνα, οι οποίοι απέχουν από τις συμβατικές προσεγγίσεις που απευθύνονται κυρίως σε ολόκληρες επιχειρήσεις ή ολόκληρα ηλεκτρονικά καταστήματα. Η δοκιμή αρκετών αλγορίθμων μηχανικής μάθησης και τεχνικών επιλογής χαρακτηριστικών οδηγεί σε ένα τελικό μοντέλο πρόβλεψης με ακρίβεια κοντά στο 70%.

Συμπερασματικά, η παρούσα μελέτη όχι μόνο καταδεικνύει την εφικτότητα της πρόβλεψης αποχώρησης πελατών σε επίπεδο εμπορικής επωνυμίας, αλλά προτείνει και πρακτικές εφαρμογές, όπως η ενσωμάτωση στην πλατφόρμα eRAM. Επιπλέον, περιγράφει μελλοντικές κατευθύνσεις για τη βελτίωση της ακρίβειας πρόβλεψης και τη διεύρυνση του εύρους της χρησιμότητας του μοντέλου.

Λέξεις Κλειδιά: πρόβλεψη απώλειας πελατών, ηλεκτρονικό εμπόριο, μηχανική μάθηση

Abstract

This master's thesis delves into the realm of Customer Churn Prediction, focusing on the novel approach of forecasting potential customer churn at the level of consumer goods brands in the context of electronic commerce. The central aim is to identify customers likely to disengage from a specific brand, offering insights valuable for product manufacturers to shape future strategies and promotional actions in collaboration with retailers.

Drawing upon real transactional data from a prominent international online pharmacy, the study refrains from including demographic, geographic, or personal customer information and uses solely transactional data instead. It leverages SQL for data acquisition and Python for dataset construction, machine learning model training, and result visualization.

The investigation reveals that predicting customer churn at a brand level is possible, opening new avenues for research, departing from conventional approaches that primarily address businesses or electronic stores. Several machine learning algorithms and feature selection techniques are tested, leading to a final predictive model with an accuracy nearing 70%.

In conclusion, this research not only demonstrates the efficacy of predicting customer churn at the brand level but also proposes practical applications, such as integration into the eRAM platform. Furthermore, it outlines future pathways for enhancing predictive accuracy and broadening the scope of the model's utility.

Keywords: churn prediction, ecommerce, machine learning

Ευχαριστίες

Με την ολοκλήρωση της διπλωματικής μου εργασίας θα ήθελα να ευχαριστήσω όλους όσους συνέβαλαν άμεσα ή έμμεσα στην εκπόνησή της.

Ευχαριστώ θερμά τον επιβλέποντα καθηγητή μου, κύριο Μιχαήλ Φιλιππάκη, για την πλήρη εμπιστοσύνη που μου έδειξε κατά την ανάθεση της διπλωματικής εργασίας καθώς και για την καθοδήγησή του σε οτιδήποτε χρειάστηκα κατά τη διάρκεια της συνεργασίας μας.

Ευχαριστώ ιδιαιτέρως τον κο Γιώργο Πάντζαρη, επικεφαλή του Engineering στην Convert Group, όπου και εργάζομαι, για την αρχική ιδέα σχετικά με το θέμα της εργασίας καθώς και για την ευγενική παραχώρηση όλων των δεδομένων που χρειάστηκα.

Ένα μεγάλο ευχαριστώ οφείλω επιπλέον στην οικογένεια και στους φίλους(-ες) μου για την ηθική υποστήριξη και την κατανόηση τους σε όλη τη διάρκεια των μεταπτυχιακών σπουδών μου. Τέλος, θα ήθελα να ευχαριστήσω το γάτο μου, Φούντα, ο οποίος στεκόταν κυριολεκτικά στο πλάι μου σε κάθε λεπτό συγγραφής της εργασίας μου.

Μαρία-Μαλεβή Παπαδάκη

Πειραιάς, Φεβρουάριος 2024

Περιεχόμενα

Λίστα Εικόνων-Πινάκων-Αποτελεσμάτων Κώδικα

Εικόνες

Πίνακες

Κεφάλαιο 1 - Εισαγωγή

1.1 Αντικείμενο της Διπλωματικής

Σε όλη τη διαθέσιμη βιβλιογραφία σχετικά με την επιχειρησιακή ευφυΐα είναι κοινή παραδοχή πως η προσέλκυση νέων πελατών κοστίζει σημαντικά περισσότερο από την διατήρηση των υπαρχόντων. Ως εκ τούτου, εκατοντάδες διαθέσιμες επιστημονικές εργασίες επιχειρούν να ανιχνεύσουν, στο σύνολο της πελατειακής βάσης μιας επιχείρησης, τους πελάτες εκείνους που έχουν τη μεγαλύτερη πιθανότητα να πάψουν να αγοράζουν τα προϊόντα ή τις υπηρεσίες μιας επιχείρησης σε μια ορισμένη στιγμή στο μέλλον. Η παραπάνω πρόβλεψη είναι ευρέως γνωστή ως 'Customer Churn Prediction'.

Η προσπάθεια αυτή μπορεί να αφορά οποιοδήποτε κλάδο. Σχετικές εργασίες αναφέρονται σε κλάδους όπως τηλεπικοινωνίες, ηλεκτρονική τραπεζική, ηλεκτρονικό εμπόριο, υπηρεσίες από επιχείρηση προς επιχείρηση (B2B), συνδρομητικές διαδικτυακές τηλεοπτικές υπηρεσίες (streaming), συνδρομητικές υπηρεσίες σε ηλεκτρονικά παιχνίδια (gaming), συνδρομές σε ηλεκτρονικά μέσα ενημέρωσης, αγοραπωλησία επενδυτικών προϊόντων και πολλά ακόμα. Αν και κάθε κλάδος έχει τις ιδιαιτερότητές του, συνήθως οι προσεγγίσεις είναι παρόμοιες, υιοθετώντας εργαλεία από τη μηχανική μάθηση, τη στατιστική, τη θεωρεία πιθανοτήτων και τα μεγάλα δεδομένα.

Αναφορικά με την πρόβλεψη της πιθανής αποχώρησης πελατών στον κλάδο του ηλεκτρονικού εμπορίου υπάρχει πληθώρα σχετικών δημοσιεύσεων. Όλες ωστόσο προσεγγίζουν το πρόβλημα θεωρώντας ως επιχείρηση που θέλει να κρατήσει τους πελάτες της το εκάστοτε ηλεκτρονικό κατάστημα. Ένα ηλεκτρονικό κατάστημα συνήθως είναι ο εκπρόσωπος στο αγοραστικό κοινό ενός αριθμού από προμηθευτές προϊόντων. Για παράδειγμα ένα ηλεκτρονικό φαρμακείο πουλάει καλλυντικά της μάρκας Α, καλλυντικά της μάρκας Β, συμπληρώματα διατροφής της μάρκας Δ, βρεφικές τροφές της μάρκας Ε κοκ. Μέχρι στιγμής δεν συναντάμε κάποια προσπάθεια πρόβλεψης απώλειας πελατών σε επίπεδο μάρκας καταναλωτικών αγαθών. Θεωρώντας μία μάρκα και όχι ένα κατάστημα ως την «επιχείρηση» για την οποία θα γίνει η παραπάνω πρόβλεψη, ανοίγεται ένα νέο πιθανό πεδίο έρευνας το οποίο θα μπορούσε να καταλήξει σε χρήσιμη πληροφορία για ένα κατασκευαστή προϊόντων που θα επηρεάσει τη στρατηγική του για το μέλλον και τις πιθανές προωθητικές ενέργειες που θα σχεδιάσει σε συνεργασία με τους πωλητές λιανικής.

1.2 Σκοπός της Διπλωματικής

Η παρούσα μεταπτυχιακή διπλωματική εργασία έχει σκοπό της διερεύνηση της δυνατότητας να προβλέψουμε τη μελλοντική «απομάκρυνση» πελατών από μία συγκεκριμένη επωνυμία καταναλωτικών αγαθών στα πλαίσια των αγορών του σε ένα ηλεκτρονικό κατάστημα. Η απομάκρυνση αυτή μπορεί να μεταφραστεί ως 2 σενάρια: 1) ο πελάτης έπαψε εντελώς τις αγορές του στο συγκεκριμένο ηλεκτρονικό κατάστημα, 2) ο πελάτης συνεχίζει να αγοράζει από το ηλεκτρονικό κατάστημα αλλά στρέφεται σε διαφορετικές επωνυμίες.

Τα δεδομένα που χρησιμοποιούνται είναι πραγματικά και προέρχονται από αληθινό μεγάλο ηλεκτρονικό φαρμακείο του εξωτερικού. Πρόκειται για δεδομένα συναλλαγών και μόνο, χωρίς να παρέχονται δημογραφικά, γεωγραφικά ή άλλα προσωπικά στοιχεία για τους πελάτες.

Έτσι εξετάζεται επιπλέον αν θα μπορούσε να γίνει η συγκεκριμένη πρόβλεψη έχοντας στη διάθεσή μας μόνο τα ελάχιστα αυτά δεδομένα μέσω των οποίων κατασκευάζουμε εμείς τα χαρακτηριστικά του τελικού συνόλου δεδομένων για κάθε πελάτη.

Για να επιτευχθούν τα παραπάνω χρησιμοποιούνται η γλώσσα SQL, για την απόκτηση όλων των απαραίτητων πρωταρχικών δεδομένων, και η γλώσσα python για την κατασκευή του τελικού συνόλου, την εκπαίδευση μοντέλων μηχανικής μάθησης και την οπτικοποίηση των αποτελεσμάτων.

1.3 Δομή της Διπλωματικής

Το **Κεφάλαιο 2** αποτελείται από δύο μεγάλες υποενότητες. Στην πρώτη γίνεται επισκόπηση ενός αριθμού δημοσιεύσεων που αφορούν την πρόβλεψη απώλειας πελατών (customer churn prediction) σε διάφορους κλάδους. Εδώ έχει γίνει προσπάθεια να καλυφτούν όσο γίνεται περισσότεροι κλάδοι αλλά και προσεγγίσεις. Στη $2^{η}$ υποενότητα παρουσιάζονται συνοπτικά όλες οι μέθοδοι που θα χρησιμοποιηθούν στο πρακτικό μέρος της εργασίας. Παρουσιάζονται έτσι αλγόριθμοι μηχανικής μάθησης, μέθοδοι που χρησιμοποιούνται παράλληλα για την βελτίωση της απόδοσης αυτών των αλγορίθμων καθώς και οι μέθοδοι με τις οποίες θα αξιολογηθούν τα τελικά αποτελέσματα.

Στο **Κεφάλαιο 3** γίνεται συνοπτική περιγραφή του πλαισίου από το οποίο προέρχονται τα δεδομένα της εργασίας. Γίνεται συνοπτική περιγραφή της εταιρείας ‘Convert Group’, των προϊόντων της και κυρίως του εργαλείου eRAM, από τη βάση δεδομένων του οποίου έχουν παρθεί τα δεδομένα. Παρέχεται επιπλέον συνοπτική περιγραφή της δομής της βάσης δεδομένων για καλύτερη μετέπειτα κατανόηση. Τέλος περιγράφεται συνοπτικά το κεφάλαιο που ακολουθεί.

Το **Κεφάλαιο 4** περιγράφει, ύστερα από τα παραπάνω, το πρακτικό κομμάτι της μελέτης. Αρχικά παρουσιάζεται η διαδικασία επιλογής συνόλου δεδομένων, μεταξύ των διαθέσιμων υποψηφίων, και κατασκευής του τελικού συνόλου δεδομένων από τα αρχικά δεδομένα συναλλαγών. Στη συνέχεια περιγράφεται η προεπεξεργασία των δεδομένων, η εφαρμογή διάφορων τεχνικών μείωσης διαστάσεων και, τέλος, η εκπαίδευση ενός αριθμού διαφορετικών μοντέλων ταξινόμησης.

Τα αποτελέσματα αναλύονται και σχολιάζονται, τέλος στο **Κεφάλαιο 5**. Εδώ έχουμε πίνακες όπου συνοψίζονται ανά μοντέλο οι διάφορες μετρικές αξιολόγησης. Παρέχονται επιπλέον οι πίνακες σύγχυσης και οι καμπύλες λειτουργικού χαρακτηριστικού δέκτη (ROC) ανά μοντέλο. Καταλήγουμε στην υπεροχή ενός συνδυαστικού μοντέλου πρόβλεψης εκπαιδευμένο με τη χρήση όλων των χαρακτηριστικών του συνόλου δεδομένων του οποίου η ακρίβεια πρόβλεψης πλησιάζει στο 70%. Στην τελευταία ενότητα προτείνεται ένας πιθανός τρόπος ενσωμάτωσης του μοντέλου στην πλατφόρμα του eRAM καθώς και μελλοντικοί τρόποι βελτίωσης της ακρίβειάς του και επέκτασης του τρόπου χρήσης του.

Κεφάλαιο 2 - Θεωρητικό Υπόβαθρο

2.1 Πρόβλεψη Απώλειας Πελατών - Επισκόπηση Βιβλιογραφίας

Είναι κοινή παραδοχή σε όλη την υπάρχουσα βιβλιογραφία ότι η προσέλκυση νέων πελατών κοστίζει σημαντικά περισσότερο από την διατήρηση των υπαρχόντων. Σύμφωνα μάλιστα με το [1] μία αύξηση της τάξεως του 5% στη διατήρηση υπαρχόντων πελατών μπορεί να οδηγήσει σε αύξηση 25%-100% στα κέρδη της εταιρείας. Επομένως γίνεται σημαντική προσπάθεια από τις επιχειρήσεις για τον εντοπισμό πελατών που κινδυνεύουν να αποχωρήσουν και την παροχή κινήτρων σε αυτούς ώστε να αλλάξουν την απόφασή τους.

Υπάρχουν πολλοί ορισμοί για την «αποχώρηση» πελάτη ανά τη βιβλιογραφία. Κατά κοινή ομολογία, πρόκειται για μία κατάσταση κατά την οποία η συνεισφορά του στο τζίρο μιας εταιρείας μειώνεται ή διακόπτεται εντελώς. Αλλού αναφέρεται επίσης ως η διακοπή της πελατειακής σχέσης μεταξύ ενός πελάτη και μιας επιχείρησης. Γενικά μπορεί να χωριστεί σε 2 μεγάλες κατηγορίες, ανάλογα με το αν ο πελάτης έχει με την εκάστοτε επιχείρηση σχέση που διέπεται από κάποιο συμβόλαιο ή όχι. Πρόβλεψη απώλειας πελάτη, επιπλέον, ορίζεται στο [2] ως η κατασκευή ενός μοντέλου που επιτρέπει την εκτίμηση της πιθανότητας της μελλοντικής απώλειας πελατών βάσει της ιστορικής γνώσης που διαθέτουμε για τους πελάτες μιας επιχείρησης. Οι ερευνητές, προς αυτό το σκοπό, χρησιμοποιούν κυρίως στατικά-δημογραφικά στοιχεία, στοιχεία συναλλαγών και στοιχεία αλληλεπιδράσεων ενός πελάτη για να προβλέψουν τη μελλοντική του συμπεριφορά [3]. Ρυθμός απώλειας (churn rate) ορίζεται, τέλος, σύμφωνα με το [4] ως το ετήσιο ποσοστό με το οποίο οι πελάτες παύουν να εγγράφονται σε μια υπηρεσία ή τερματίζουν μια επιχειρηματική σχέση.

Η βιβλιογραφία πάνω σε προσπάθειες πρόβλεψης της πιθανής απώλειας πελατών μιας επιχείρησης είναι πολύ μεγάλη. Η πλειοψηφία των προσπαθειών χρησιμοποιεί σύνολα δεδομένων από τους κλάδους των υπηρεσιών, με τις τηλεπικοινωνίες και την ηλεκτρονική τραπεζική να κατέχουν το μεγαλύτερο μέρος των βιβλιογραφικών αναφορών. Επιπλέον οι μέθοδοι που χρησιμοποιούνται αφορούν κυρίως μηχανική μάθηση ή χρήση χρονοσειρών. Ακολουθεί σύντομη ανάλυση εργασιών που αφορούν στην πρόβλεψη απώλειας πελατών ανάλογα με τον τομέα δραστηριότητας που αυτές αφορούν.

Ηλεκτρονικό Εμπόριο

Η ιδιαιτερότητα στον κλάδο του ηλεκτρονικού εμπορίου, που είναι και το επίκεντρο του ενδιαφέροντος στην παρούσα εργασία, είναι ότι μιλάμε για πελατειακές σχέσεις που δεν επισφραγίζονται από κάποιου είδους σύμβαση. Ένας «churner» στο ηλεκτρονικό εμπόριο μπορεί να είναι πελάτης που δεν θα ξαναγοράσει ποτέ ξανά από το κατάστημα. Μπορεί όμως εξίσου να είναι και ένας πελάτης που θα αγοράσει μεν ξανά αλλά ύστερα από ένα τόσο μεγάλο χρονικό διάστημα έτσι ώστε στην παρούσα στιγμή να μπορούμε με ασφάλεια να τον κατατάξουμε στους πελάτες που έχουν χαθεί οριστικά.

Στη δημοσίευση [5] επιχειρείται η πρόβλεψη της πιθανότητας των πελατών ενός ηλεκτρονικού καταστήματος να πάψουν να είναι πελάτες του. Οι συγγραφείς χρησιμοποιούν μηχανική μάθηση, και συγκεκριμένα μηχανές διανυσμάτων υποστήριξης (SVM) με πυρήνα συνάρτησης ακτινικής βάσης (RBF).Το σύνολο δεδομένων είναι οι πελάτες ενός ηλεκτρονικού καταστήματος και τα χαρακτηριστικά που προκύπτουν από τις συναλλαγές τους. Η κύρια εργασία ταξινόμησης περιλαμβάνει τον προσδιορισμό του εάν η τρέχουσα περίοδος σύνδεσης

είναι η τελευταία του χρήστη. Για να αντιμετωπιστεί το ζήτημα των εξαιρετικά ανισόρροπων δεδομένων μεταξύ των 2 κλάσεων χρησιμοποιείται η μέθοδος της υποδειγματοληψίας (downsampling) στην κλάση με τις περισσότερες παρατηρήσεις. Μετά την κατασκευή του μοντέλου πρόβλεψης οι συγγραφείς διεξάγουν ένα πείραμα στο οποίο συμμετέχουν χρήστες με πολυάριθμες συναλλαγές, παρακολουθώντας τους για ένα μήνα για να αξιολογήσουν την ακρίβεια των προβλέψεων με βάση το ιστορικό αγορών τους. Η μελέτη υπογραμμίζει την ανάγκη για ύπαρξη πολλαπλών συνεδριών των χρηστών για τον ακριβή εντοπισμό και την κατανόηση των μοτίβων συμπεριφοράς τους που θα οδηγήσει τελικά σε πολύτιμες πληροφορίες για την ενίσχυση των στρατηγικών διατήρησης πελατών στον τομέα του ηλεκτρονικού εμπορίου.

Παρόμοια προσέγγιση, με χρήση SVM, χρησιμοποιείται και στο [6]. Τα χαρακτηριστικά για κάθε πελάτη-παρατήρηση προκύπτουν από τα στοιχεία των συναλλαγών του. Ως χαμένοι «churners» χαρακτηρίζονται οι πελάτες που δεν έχουν αγοράσει από ένα ηλεκτρονικό κατάστημα για διάστημα μεγαλύτερο του μέσου διαστήματος που μεσολαβεί μεταξύ 2 αγορών ενός πελάτη στο ίδιο κατάστημα. Με παρόμοιο τρόπο θα οριστούν και οι «churners» στο πρακτικό μέρος της παρούσας εργασίας.

Μία ενδιαφέρουσα προσέγγιση, πολύ βοηθητική στα πλαίσια αυτής της εργασίας, συναντάμε στο διαδικτυακό άρθρο [7]. Σε αυτό αναλύεται βήμα προς βήμα η διαδικασία της πρόβλεψης μελλοντικής απώλειας πελατών με το λογαριασμό Google Analytics ενός ηλεκτρονικού καταστήματος ως πηγή δεδομένων και την γλώσσα Python ως εργαλείο για την εξαγωγή, τον καθαρισμό και την ανάλυση των δεδομένων.

Ηλεκτρονική τραπεζική

Πρόκειται για τυπικό δείγμα πελατειακής σχέσης που επισφραγίζεται από κάποιο συμβόλαιο. Η αποχώρηση πελατών στον συγκεκριμένο κλάδο μπορεί να συνίσταται στο πλήρες κλείσιμο των τραπεζικών τους προϊόντων ή απλώς στην παύση χρήσης τους. Έτσι ο εντοπισμός της πραγματικής αποχώρησης αποτελεί ένα είδος πρόκλησης.

Στην εργασία [8] οι συγγραφείς δημιουργούν ένα μοντέλο πρόβλεψης στον κλάδο της τραπεζικής. Τα δεδομένα που χρησιμοποιούν αφορούν 3 εκατομμύρια πελάτες τράπεζας αναφοράς για την περίοδο 2011-2015. Πρόκειται για συνδυασμό δομημένων (όπως στοιχεία από το λογισμικό διαχείρισης πελατειακών σχέσεων, στοιχεία συναλλαγών κτλ.) και μη δομημένων στοιχείων (όπως επισκέψεις σελίδων, τηλεφωνικές συνομιλίες κτλ.). Προηγουμένως διεξάγουν τμηματοποίηση των πελατών και εστιάζουν σε μία από τις 3 κατηγορίες που προκύπτουν. Οι τεχνολογίες που χρησιμοποιούν είναι το datameer – μία εμπορική εκδοχή του οικοσυστήματος Hadoop – και η πλατφόρμα SAS για την ανάλυση των δεδομένων. Η έρευνα καταλήγει στην ύπαρξη στενής σύνδεσης μεταξύ της συμπεριφοράς ενός πελάτη της τράπεζας και της ενδεχόμενης απόφασής του να αποχωρήσει.

Η μελέτη [9] επικεντρώνεται στον εντοπισμό των πελατών με αξία στον τραπεζικό τομέα. Χρησιμοποιείται το μοντέλο RFM, για την εκτίμηση της αξίας του πελάτη, και συσταδοποίηση με τον αλγόριθμο k-means (με χαρακτηριστικά όπως πρόσφατες συναλλαγές, συχνότητα συναλλαγών, κερδοφορία). Αναγνωρίζεται ότι η μείωση των πελατών στον τραπεζικό τομέα χαρακτηρίζεται συχνά από την αδράνεια του πελάτη και όχι από το πλήρες κλείσιμο του λογαριασμού του. Μετά την κατασκευή μιας χρονοσειράς με τα ποσοστά μείωσης πελατών κατά 12 χρονικά διαστήματα, χρησιμοποιούνται το ARIMA ως γραμμικό μοντέλο και ένα νευρωνικό δίκτυο ως μη γραμμικό μοντέλο για την πρόβλεψη μελλοντικών ποσοστών μείωσης

πελατών που έχουν υψηλή αξία. Στο βήμα της πρόβλεψης, συγκρίνονται τα δύο μοντέλα, με το νευρωνικό δίκτυο να εμφανίζει καλύτερη επίδοση ως μη γραμμικό μοντέλο.

Το [10] αναφέρεται στην πρόκληση του να προβλέψουμε και να ταξινομήσουμε τους «churners» στην ηλεκτρονική τραπεζική, εστιάζοντας σε πελάτες με μειωμένη διαδικτυακή δραστηριότητα παρόλο που συνεχίζουν να διατηρούν τραπεζικούς λογαριασμούς στην τράπεζα. Η προτεινόμενη διαδικασία επτά σταδίων περιλαμβάνει την απόκτηση και τον καθαρισμό δεδομένων πελατών, την επιλογή χαρακτηριστικών, την εκπαίδευση μοντέλων πρόβλεψης και την αξιολόγηση της αποτελεσματικότητάς τους. Στόχος είναι να αποτραπεί η έξοδος των πελατών από τα ηλεκτρονικά κανάλια, σε συνάρτηση με την τάση ψηφιακού μετασχηματισμού στον τραπεζικό κλάδο. Η εν εξελίξει έρευνα στοχεύει να εφαρμόσει και να βελτιώσει την προτεινόμενη λύση σε περιβάλλον παραγωγής, παρακολουθώντας και συγκρίνοντας μοντέλα πρόβλεψης, λαμβάνοντας υπόψη μη τυπικά χαρακτηριστικά όπως η κοινωνική επιρροή και διερευνώντας βέλτιστες στρατηγικές διατήρησης για διαφορετικές ομάδες πελατών

Τηλεπικοινωνίες

Άλλη μία κλασσική περίπτωση πελατειακής σχέσης που διέπεται από κάποιο συμβόλαιο. Πρόκειται πιθανότατα για τον κλάδο που αφορούν οι περισσότερες διαθέσιμες δημοσιεύσεις αναφορικά με την πρόβλεψη μελλοντικής απώλειας πελατών. Η διαφορά με τον κλάδο της τραπεζικής είναι πως εδώ απώλεια σημαίνει ρητά πλήρη λήξη της πελατειακής σχέσης. Ο πελάτης ενός παρόχου τηλεπικοινωνιών θα πρέπει να λήξει το συμβόλαιό του πριν απευθυνθεί σε έναν ανταγωνιστή πάροχο.

Η δημοσίευση [11] ασχολείται με τη σύγκριση πέντε δημοφιλών αλγορίθμων μηχανικής μάθησης – συγκεκριμένα τεχνητά νευρωνικά δίκτυα πολλαπλών επιπέδων, δέντρα αποφάσεων, μηχανές διανυσμάτων υποστήριξης (SVM), αφελής ταξινομητής bayes (naïve bayes) και λογιστική παλινδρόμηση - για την πρόβλεψη απώλειας πελατών στον κλάδο των τηλεπικοινωνιών χρησιμοποιώντας ένα δημόσια διαθέσιμο σύνολο δεδομένων από το UCL. Η μελέτη χρησιμοποιεί προσομοιώσεις Monte Carlo για εκτεταμένη βελτιστοποίηση (tuning) παραμέτρων μέσω διασταυρούμενης επικύρωσης (cross validation). Αναδεικνύει τελικά τις μηχανές διανυσμάτων υποστήριξης (SVM) με ενίσχυση (boosting) ως την πιο αποτελεσματική προσέγγιση, βάσει των μετρικών αξιολόγησης που προκύπτουν από κάθε μέθοδο.

Στο [12] επιχειρείται η σύγκριση της απόδοσης κλασσικών αλγορίθμων μηχανικής μάθησης (δέντρα απόφασης, δίκτυο bayes, αφελής ταξινομητής bayes και multilayer perceptrons) και αρχιτεκτονικών βαθιάς μάθησης (συνελικτικά νευρωνικά δίκτυα - CNN, επαναλαμβανόμενα νευρωνικά δίκτυα - RNN) σε 2 γνωστά σύνολα δεδομένων από των κλάδο των τηλεπικοινωνιών. Όλες οι δοκιμές εκτελούνται μέσω του εργαλείου MATLAB. Οι συγγραφείς καταλήγουν σε καλύτερη απόδοση των συνελικτικών νευρωνικών δικτύων στην πρόβλεψη αποχώρησης πελατών, έναντι των υπολοίπων μεθόδων. Το εργαλείο MATLAB χρησιμοποιείται και στο [13]. Εδώ εισάγεται ο IBA (βελτιωμένος αλγόριθμος BAT) ο οποίος βελτιστοποιεί την απόδοση του μοντέλου ELM (αλγόριθμος μηχανικής μάθησης της κατηγορίας των νευρωνικών δικτύων).

Στο [14] χρησιμοποιείται και πάλι η κλασσική προσέγγιση του δέντρου απόφασης. Η διαφορά εδώ είναι πως οι συγγραφείς ενσωματώνουν στο δέντρο που κατασκευάζουν τις έννοιες του κόστους και του κέρδους διότι, όπως αναφέρουν, δεν αρκεί να εντοπίσουμε απλά όλους του πιθανούς «churners». Πιο σημαντικό είναι να εντοπιστούν οι πιθανοί «churners» που έχουν μεγαλύτερη αξία για μια επιχείρηση και συνεπώς αξίζει η προσπάθεια για τη

διατήρησή τους. Η παράμετρος του κέρδους στη διαδικασία της εν λόγω πρόβλεψης ενσωματώνεται επίσης σε μοντέλο λογιστικής παλινδρόμησης στο [15], των ίδιων συγγραφέων.

Οι συγγραφείς του [16] συνδυάζουν τους bagging και boosting αλγορίθμους και εισάγουν έτσι 3 νέους συνδυαστικούς αλγορίθμους: BaBag (bagged bagging), BoBag (boosted bagging) and BNNGA (bagging νευρωνικού δικτύου με μάθηση βασισμένη σε γενικό αλγόριθμο) που αυξάνουν σημαντικά την απόδοση βασικών αλγορίθμων ταξινόμησης.

Στην εργασία [17] οι συγγραφείς επικεντρώνονται στο αρχικό πρόβλημα της επιλογής ενός συνόλου ταξινομητών για την πρόβλεψη της αποχώρησης πελατών σε μια εταιρεία τηλεπικοινωνιών. Υπογραμμίζουν την ανάγκη ο αριθμός αυτός να είναι σχετικά μικρός προκειμένου τα αποτελέσματα να είναι εξηγήσιμα. Χρησιμοποιούν ένα δυαδικό σύνολο δεδομένων, το χωρίζουν σε δεδομένα εκπαίδευσης και δοκιμής και χρησιμοποιούν διασταυρούμενη επικύρωση με k-folds για την αξιολόγηση. Στη μελέτη εξετάζονται 28 ταξινομητές που περιέχονται στο εργαλείο Weka και ως κύρια μετρική αξιολόγησης χρησιμοποιείται ο συντελεστής συσχέτισης Matthews (MCC - Matthews Correlation Coefficient). Στα συμπεράσματα, επισημαίνεται η αποτελεσματικότητα της προτεινόμενης μεθόδου, ειδικότερα χρησιμοποιώντας τον εξελικτικό αλγόριθμο MO-EoC.

Στο [18] χρησιμοποιείται ένα σύνολο δεδομένων πελατών από μια γαλλική εταιρεία τηλεπικοινωνιών. Αυτό υποβάλλεται σε προεπεξεργασία χρησιμοποιώντας ασαφείς (fuzzy) μεθόδους συσταδοποίησης (FCM, PCM, PFCM). Έπειτα οι ομαδοποιημένοι πελάτες χωρίζονται σε σύνολα εκπαίδευσης και δοκιμής που χρησιμοποιούνται για την εκπαίδευση συνδυαστικών μοντέλων ταξινόμησης (bagging, boosting, random rubspace), με το boosting και το possibility fuzzy c-means (PFCM) να υπερτερούν των άλλων.

Το [19] εστιάζει στην ανάλυση μεγάλων δεδομένων, όπως το ιστορικό τηλεφωνικών κλήσεων, το ιστορικό διαδικτυακών συνεδριών, το προφίλ του πελάτη και τα στοιχεία κοινωνικής δικτύωσής του χρησιμοποιώντας μια διαδικασία απόκτησης-μετασχηματισμού-φόρτωσης (ETL) και αποθήκευση με το σύστημα αποθήκευσης Hadoop (HDFS). Τα χαρακτηριστικά εξετάζονται χρησιμοποιώντας Hive/Spark SQL και εφαρμόζεται συσταδοποίηση για τον εντοπισμό ομάδων πελατών με υψηλά ποσοστά αποχώρησης. Σύμφωνα με τους συγγραφείς η εργασία παρουσιάζει τρεις βασικές συνεισφορές. Πρώτον, την εισαγωγή της SDSCM, μιας νέας μεθόδου συσταδοποίησης που ενισχύει την ακρίβεια και μετριάζει τους λειτουργικούς κινδύνους. Δεύτερον, την προσαρμογή της σε μεγάλα δεδομένα και την εφαρμογή της μέσω του πλαισίου MapReduce. Τέλος την εφαρμογή της νέας μεθόδου καθώς και του αλγορίθμου k-means σε μεγάλα δεδομένα για την επίλυση του προβλήματος της πρόβλεψης μελλοντικής απώλειας πελατών σε μία μεγάλη κινεζική εταιρεία τηλεπικοινωνιών.

Η εργασία [20] αποτελεί μία ακόμα προσέγγιση μεγάλων δεδομένων στον κλάδο των τηλεπικοινωνιών. Σε αυτή χρησιμοποιείται ένα συνελικτικό δίκτυο 2 διαστάσεων (2D CNN). Το Apache Spark χρησιμοποιείται για παράλληλη επεξεργασία. Το σύνολο δεδομένων που χρησιμοποιείται είναι το Telco Customer Churn από το Kaggle [21]. Τα αποτελέσματα δείχνουν υψηλή βαθμολογία ακρίβειας > 90%. Γενικά η εργασία δίνει έμφαση στη σχολαστική προεπεξεργασία για το μοντέλο CNN, προσφέροντας πολύτιμες γνώσεις για την επιτυχή πρόβλεψη απώλειας πελατών στο πλαίσιο του ηλεκτρονικού επιχειρείν.

Η δημοσίευση [22] αναφέρεται στην περίπτωση όπου κατασκευάζεται μοντέλο με τα δεδομένα μίας μεγάλης εταιρείας τηλεπικοινωνιών το οποίο μπορεί μετά να χρησιμοποιηθεί για την εξαγωγή προβλέψεων σε δεδομένα μικρών εταιρειών που δεν έχουν αρκετά δεδομένα

για εκπαίδευση μοντέλου. Τα διεταιρικά (cross-company) μοντέλα πρόβλεψης απώλειας πελατών αποτελούν, σύμφωνα με τους συγγραφείς, μεγάλη πρόκληση λόγω της ανομοιογένειας κάθε φορά των δεδομένων σε διαφορετικές εταιρείες. Προς αυτή την κατεύθυνση εφαρμόζουν μετασχηματισμούς ώστε να μπορούν όλα τα δεδομένα να εφαρμόσουν στο μοντέλο πρόβλεψης. Πέραν των μετασχηματισμών, συγκρίνουν την απόδοση διάφορων βασικών αλγορίθμων ταξινόμησης, σε δημόσια προσβάσημα σχετικά σύνολα δεδομένων, κρατώντας κυρίως τις προκαθορισμένες παραμέτρους του εργαλείου Rapidminer.

Στο [23],από την άλλη, παρουσιάζεται μία διαφορετική προσέγγιση, ο αλγόριθμος CCPBI-TAMO, για δυναμική πρόβλεψη της πιθανής απώλειας πελατών, με εστίαση και πάλι στον κλάδο των τηλεπικοινωνιών. Χρησιμοποιώντας ανάλυση κειμένου και αλγορίθμους βελτιστοποίησης γίνεται διάκριση μεταξύ των «churners» και των «non-churners». Η τεχνική CPIO-FS απλοποιεί την επιλογή χαρακτηριστικών, ενώ το μοντέλο LSTM-SAE ταξινομεί αποτελεσματικά τα δεδομένα. Η τεχνική Sunflower (SFO) ενισχύει την απόδοση του μοντέλου μέσω της βελτιστοποίησης των παραμέτρων. Η προσομοίωση με διαφορετικά σύνολα δεδομένων δίνει υψηλή ακρίβεια πρόβλεψης, άνω του 92% σε όλες τις περιπτώσεις.

Άξια αναφοράς είναι και η εργασία [24]. Σε αυτή γίνεται αξιοποίηση της θεωρίας των κοινωνικών δικτύων για την αποτελεσματικότερη πρόβλεψη στον κλάδο των τηλεπικοινωνιών. Όπως αναφέρεται η χρήση στοιχείων κοινωνικής δικτύωσης των πελατών είναι πολύ χρήσιμη καθώς πελάτες που συνδέονται στενά με πελάτες που έχουν αποχωρήσει έχουν μεγαλύτερη πιθανότητα να αποχωρήσουν και οι ίδιοι, σύμφωνα με τη βασική αρχή της θεωρίας κοινωνικών δικτύων ότι η επαφή ανάμεσα σε ανθρώπους με κοινά χαρακτηριστικά είναι πολύ συχνότερη. Το φαινόμενο αυτό είναι γνωστό στη βιβλιογραφία ως ομοφυλία.

<u>Υπηρεσίες από Επιχείρηση προς Επιχείρηση (B2B)</u>

Στο [25] γίνεται πρόβλεψη απώλειας πελατών στα δεδομένα μίας μεγάλης εταιρείας που παρέχει υπηρεσίες σε άλλες επιχειρήσεις (B2B – Business to Business). Κάθε πελάτης αντιμετωπίζεται ως μια ξεχωριστή χρονοσειρά στην οποία φαίνεται για κάθε μονάδα χρόνου ο αριθμός των υπηρεσιών που αγόρασε. Ως πιθανοί αποχωριτές ορίζονται αυτοί που η χρονοσειρά τους έχει κάποια συγκεκριμένα χαρακτηριστικά. Βασικό εύρημα του άρθρου είναι πως μεγαλύτερη πιθανότητα αποχώρησης έχουν οι πελάτες που αντιμετώπισαν προβλήματα στις υπηρεσίες που αγόρασαν για αυτό και εταιρείες του κλάδου πρέπει να είναι συνεχώς σε εγρήγορση και να επανορθώνουν έγκαιρα τέτοιου είδους αστοχίες.

Το [26] αναφέρεται επίσης στην περίπτωση του B2B. Εδώ γίνεται σαφής διάκριση από τις εταιρείες που παρέχουν προϊόντα ή υπηρεσίες απευθείας στους καταναλωτές (B2C - Business to Customer). Η επιλογή του σε ποιους πελάτες θα γίνει εστίαση για τη διατήρησή τους είναι μεγαλύτερης σημασίας σε μία B2B εταιρεία καθώς τα κόστη διατήρησης είναι πολύ μεγαλύτερα. Συνεπώς στη συγκεκριμένη εργασία ο στόχος δεν είναι απλά να βρεθούν οι πελάτες της επιχείρησης που έχουν μεγάλη πιθανότητα να αποχωρήσουν. Πρέπει επιπρόσθετα να βρεθεί το υποσύνολο αυτών για τους οποίους μια εκστρατεία διατήρησης έχει πιθανότητες επιτυχίας, διαφυλάσσοντας έτσι τους πόρους της επιχείρησης. Γενικά κατά τη βιβλιογραφική τους έρευνα οι συγγραφείς συμπεραίνουν ότι μέχρι τώρα έχει δοθεί ερευνητικά μεγαλύτερη έμφαση σε σχετικές προβλέψεις που αφορούν εταιρείες B2C.

Λοιποί Κλάδοι

Στο [27] περιγράφεται ένα πιο συνολικό σύστημα μεγάλων δεδομένων που μπορεί να υποστηρίζει όχι μόνο εργασίες πρόβλεψης απώλειας πελατών αλλά και άλλα ήδη αποφάσεων που βασίζονται σε δεδομένα. Τα δεδομένα που χρησιμοποιούνται αφορούν πελάτες συνδρομητικών διαδικτυακών τηλεοπτικών υπηρεσιών (video streaming). Το σύστημα συνδυάζει το Spark και Hadoop clusters. Μεγάλη σημασία δίνεται στην επεξεργασία των δεδομένων, τα οποία προέρχονται από διαφορετικές πηγές. Επίσης τα σύνολα δεδομένων εκπαίδευσης και επικύρωσης επαναπροσδιορίζονται κάθε 2 εβδομάδες.

Για τον κλάδο των διαδικτυακών παιχνιδιών (online gaming) παίρνουμε κάποιες ενδιαφέρουσες πληροφορίες από το [28]. Στα διαδικτυακά παιχνίδια απαιτείται πάντα η τμηματοποίηση καθώς υπάρχουν πολλοί χρήστες χωρίς ιδιαίτερη αξία για το προϊόν πχ δωρεάν, ανενεργοί. Ως «churners» ορίζονται από τους συγγραφείς οι χρήστες που είναι ανενεργοί για διάστημα μεγαλύτερο εκείνου που μεσολαβεί ανάμεσα σε 2 ανανεώσεις περιεχομένου του παιχνιδιού. Αν ένας χρήστης δεν δείξει κανένα ενδιαφέρον ως προς αυτό θεωρείται χαμένος. Την πρόβλεψη μελλοντικών αποχωρήσεων χρησιμοποιούνται οι μακροχρόνιοι χρήστες με τη μεγαλύτερη χρηματική αξία και στο μοντέλο ενσωματώνεται η έννοια του κέρδους.

Οι κλάδοι όπου μπορεί να γίνει η παραπάνω πρόβλεψη είναι φυσικά πολύ περισσότεροι. Στη βιβλιογραφία συναντάμε ακόμη εργασίες που αφορούν χρήστες σε ηλεκτρονικές βιβλιοθήκες [29], μέλη σε ιδιωτικές λέσχες [30] συνδρομές σε εφημερίδες [31], εταιρείες πώλησης επενδυτικών προϊόντων [32], ακόμη και πλατφόρμες διαδικτυακών γνωριμιών [3]. Κάθε κλάδος έχει τις ιδιαιτερότητές του αλλά ο πυρήνας της διαδικασίας παραμένει ο ίδιος: συλλογή δεδομένων πελατών, εντοπισμός πελατών με μεγάλη πιθανότητα να αποχωρήσουν, επιλογή εκείνων των οποίων η αποχώρηση είναι περισσότερο επιζήμια για τον οργανισμό και, τέλος, στόχευση με κάποια προωθητική ενέργεια με σκοπό τη διατήρηση.

2.2 Επισκόπηση των Μεθόδων που χρησιμοποιούνται στην Εργασία

2.2.1 Μέθοδοι Μηχανικής Μάθησης για Ταξινόμηση

Το πρόβλημα της κατάταξης πελατών σε churners και μη churners αποτελεί ουσιαστικά ένα πρόβλημα δυαδικής ταξινόμησης. Στη βιβλιογραφία υπάρχει πληθώρα αλγορίθμων για την κατασκευή μοντέλων μηχανικής μάθησης για εργασίες τέτοιου είδους. Παρακάτω αναλύονται συνοπτικά οι αλγόριθμοι που δοκιμάζονται στο πρακτικό μέρος της παρούσας εργασίας. Παρουσιάζονται επιπλέον και κάποιες βοηθητικές τεχνικές που χρησιμοποιούνται στα πλαίσια της κατασκευής μοντέλων για την καλύτερή, τελικά, απόδοσή τους.

Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση είναι μια ευρέως χρησιμοποιούμενη στατιστική μέθοδος για τη μοντελοποίηση της σχέσης μεταξύ μιας δυαδικής εξαρτημένης μεταβλητής - δηλαδή μίας μεταβλητής που παίρνει τιμές 0 ή 1- και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Η τεχνική εισήχθη για πρώτη φορά το 1958 [33] και έκτοτε έχει χρησιμοποιηθεί σε πολλούς τομείς, συμπεριλαμβανομένης της ιατρικής, του μάρκετινγκ και των οικονομικών. Αναλύεται επίσης εκτενώς σε πολλά βιβλία και επιστημονικές δημοσιεύσεις όπως στο [34].

Έχει σκοπό παρόμοιο με εκείνον της κλασικής Γραμμικής Παλινδρόμησης με τη διαφορά ότι στην περίπτωση της Λογιστικής Παλινδρόμησης η εξαρτημένη μεταβλητή

είναι κατηγορική και όχι ποσοτική. Αν και μπορεί να χρησιμοποιηθεί στο πλαίσιο της μηχανικής μάθησης, δεν είναι αυστηρά μέθοδος μηχανικής μάθησης, καθώς βασίζεται κυρίως σε στατιστικές μεθόδους.

Το μοντέλο λογιστικής παλινδρόμησης υποθέτει ότι ο φυσικός λογάριθμος της πιθανότητας η δυαδική τιμή να πάρει την τιμή 1 είναι γραμμική συνάρτηση των ανεξάρτητων μεταβλητών. Μαθηματικά αυτό μπορεί να γραφτεί ως:

$$l_n(p) = \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta n\, xn + \varepsilon$$

όπου β_0 είναι σταθερά και $\beta_1,\beta_2\ldots\beta_n$ οι συντελεστές των ανεξάρτητων μεταβλητών x_1, x_2, ..., x_n αντίστοιχα. Για να εκτιμηθούν οι τιμές των συντελεστών β_0, β_1, β_2, ..., β_n, συνήθως χρησιμοποιείται η μέθοδος της μέγιστης πιθανοφάνειας ή η μέθοδος των ελαχίστων τετραγώνων. Η πιθανότητα p υπολογίζεται στη συνέχεια μέσω του τύπου:

$$p = e^{l_n(p)}$$

Όταν το μοντέλο εκπαιδεύεται σε ένα σύνολο δεδομένων εκπαίδευσης, μπορεί να χρησιμοποιηθεί για να κάνει προβλέψεις σε νέες παρατηρήσεις, υπολογίζοντας την πιθανότητα η εξαρτημένη μεταβλητή να είναι 1 βάσει των τιμών των ανεξάρτητων μεταβλητών. Η απόφαση σχετικά με το αν η εξαρτημένη μεταβλητή θα έχει τιμή 0 ή 1, γίνεται μέσω ενός κατωφλίου για τις πιθανότητες που ορίζεται από τον χρήστη ή από τις απαιτήσεις του προβλήματος.

Μερικά πλεονεκτήματα της μεθόδου σε σύγκριση με άλλες μεθόδους στατιστικής ή μηχανικής μάθησης είναι τα ακόλουθα:

- είναι μια απλή και κατανοητή μέθοδος που μπορεί να εφαρμοστεί σε ένα ευρύ φάσμα προβλημάτων
- απαιτεί σχετικά λίγη προετοιμασία των δεδομένων και μπορεί να χειριστεί τόσο κατηγορικές όσο και συνεχείς μεταβλητές πρόβλεψης
- η προβλεπόμενη πιθανότητα η εξαρτημένη μεταβλητή να βρίσκεται σε μία από τις δύο κατηγορίες μπορεί να χρησιμοποιηθεί για τη λήψη τεκμηριωμένων αποφάσεων.
- μπορεί εύκολα να επεκταθεί για να χειριστεί προβλήματα ταξινόμησης πολλαπλών κλάσεων

Παρουσιάζει ωστόσο και κάποια μειονεκτήματα, όπως:

- υποθέτει ότι η σχέση μεταξύ των εξαρτημένων μεταβλητών και της ανεξάρτητης μεταβλητής είναι γραμμική. Αυτό μπορεί να μην ισχύει σε ορισμένες περιπτώσεις και μπορεί να απαιτούνται πιο πολύπλοκα μοντέλα
- υποθέτει ότι οι μεταβλητές πρόβλεψης είναι ανεξάρτητες η μία από την άλλη. Εάν υπάρχει συσχέτιση μεταξύ τους, οι εκτιμήσεις των συντελεστών παλινδρόμησης μπορεί να είναι αναξιόπιστες.
- προϋποθέτει ότι η εξαρτημένη μεταβλητή είναι δυαδική ή κατηγορική. Δεν μπορεί να χρησιμοποιηθεί για τη μοντελοποίηση συνεχών μεταβλητών.
- είναι επιρρεπής στην υπερπροσαρμογή στα δεδομένα εκπαίδευσης (overfitting) εάν το μοντέλο είναι πολύ περίπλοκο σε σχέση με τον όγκο των διαθέσιμων δεδομένων – αν δηλαδή έχουμε περισσότερες ανεξάρτητες μεταβλητές από ότι αριθμό παρατηρήσεων στο δείγμα.

Σε σύγκριση με άλλες μεθόδους μηχανικής μάθησης, η λογιστική παλινδρόμηση είναι γενικά λιγότερο ισχυρή όσον αφορά την ακρίβεια πρόβλεψης, αλλά συχνά είναι ευκολότερη στην ερμηνεία και έχει χαμηλότερες υπολογιστικές απαιτήσεις.

Δέντρα Απόφασης

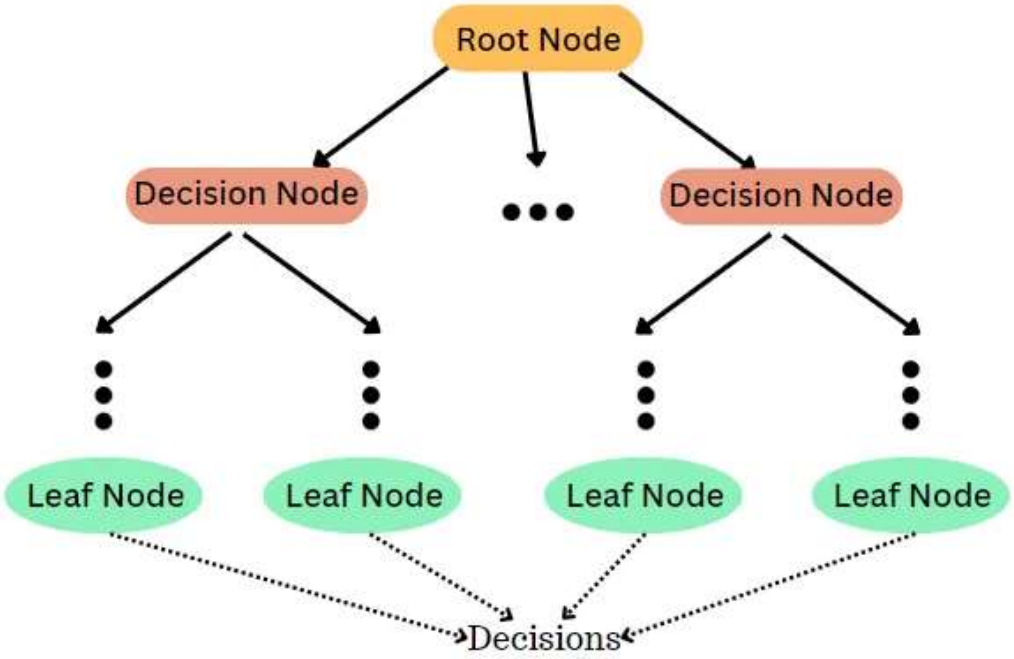
Τα δέντρα απόφασης είναι μια ευρέως χρησιμοποιούμενη μέθοδος εποπτευόμενης μηχανικής μάθησης. Είναι ιδιαίτερα αποτελεσματική τόσο για εργασίες ταξινόμησης όσο και για εργασίες παλινδρόμησης. Ως έννοια συναντάται για πρώτη φορά τη δεκαετία του 1980 στο [35], όπου εισάγεται για πρώτη φορά ο αλγόριθμος κατασκευής δέντρων ID3. Έκτοτε έχουν εξελιχθεί πολλοί αλγόριθμοι κατασκευής δέντρων αποφάσεων, όπως ο CART και ο C4.5.

Ένας τέτοιος αλγόριθμος λειτουργεί διαχωρίζοντας αναδρομικά ένα σύνολο δεδομένων εκπαίδευσης σε υποσύνολα με βάση τις τιμές των χαρακτηριστικών εισόδου, οδηγώντας τελικά στη δημιουργία μιας δομής που μοιάζει με δέντρο. Σε κάθε κόμβο του δέντρου λαμβάνεται μια απόφαση με βάση την τιμή ενός συγκεκριμένου χαρακτηριστικού. Η απόφαση καθορίζεται με τη βελτιστοποίηση ενός κριτηρίου όπως, για παράδειγμα, η ελαχιστοποίηση της εντροπίας των υποσυνόλων που προκύπτουν. Εντροπία (H) είναι ένα μέτρο ακαθαρσίας ή αταξίας σε ένα σύνολο και ορίζεται ως:

$$\mathrm{H(S)} = -\sum_{i=1}^{c} pi \log_2(pi)$$

όπου p_i είναι η αναλογία των δειγμάτων που ανήκουν στην κατηγορία i στο σύνολο S. Η απόφαση είναι συνήθως δυαδική, με αποτέλεσμα τον διαχωρισμό του συνόλου δεδομένων σε δύο υποσύνολα. Αυτή η διαδικασία επαναλαμβάνεται μέχρι να ικανοποιηθεί ένα κριτήριο διακοπής, όπως ένα μέγιστο βάθος δέντρου ή ένας ελάχιστος αριθμός δειγμάτων σε έναν κόμβο.

Η διαδικασία κατασκευής ενός δέντρου απόφασης στοχεύει στη μεγιστοποίηση της ομοιογένειας των κλάσεων στα υποσύνολα που προκύπτουν. Η τελική απόφαση λαμβάνεται διασχίζοντας το δέντρο από τη ρίζα του (root node) έως ένα φύλλο του (leaf node). Η κλάση που επικρατεί στον αντίστοιχο κόμβο γίνεται και η τελικά προβλεπόμενη.



Εικόνα 1-Σχηματική Αναπαράσταση Δέντρου Απόφασης [36]

Παρά την τάση τους να προσαρμόζονται υπερβολικά στα δεδομένα εκπαίδευσης (overfitting), τα δέντρα αποφάσεων συχνά αποδίδουν καλά στην πράξη. Τεχνικές όπως το κλάδεμα (pruning), ο περιορισμός του βάθους του δέντρου ή η απαίτηση για ύπαρξη ενός ελάχιστου αριθμού δειγμάτων ανά τελικό κόμβο μπορούν να βοηθήσουν στον μετριασμό της υπερπροσαρμογής. Ένα ακόμα πλεονέκτημα των δέντρων απόφασης, συγκριτικά με άλλες μεθόδους ταξινόμησης, είναι πως είναι εύκολα ερμηνεύσιμα καθώς μπορούν να παρασταθούν οπτικά.

Αφελής Ταξινομητής Bayes (Naïve Bayes)

Ο αφελής ταξινομητής Bayes (naïve Bayes) είναι ένα δημοφιλές μοντέλο μηχανικής μάθησης που κατατάσσεται στην κατηγορία πιθανοτικών ταξινομητών. Χρησιμοποιείται ευρέως για διάφορες εργασίες, όπως ο εντοπισμός ανεπιθύμητης αλληλογραφίας, η ανάλυση συναισθημάτων και η κατηγοριοποίηση εγγράφων. Ο αλγόριθμος βασίζεται στο θεώρημα του Bayes, το οποίο αποτελεί θεμελιώδη αρχή της θεωρίας πιθανοτήτων.

Η βασική ιδέα πίσω από το naïve Bayes είναι να κάνει προβλέψεις υπολογίζοντας την πιθανότητα μιας συγκεκριμένης κλάσης δεδομένων των χαρακτηριστικών εισόδου. Ας συμβολίσουμε την κλάση ως C και τα χαρακτηριστικά ως x_1, x_2, ..., x_n. Σύμφωνα με το θεώρημα του Bayes, η πιθανότητα μιας κλάσης δεδομένων των χαρακτηριστικών μπορεί να εκφραστεί ως:

$$P(C|x1, x2, \dots, xn) = \frac{P(x1, x2, \dots, xn|C)P(C)}{P(x1, x2, \dots, xn)}$$

Η "αφελής" υπόθεση είναι ότι τα χαρακτηριστικά είναι υπό συνθήκη ανεξάρτητα δεδομένης της κλάσης. Αυτό απλοποιεί την έκφραση σε:

$$P(C|x1, x2, \dots, xn) \propto P(C) \prod\nolimits_{i=1}^{n} P(xi|C)$$

Το μοντέλο υπολογίζει την πιθανότητα P(xi|C) και την προηγούμενη πιθανότητα P(C) από τα δεδομένα εκπαίδευσης. Η τελική πρόβλεψη γίνεται επιλέγοντας την κλάση με τη μεγαλύτερη πιθανότητα. Υπάρχουν διάφορες εκδοχές των ταξινομητών naïve Bayes, ανάλογα με τις ιδιαιτερότητες κάθε προβλήματος, όπως Gaussian naïve Bayes, Πολυωνυμικός (Multinomial) naïve Bayes και Bernoulli naïve Bayes. Παρότι οι υποθέσεις του μπορεί να φαίνονται υπερβολικά απλουστευτικές, συχνά παρουσιάζει καλή απόδοση στην πράξη και αποτελεί βάση για πιο σύνθετα μοντέλα.

Μηχανές Διανυσμάτων Υποστήριξης (SVM)

Οι μηχανές διανυσμάτων υποστήριξης (Support Vector Machines-SVM) είναι ισχυροί και ευέλικτοι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιούνται ευρέως για εργασίες ταξινόμησης και παλινδρόμησης. Προτάθηκαν το 1963 ως γραμμικοί ταξινομητές, ωστόσο άρχισαν να χρησιμοποιούνται ευρέως τη δεκαετία του 90, όταν ενισχύθηκαν, από τους αρχικούς εμπνευστές τους [37], με το κόλπο του πυρήνα (kernel trick), που επέτρεψε την εφαρμογή τους και σε μη γραμμικώς διαχωρίσιμα προβλήματα.

Ο κύριος στόχος των SVM είναι να βρεθεί μία υπερεπιφάνεια που να διαχωρίζει αποτελεσματικά τα δεδομένα σε διάφορες κατηγορίες ενώ ταυτόχρονα να μεγιστοποιεί την απόσταση μεταξύ τους.

Η συνάρτηση απόφασης ενός γραμμικού SVM εκφράζεται ως:

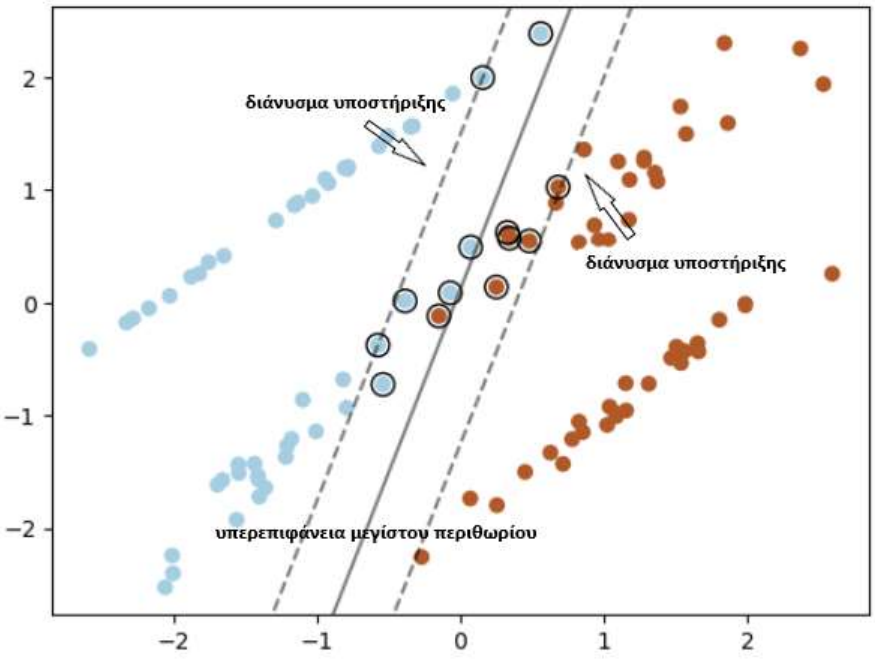
$$f(X) = sign(xX + b)$$

όπου w είναι το διάνυσμα των βαρών, X το διάνυσμα χαρακτηριστικών, b ο όρος προσαρμογής, και sign η συνάρτηση επιστροφής προσήμου. Η βέλτιστη υπερεπιφάνεια καθορίζεται από την επίλυση του προβλήματος ελαχιστοποίησης:

$$\min_{w,b} \frac{1}{2}||w||^2 \; \upsilon\pi\acute{o} \; \tau\eta\nu \; \pi\rho o\ddot{\upsilon}\pi\acute{o}\theta\varepsilon\sigma\eta \; Yi(wXi + b) \geq 1 \; \gamma\iota\alpha \; i = 1, \ldots, N$$

Αυτή η συνάρτηση στοχεύει στο να ελαχιστοποιήσει τη νόρμα του διανύσματος βαρών (w) ενώ εξασφαλίζει παράλληλα ότι κάθε σημείο δεδομένων (X_i) ταξινομείται σωστά με ένα περιθώριο μεγαλύτερο ή ίσο του 1.

Ο σκοπός των μηχανών διανυσμάτων υποστήριξης είναι να εντοπίσουν τη γραμμή που απέχει όσο το δυνατόν περισσότερο από τα παραδείγματα των διαφορετικών κλάσεων. Στην εικόνα πρόκειται για την ευθεία ανάμεσα στις διακεκομμένες γραμμές. Αυτή η γραμμή, γνωστή ως "σύνορο μέγιστου περιθωρίου," ορίζεται σε γραμμικώς διαχωρίσιμα προβλήματα από έναν πεπερασμένο αριθμό παραδειγμάτων του συνόλου εκπαίδευσης, τα οποία αναφέρονται ως "διανύσματα υποστήριξης." Στην περίπτωση μη γραμμικά διαχωρίσιμων προβλημάτων τα svms μπορούν να μετασχηματίσουν τον αρχικό χώρο υποθέσεων, με τη βοήθεια των συναρτήσεων πυρήνα, έτσι ώστε να μετατραπούν σε προβλήματα γραμμικώς διαχωρίσιμα. Η έννοια της ευθείας, σε μη δισδιάστατους χώρους αντικαθίσταται από εκείνη της υπερεπιφάνειας [38]. Πυρήνες που συνήθως χρησιμοποιούνται είναι ο γραμμικοί, πολυωνυμικοί και ακτινικής συνάρτησης βάσης (Radial Kernel Function-RBF), ανάλογα κάθε φορά με την κατανομή των δεδομένων.

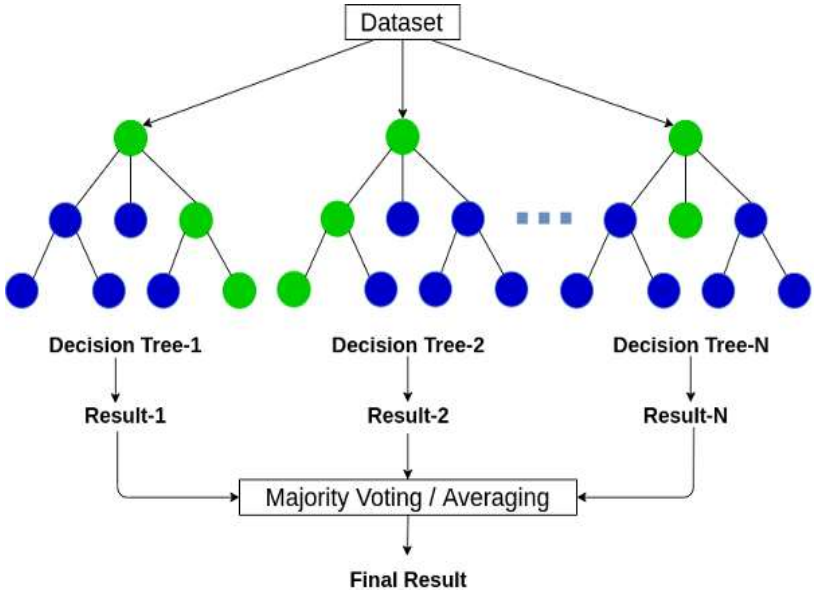


Εικόνα 2-Σχηματική Αναπαράσταση SVM (με matplotlib)

Τυχαία Δάση (Random Forest)

Τα τυχαία δάση ανήκουν στην κατηγορία των συνδυαστικών (ensemble) μεθόδων ταξινόμησης, όπου πολλά μοντέλα συνδυάζονται για να βελτιώσουν την προβλεπτική απόδοση. Συναντώνται πρώτη φορά το 2001 στο [39] και έκτοτε έχουν κερδίσει δημοτικότητα για τη δυνατότητά τους να αντιμετωπίζουν πολύπλοκα σύνολα δεδομένων και να

αντιμετωπίζουν την τάση των μεμονωμένων δέντρων απόφασης στην υπερβολική προσαρμογή στα δεδομένα εκπαίδευσης (overfitting).



Εικόνα 3-Σχηματική Αναπαράσταση Τυχαίου Δάσους [40]

Ένα τυχαίο δάσος αποτελείται από πολλά δέντρα απόφασης τα οποία κατασκευάζονται ανεξάρτητα το ένα από το άλλο. Για την εισαγωγή ποικιλίας, το τυχαίο δάσος χρησιμοποιεί δειγματοληψία μέσα στο ίδιο το δείγμα (bootstrap). Για κάθε δέντρο, επιλέγεται τυχαία ένα υποσύνολο των δεδομένων εκπαίδευσης με επανάληψη. Αυτή η διαδικασία δημιουργεί διαφορετικά σύνολα δεδομένων εκπαίδευσης για κάθε δέντρο, ενισχύοντας τη συνολική ανθεκτικότητα του μοντέλου. Κατά την κατασκευή κάθε δέντρου, λαμβάνεται υπόψη μόνο ένα τυχαίο υποσύνολο χαρακτηριστικών σε κάθε διακλάδωση. Αυτό ενισχύει περαιτέρω την ποικιλία των δέντρων και εμποδίζει την υπεροχή συγκεκριμένων χαρακτηριστικών, συμβάλλοντας σε ένα πιο γενικευμένο μοντέλο.

Οι προβλέψεις των ατομικών δέντρων συνδυάζονται μέσω ενός μηχανισμού ψηφοφορίας. Για την κατηγοριοποίηση, η κατηγορία που λαμβάνει την πλειονότητα των ψήφων γίνεται η τελική πρόβλεψη. Σε προβλήματα παλινδρόμησης, λαμβάνεται ο μέσος όρος των προβλέψεων των ατομικών δέντρων. Έστω T ο αριθμός των δέντρων στο δάσος, K ο αριθμός των κατηγοριών και t ο δείκτης κάθε δέντρου. Η πρόβλεψη για ένα νέο δείγμα X δίνεται από την εξίσωση:

$$\hat{Y}(X) = \text{argmax}_k(\frac{1}{T}\sum_{t=1}^{T} I(Y_t\,(X) = k)$$

όπου $Y_t(X)$ είναι η πρόβλεψη του t-οστού δέντρου, και $I(\cdot)$ είναι η συνάρτηση δείκτη.

Μέθοδοι Ενίσχυσης (Boosting)

Η ενίσχυση (boosting) είναι μια ισχυρή συνδυαστική τεχνική εκμάθησης που ανήκει στην οικογένεια των αλγορίθμων μηχανικής μάθησης. Έχει σχεδιαστεί για να βελτιώνει την ακρίβεια και την απόδοση των αδύναμων μοντέλων ταξινόμησης. Οι αλγόριθμοι ενίσχυσης εκπαιδεύουν επαναληπτικά μια σειρά αδύναμων ταξινομητών και συνδυάζουν τις προβλέψεις τους για να σχηματίσουν ένα ισχυρό και ακριβές μοντέλο.

Στον πυρήνα του, το boosting εστιάζει στη διαδοχική δημιουργία μιας σειράς αδύναμων μοντέλων ταξινόμησης, καθένα από τα οποία δίνει έμφαση στους τομείς όπου τα προηγούμενα μοντέλα είχαν κακή απόδοση. Αυτή η προσαρμοστικότητα είναι βασικό χαρακτηριστικό των μεθόδων ενίσχυσης. Ο πρωταρχικός στόχος είναι να βελτιωθεί η προγνωστική ικανότητα του συνολικού μοντέλου δίνοντας μεγαλύτερη βαρύτητα σε εσφαλμένα ταξινομημένα δείγματα, μαθαίνοντας αποτελεσματικά από τα λάθη και βελτιώνοντας το μοντέλο με κάθε επανάληψη.

Οι ενισχυτικές μέθοδοι αποτελούνται από 3 βασικά μέρη:

1. **Αδύναμοι Ταξινομητές:** μια σειρά από μοντέλα με απόδοση καλύτερη από την τυχαία απόδοση κατηγορίας αλλά όχι αρκετά ισχυρά ώστε να δώσουν ακριβείς προβλέψεις από μόνα τους. Συνήθως πρόκειται για δέντρα απόφασης, γραμμικά μοντέλα κτλ.
2. **Βάρη:** σε κάθε παρατήρηση από τα δεδομένα εκπαίδευσης αποδίδεται ένα βάρος που αντικατοπτρίζει τη σημασία του σε επόμενες επαναλήψεις. Στα λανθασμένα ταξινομημένα δείγματα αποδίδονται υψηλότερα βάρη ώστε να δοθεί σε αυτά μεγαλύτερη έμφαση στα επόμενα μοντέλα.
3. **Συνδυαστικό (Ensemble) Μοντέλο:** το τελικό μοντέλο είναι ένα σύνολο αδύναμων ταξινομητών όπου ο καθένας συμβάλλει στη συνολική πρόβλεψη με βάση την απόδοσή του και το βάρος του

Τα τελευταία χρόνια έχουν αναπτυχθεί αρκετοί αλγόριθμοι ενίσχυσης, με μερικούς από τους πιο δημοφιλείς να είναι οι AdaBoost, Gradient Boosting, XGBoost και LightGBM. Κάθε αλγόριθμος έχει τη δική του προσέγγιση για να ενισχύσει τους αδύναμους ταξινομητές και να συνδυάσει τις προβλέψεις τους.

Ο αλγόριθμος προσαρμοστικής ενίσχυσης (Adaptive Boosting - AdaBoost) εισήχθη το 1997 στο [41] ως ένας από τους πρώτους αλγορίθμους της κατηγορίας. Αποδίδει μεγαλύτερα βάρη στα λανθασμένα ταξινομημένα στιγμιότυπα και συνδυάζει αδύναμους ταξινομητές για να διορθώνει τα λάθη επαναληπτικά. Η ενίσχυση κλίσης (gradient boosting), η οποία εισήχθη στο [42], βασίζεται στην ιδέα της ελαχιστοποίησης μιας συνάρτησης απώλειας με την προσθήκη αδύναμων ταξινομητών με τρόπο που μοιάζει με διαβάθμιση κλίσης (gradient descent). Οι δημοφιλείς υλοποιήσεις της περιλαμβάνουν τους αλγορίθμους ακραίας ενίσχυσης κλίσης (Extreme Gradient Boosting - XGBoost) και ελαφριάς ενίσχυσης κλίσης (Light Gradient Boosting Machine - LightGBM), οι οποίοι εμφανίζονται για πρώτη φορά στα [43] και [44] αντίστοιχα.

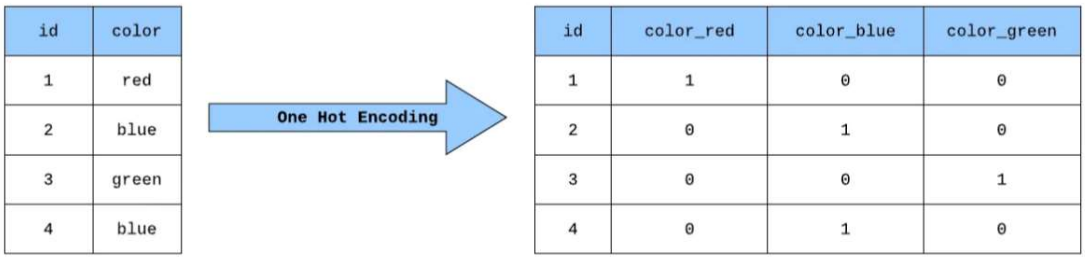
2.2.1 Βοηθητικές Μέθοδοι για τη Μηχανική Μάθηση

Κωδικοποίηση One-hot

Στα πλαίσια της μηχανικής μάθησης, η κωδικοποίηση αποτελεί θεμελιώδη τεχνική για τη μετατροπή κατηγορικών δεδομένων σε μορφή κατάλληλη για αριθμητική ανάλυση και εκπαίδευση μοντέλων. Οι κατηγορικές μεταβλητές μπορεί να είναι διατάξιμες (ordinal), όπως πχ η γνώμη ενός πελάτη για κάποιο προϊόν (κακή-μέτρια-καλή) ή ονομαστικές (nominal), όπου οι διαφορετικές τιμές που μπορεί να πάρουν δεν σχετίζονται μεταξύ τους.

Στα πλαίσια αυτής της εργασίας συναντάμε nominal μεταβλητές. Για αυτές εφαρμόζεται η κωδικοποίηση one-hot (one-hot encoding). Η διαδικασία περιλαμβάνει τη μετατροπή κάθε τιμής που μπορεί να πάρει η κατηγορική μεταβλητή σε μια δυαδική διανυσματική

αναπαράσταση, δημιουργώντας έναν πίνακα "one-hot" όπου κάθε τιμή αντιστοιχεί σε μια μοναδική στήλη. Για ένα δεδομένο χαρακτηριστικό με κατηγορικές τιμές, εάν μια παρατήρηση ανήκει σε μια συγκεκριμένη τιμή από αυτές, η τιμή της αντίστοιχης στήλης ορίζεται σε 1 και όλες οι άλλες ορίζονται στο 0. Η διαδικασία γίνεται πιο κατανοητή στην εικόνα που ακολουθεί [45], όπου η nominal μεταβλητή 'color' μετατρέπεται με τη μέθοδο one-hot σε 3 νέες δυαδικές μεταβλητές, όσες και οι μοναδικές τιμές (red, blue, green) που μπορούσε να πάρει η αρχική μεταβλητή.



Εικόνα 4-Παράδειγμα Κωδικοποίησης One-Hot [45]

Η κωδικοποίηση one-hot αποτελεί ένα βασικό βήμα προεπεξεργασίας στους αγωγούς μηχανικής εκμάθησης, διευκολύνοντας την ενσωμάτωση κατηγορικών πληροφοριών σε μοντέλα σχεδιασμένα να λειτουργούν με αριθμητικά δεδομένα. Αν και χειρίζεται αποτελεσματικά κατηγορικά δεδομένα, οι επαγγελματίες θα πρέπει ωστόσο να προσέχουν την πιθανή αύξηση των διαστάσεων, ειδικά με έναν μεγάλο αριθμό μοναδικών τιμών που μπορεί να πάρει μια κατηγορική μεταβλητή.

Βελτιστοποίηση Παραμέτρων (Tuning)

Η επιλογή παραμέτρων με αναζήτηση πλέγματος (grid search) είναι μια συστηματική μέθοδος που χρησιμοποιείται στη μηχανική μάθηση για τον προσδιορισμό και τον βελτιστοποίηση των παραμέτρων ενός μοντέλου. Οι παράμετροι καθορίζονται πριν από την εκπαίδευση ενός μοντέλου και δεν επηρεάζονται από τα δεδομένα εκπαίδευσης. Παραδείγματα περιλαμβάνουν το ρυθμό μάθησης σε έναν αλγόριθμο ενίσχυσης κλίσης, το μέγιστο βάθος σε ένα δέντρο απόφασης, τον αριθμό των χρησιμοποιούμενων δέντρων σε ένα μοντέλο τυχαίου δάσους, τη συνάρτηση πυρήνα σε μία μηχανή διανυσμάτων υποστήριξης ή τον αριθμό των στρωμάτων σε ένα νευρωνικό δίκτυο.

Η μέθοδος αναζήτησης πλέγματος ξεκινά με τον καθορισμό ενός εύρους τιμών για κάθε παράμετρο. Ο αλγόριθμος αξιολογεί συστηματικά την απόδοση του μοντέλου για κάθε δυνατό συνδυασμό των τιμών αυτών, βάσει κάποιας μετρικής απόδοσης, όπως ακρίβεια ταξινόμησης, f1 score και άλλοι, ανάλογα με τον τύπο του προβλήματος. Ο συνδυασμός που οδηγεί στην καλύτερη επίδοση, σύμφωνα με την επιλεγμένη μετρική, θεωρείται η βέλτιστη ρύθμιση, και οι τιμές του είναι αυτές που τελικά επιλέγονται για την εκπαίδευση του μοντέλου.

Μείωση Διαστάσεων (Ανάλυση Κυρίων Συνιστωσών)

Η ανάλυση κυρίων συνιστωσών (Principal Component Analysis - PCA) είναι στατιστική μέθοδος που στοχεύει στη μείωση των διαστάσεων και την εξαγωγή χαρακτηριστικών σε σύνολα δεδομένων πολλών μεταβλητών. Χρησιμοποιώντας εργαλεία από τη γραμμική άλγεβρα επιδιώκει να μετατρέψει τις αρχικές μεταβλητές ενός συνόλου δεδομένων σε ένα νέο σύνολο ασύνδετων μεταβλητών που ονομάζονται κύριες συνιστώσες. Αυτός ο μετασχηματισμός επιτυγχάνεται με τον υπολογισμό των ιδιοδιανυσμάτων και των ιδιοτιμών

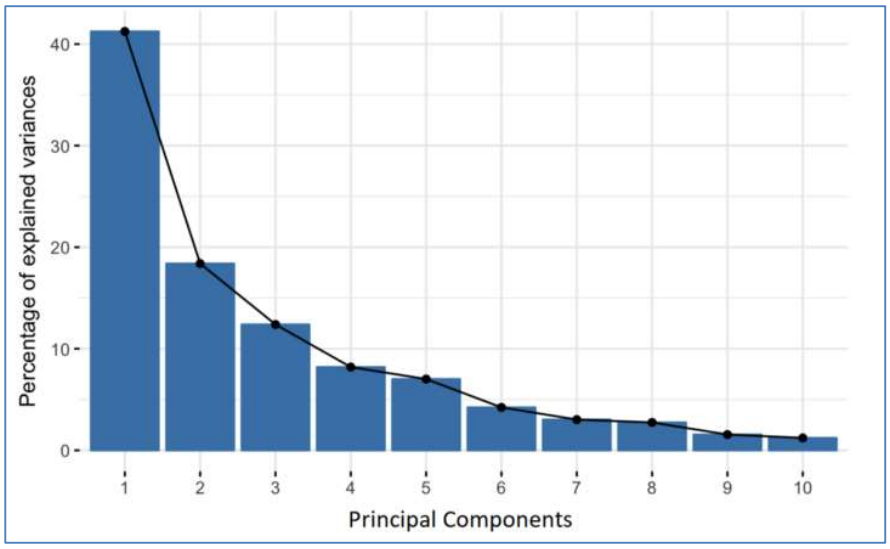
του πίνακα συνδιακύμανσης των αρχικών δεδομένων. Η μέθοδος αναλύεται με μεγάλη λεπτομέρεια σε διάφορες δημοσιεύσεις, όπως οι [46], [47]. Συνοπτικά συνίσταται στα παρακάτω βήματα.

Έστω ένα σύνολο δεδομένων που απεικονίζεται στον πίνακα X, όπου κάθε γραμμή αντιστοιχεί σε μία παρατήρηση και κάθε στήλη σε ένα χαρακτηριστικό. Το πρώτο βήμα στην ανάλυση κυρίων συνιστωσών περιλαμβάνει το κεντράρισμα των δεδομένων αφαιρώντας το μέσο όρο κάθε μεταβλητής. Ο πίνακας συνδιακύμανσης C υπολογίζεται στη συνέχεια μέσω του τύπου:

$$C = \frac{1}{n} X^T X$$

όπου n το σύνολο των παρατηρήσεων. Τα ιδιοδιανύσματα v και οι ιδιοτιμές λ του C υπολογίζονται, και τα ιδιοδιανύσματα ταξινομούνται κατά φθίνουσα σειρά, ανάλογα με τις ιδιοτιμές τους.

Οι κύριες συνιστώσες κατασκευάζονται ως γραμμικοί συνδυασμοί των αρχικών μεταβλητών, με κάθε μία να συνεισφέρει κατά ένα ποσοστό στη συνολική διακύμανση των δεδομένων. Οι συνιστώσες δίνονται από το $PC_i = Xvi$, όπου v_i το ιοστό ιδιοδιάνυσμα. Το αθροιστικό ποσοστό διακύμανσης που εξηγείται από τις πρώτες k κύριες συνιστώσες χρησιμοποιείται συχνά για την αξιολόγηση της μείωσης διαστάσεων που επιτυγχάνεται.



Εικόνα 5-Παράδειγμα Διαγράμματος Αθροιστικής Μεταβλητότητας [48]

Μαθηματικά, ο μετασχηματισμός στο νέο σύστημα συντεταγμένων εκφράζεται ως Y=XV, όπου Y ο πίνακας των κύριων συνιστωσών, V ο πίνακας των ιδιοδιανυσμάτων, και X ο αρχικός πίνακας δεδομένων. Η μείωση των διαστάσεων επιτυγχάνεται με την επιλογή των πρώτων k στηλών του Y, διατηρώντας τις πιο σημαντικές πληροφορίες στα δεδομένα.

Η χρησιμότητα της μεθόδου εκτείνεται πέρα από τη μείωση διαστάσεων. Χρησιμεύει επιπλέον ως εργαλείο για τη μείωση του θορύβου, την αναγνώριση προτύπων και την οπτικοποίηση δεδομένων. Παρά τη μαθηματική της κομψότητα, οι χρήστες πρέπει να είναι προσεκτικοί. Η PCA υποθέτει ότι οι κύριες συνιστώσες με τη μεγαλύτερη διακύμανση περιέχουν και τη μεγαλύτερη/χρησιμότερη ποσότητα πληροφορίας. Η παραδοχή αυτή μπορεί να μην ευθυγραμμίζεται πάντα με τις ανάγκες συγκεκριμένων προβλημάτων.

Επιλογή Χαρακτηριστικών

Η επιλογή χαρακτηριστικών επιτυγχάνεται με διάφορες τεχνικές και αποσκοπεί στην επιλογή ενός υποσυνόλου των χαρακτηριστικών του συνόλου δεδομένων. Οι τεχνικές αυτές διαχωρίζονται σε μεθόδους τύπου φιλτραρίσματος (filter) και σε μεθόδους τύπου wrapper. Οι πρώτες βασίζονται στα ίδια τα χαρακτηριστικά των δεδομένων και εφαρμόζονται ανεξάρτητα από τα μοντέλα εκπαίδευσης που θα εφαρμοστούν στη συνέχεια. Οι δεύτερες, από την άλλη, χρησιμοποιούν το ίδιο το μοντέλο εκπαίδευσης για να αξιολογήσουν τα υποψήφια υποσύνολα χαρακτηριστικών. Στα πλαίσια της παρούσας εργασίας έχουν χρησιμοποιηθεί 4 διαφορετικές τεχνικές επιλογής χαρακτηριστικών πριν από την εκπαίδευση των διαφόρων μοντέλων ταξινόμησης.

Η επιλογή χαρακτηριστικών βάσει συσχέτισης (Correlation-Based Feature Selection-CFS) στοχεύει στην επιλογή υποσυνόλων χαρακτηριστικών που συσχετίζονται σε μεγάλο βαθμό με τη μεταβλητή-στόχο, ενώ έχουν χαμηλή αλληλοσυσχέτιση μεταξύ τους. Επιδιώκει να εντοπίσει χαρακτηριστικά που παρέχουν σημαντική πληροφορία για εργασίες πρόβλεψης/ταξινόμησης στοχεύοντας παράλληλα στον περιορισμό, όσο γίνεται, του αριθμού τους. Η μέθοδος αναλύεται σε μεγάλο βαθμό στην πηγή [49].

Η σημαντικότητα χαρακτηριστικών βάσει δέντρου (Tree-Based Feature Importance) αναφέρεται στη μεθοδολογία που χρησιμοποιείται σε αλγορίθμους βασισμένους σε δέντρα απόφασης για την αξιολόγηση της σημασίας ή της συμβολής διαφορετικών χαρακτηριστικών στην πραγματοποίηση προβλέψεων ή ταξινομήσεων. Οι αλγόριθμοι που βασίζονται σε δέντρα αποφάσεων, όπως τα τυχαία δάση, ταξινομούν τα χαρακτηριστικά με βάση τη σημασία τους για τον διαχωρισμό των κόμβων και τη λήψη αποφάσεων μέσα στα δέντρα. Χαρακτηριστικά που εμφανίζονται συχνά στην κορυφή των δέντρων ή συμβάλλουν περισσότερο στη μείωση της ακαθαρσίας θεωρούνται πιο σημαντικά.

Ο σταδιακός αποκλεισμός χαρακτηριστικών (Recursive Feature Elimination-RFE) είναι τεχνική επιλογής χαρακτηριστικών που αφαιρεί επαναληπτικά τα λιγότερο σημαντικά χαρακτηριστικά από το σύνολο δεδομένων μέχρι να ικανοποιηθεί ο καθορισμένος αριθμός χαρακτηριστικών ή ένα προκαθορισμένο κριτήριο. Λειτουργεί εκπαιδεύοντας ένα μοντέλο για το πλήρες σύνολο των χαρακτηριστικών και στη συνέχεια ταξινομεί τα χαρακτηριστικά με βάση τη σημασία τους. Στη συνέχεια, αφαιρεί τα λιγότερο σημαντικά από αυτά και επαναλαμβάνει τη διαδικασία μέχρι να επιτευχθεί ο επιθυμητός αριθμός χαρακτηριστικών.

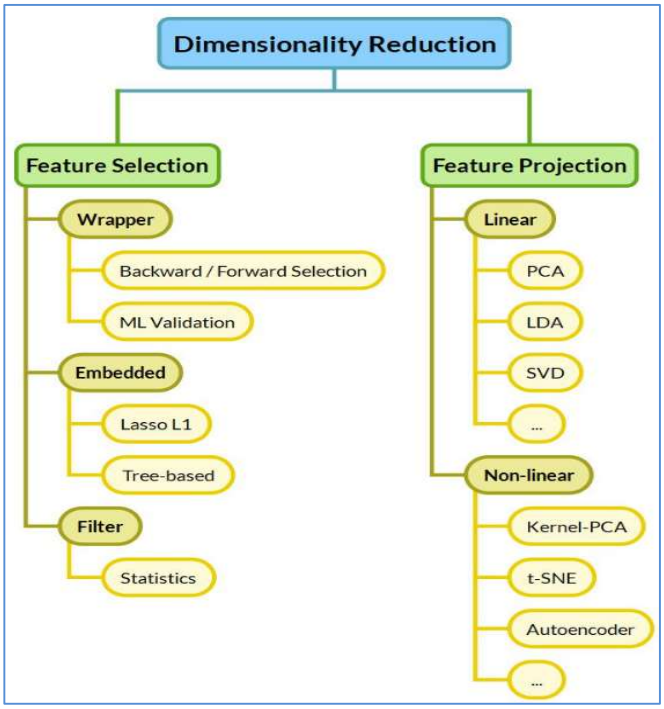
Τέλος, η κανονικοποίηση LASSO (Least Absolute Shrinkage and Selection Operator) χρησιμοποιείται στην ανάλυση παλινδρόμησης και στη μηχανική μάθηση για να επιβάλει ποινή στο απόλυτο μέγεθος των συντελεστών ενός μοντέλου, στοχεύοντας σε απλούστερα και πιο ερμηνεύσιμα μοντέλα. Η μέθοδος εισήχθη για πρώτη φορά στο [50]. Μέσω αυτής οι συντελεστές των λιγότερο σημαντικών χαρακτηριστικών συρρικνώνονται μέχρι μηδενισμού, με αποτέλεσμα τα χαρακτηριστικά αυτά ουσιαστικά να αφαιρούνται εντελώς από το μοντέλο.

Είναι σημαντικό να μη συγχέουμε τις τεχνικές επιλογής χαρακτηριστικών με τις τεχνικές μείωσης διαστάσεων, όπως η PCA που αναλύεται στην προηγούμενη υποενότητα. Και οι 2 μέθοδοι χρησιμοποιούνται στη μηχανική εκμάθηση για τη βελτίωση της απόδοσης και της αποδοτικότητας του μοντέλου, αλλά διαφέρουν ως προς τις προσεγγίσεις και τους στόχους τους.

Η επιλογή χαρακτηριστικών περιλαμβάνει την επιλογή ενός υποσυνόλου των αρχικών χαρακτηριστικών από το σύνολο δεδομένων, διατηρώντας παράλληλα τον αρχικό χώρο χαρακτηριστικών. Βοηθά στην απλοποίηση των μοντέλων, στη μείωση της

υπερπροσαρμογής στα δεδομένα εκπαίδευσης και στη βελτίωση της ερμηνευσιμότητας, δουλεύοντας με ένα υποσύνολο των πιο σχετικών χαρακτηριστικών.

Η μείωση διαστάσεων, από την άλλη πλευρά, περιλαμβάνει τη μετατροπή του αρχικού χώρου χαρακτηριστικών σε χώρο χαμηλότερης διάστασης. Στοχεύει ουσιαστικά στο μετασχηματισμό ή τη συμπίεση των αρχικών χαρακτηριστικών σε μια αναπαράσταση χαμηλότερης διάστασης, συχνά δημιουργώντας νέα σύνθετα χαρακτηριστικά που διατηρούν βασικές πληροφορίες από το αρχικό σύνολο δεδομένων.



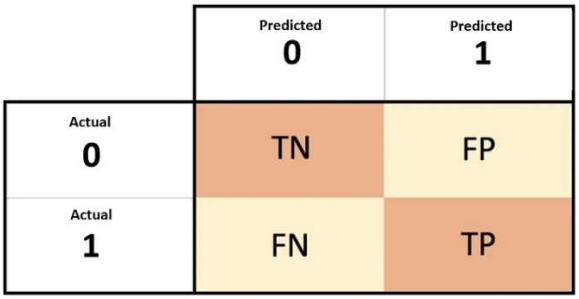
Εικόνα 6-Ταξινόμηση Τεχνικών Μείωσης Διαστάσεων [51]

2.2.2 Μέθοδοι Αξιολόγησης Απόδοσης

Στη βιβλιογραφία υπάρχει πληθώρα μετρικών αξιολόγησης για τα μοντέλα ταξινόμησης ανάλογα με τις ανάγκες που προκύπτουν από το εκάστοτε πρόβλημα. Οι ορισμοί που ακολουθούν για τις εδώ χρησιμοποιούμενες μετρικές έχουν ως πρότυπο κυρίως τα [11][17] αλλά και συνδυαστικά όλες τις πηγές της παρούσας εργασίας.

Πίνακας Σύγχυσης (Confusion Matrix)

Ο πίνακας σύγχυσης είναι ένας εύκολος τρόπος για την αναπαράσταση της επίδοσης ενός μοντέλου ταξινόμησης. Πρόκειται για ένα πίνακα n x n διαστάσεων, όπου n ο αριθμός των διαφορετικών κλάσεων στις οποίες μπορεί να ανήκουν τα δεδομένα του προβλήματος. Κάθε στοιχείο ai,j του πίνακα αντιπροσωπεύει το πλήθος των περιπτώσεων όπου μία παρατήρηση ανήκει στην κλάση i και ταξινομήθηκε από το μοντέλο στην κλάση j. Η απλούστερη μορφή του πίνακα συναντάται σε προβλήματα 2 κλάσεων, όπου και είναι 2x2 διαστάσεων, με γραμμές τις πραγματικές τιμές των παρατηρήσεων και στήλες τις προβλεπόμενες από το μοντέλο τιμές. Η μορφή του πίνακα είναι συνήθως η παρακάτω:



Εικόνα 7-Παράδειγμα Πίνακα Σύγχυσης σε Πρόβλημα Δυαδικής Ταξινόμησης [52]

όπου:

- TP (αληθώς θετικά - true positive): τα παραδείγματα που ανήκουν στην κλάση 1 και ταξινομήθηκαν στην 1
- FN (ψευδώς αρνητικά - false negative): τα παραδείγματα που ανήκουν στην κλάση 1, αλλά ταξινομήθηκαν στην κλάση 2
- FP (ψευδώς θετικά - false positive): τα παραδείγματα που ανήκουν στην κλάση 2, αλλά ταξινομήθηκαν στην κλάση 1
- TN (αληθώς αρνητικά - true negative): τα παραδείγματα που ανήκουν στην κλάση 2 και ταξινομήθηκαν στην 2

Παρατηρούμε επιπλέον ότι:

TP+TN= το σύνολο των σωστά ταξινομημένων παρατηρήσεων

FP+FN= το σύνολο των λανθασμένα ταξινομημένων παρατηρήσεων

TP+TN+ FP+FN = το σύνολο όλων των παρατηρήσεων που πέρασαν από το μοντέλο

Μετρικές Αξιολόγησης

1) Ορθότητα (Accuracy): είναι από τα συχνότερα μέτρα αξιολόγησης ενός ταξινομητή, που εκφράζει το συνολικό ποσοστό των ορθών ταξινομήσεων βάσει του τύπου:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Πρόκειται για σχετικά απλοϊκή μετρική καθώς δίνει μια καλή εικόνα για το μοντέλο μόνο σε περιπτώσεις που τα δεδομένα είναι ισορροπημένα μοιρασμένα στις διαφορετικές κλάσεις. Σε διαφορετική περίπτωση μπορεί να οδηγήσει σε εσφαλμένα συμπεράσματα και το μοντέλο, αν και με υψηλή ακρίβεια να μην ταξινομεί ποτέ σωστά τα δείγματα που ανήκουν στην μειοψηφική κλάση. Αυτό είναι σημαντικό μειονέκτημα σε προβλήματα όπου η ανίχνευση των δειγμάτων της μειοψηφικής κλάσης είναι υψίστης σημασίας, όπως πχ στην έγκαιρη ανίχνευση παθήσεων ή στον εντοπισμό απάτης. Σε τέτοια προβλήματα προτιμώνται άλλες μετρικές για την αξιολόγηση ενός μοντέλου ταξινόμησης.

2) Ακρίβεια (Precision): είναι το ποσοστό των θετικών προβλέψεων που είναι ορθές και υπολογίζεται βάσει της σχέσης:

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

3) Ανάκληση/ Ευαισθησία (Recall/Sensitivity/True Positive Rate): είναι η πιθανότητα του μοντέλου να προβλέψει ότι μια παρατήρηση ανήκει στη θετική κλάση όταν η παρατήρηση αυτή πράγματι ανήκει στη θετική κλάση.

$$Recall/Sensitivity = \frac{TP}{(TP + FN)}$$

4) Ειδικότητα (Specificity/True Negative Rate): είναι η πιθανότητα του μοντέλου να προβλέψει ότι μια παρατήρηση ανήκει στην αρνητική κλάση όταν η παρατήρηση αυτή πράγματι ανήκει στην αρνητική κλάση.

$$Specificity = \frac{TN}{(TN + FP)}$$

5) F-Measure (F1-Score): είναι ο αρμονικός μέσος των precision και recall, ο οποίος δίνεται από τη σχέση:

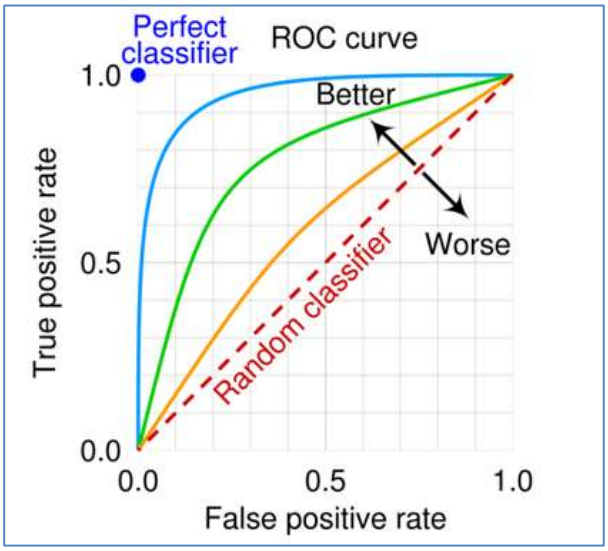
$$F1score = 2x\ \frac{(Precision\ x\ Recall)}{(Precision +\ Recall)}$$

6) Συντελεστής Συσχέτισης Matthews (Matthews Correlation Coefficient - MCC): λαμβάνει υπόψιν εξίσου τις μετρικές sensitivity και specificity, στοχεύοντας στο να παρέχει μια πιο ισορροπημένη εκτίμηση για την απόδοση του μοντέλου. Υψηλότερες τιμές της μετρικής MCC δείχνουν συνολικά καλύτερη προβλεπτική ικανότητα. Η μετρική υπολογίζεται μέσω του παρακάτω τύπου:

$$MCC = \frac{(TP\ x\ TN) - (FP\ x\ FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Καμπύλη Λειτουργικού Χαρακτηριστικού Δέκτη (ROC Curve)

Η καμπύλη λειτουργικού χαρακτηριστικού δέκτη (Receiver Operating Characteristic Curve – ROC) αποτελεί τη γραφική αναπαράσταση της σχέσης του ποσοστού των αληθώς θετικών (TP) και των ψευδώς θετικών προβλέψεων του μοντέλου για όλες τις πιθανές οριακές τιμές διαχωρισμού.



Εικόνα 8- Παράδειγμα Καμπύλης Λειτουργικού Χαρακτηριστικού Δέκτη [53]

Η καμπύλη βρίσκεται εντός ενός τετραγώνου με πλευρά μήκους 1. Η διαγώνιος μεταξύ των σημείων (0,0) και (1,1) αναπαριστά έναν ταξινομητή που προβλέπει τυχαία την κλάση. Οι

ταξινομητές που βρίσκονται πάνω από την διαγώνιο είναι καλύτεροι σε σχέση με την τυχαία πρόβλεψη. Όσο πιο μετατοπισμένο βρίσκεται κάποιο σημείο της καμπύλης προς την πάνω αριστερή γωνία του τετραγώνου, δηλαδή στο σημείο (0,1) , τόσο πιο καλός είναι ο ταξινομητής.

Το εμβαδόν κάτω από την καμπύλη (Area Under ROC Curve – AUC) χρησιμοποιείται για τη σύγκριση δύο ή περισσοτέρων ταξινομητών. Είναι μια μέτρηση που αντιπροσωπεύει τη συνολική απόδοση του μοντέλου. Κυμαίνεται από 0 έως 1, όπου υψηλότερη τιμή υποδηλώνει υψηλότερη επίδοση ταξινόμησης. Ένα AUC 0.5 υποδηλώνει τυχαία απόδοση, ενώ ένα AUC 1.0 υποδηλώνει τέλεια επίδοση.

Κεφάλαιο 3 - Μεθοδολογία Έρευνας

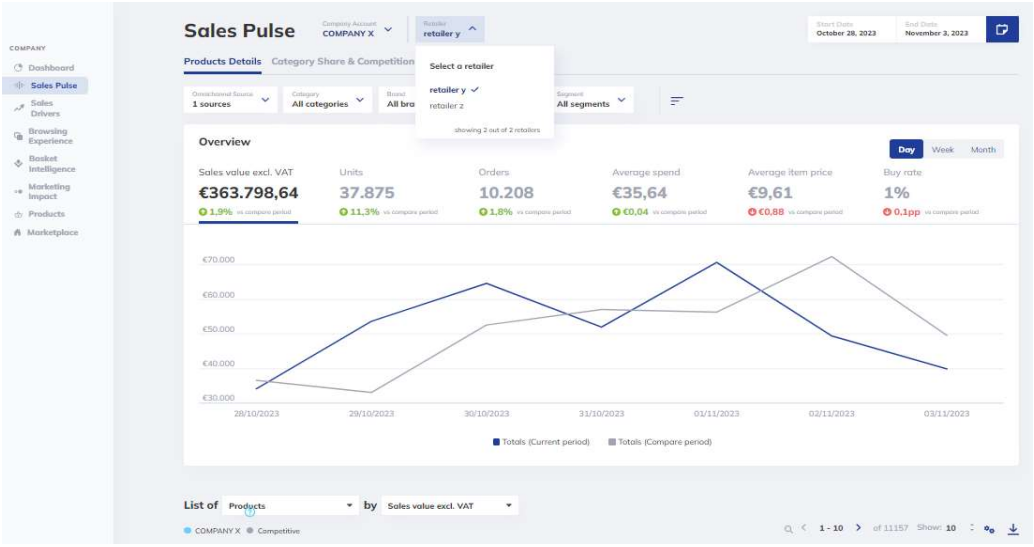
3.1 Η Convert Group

Τα δεδομένα που χρησιμοποιούνται στην παρούσα έρευνα είναι ευγενική παραχώρηση της Convert Group. Η Convert Group είναι εταιρεία ανάλυσης δεδομένων που εδρεύει στο κέντρο της Αθήνας και έχει αυτή τη στιγμή πελάτες σε 21 χώρες. Όπως αναφέρεται στην επίσημη ιστοσελίδα της [54], αποστολή της είναι να φέρει μία «επανάσταση» στον τρόπο συνεργασίας των καταστημάτων λιανικής ηλεκτρονικού, και μη, εμπορίου, με τους προμηθευτές τους, διευκολύνοντας τη λήψη αποφάσεων που βασίζονται σε δεδομένα.

Αυτή τη στιγμή παρέχει 3 λύσεις υπηρεσιών λογισμικού (Software as a Service - SaaS) σε εμπορικές επωνυμίες (brands) και καταστήματα λιανικής: το eRetail Audit Panel (παρακολούθηση ολόκληρου κλάδου ηλεκτρονικού εμπορίου για μέτρηση μεριδίων αγοράς), το eRetail Content (διαχείριση ηλεκτρονικού ραφιού) και το eRetail Audit Marketplace (δημιουργία εσόδων για καταστήματα λιανικής μέσω της πώλησης των δεδομένων τους σε προμηθευτές). Στην πλατφόρμα eRetail Audit Marketplace γίνεται μεγαλύτερη αναφορά καθώς πρόκειται για το πεδίο έρευνας της παρούσας εργασίας.

3.2 Το eRetail Audit Marketplace

Η πλατφόρμα eRetail Audit Marketplace επιτρέπει σε πωλητές λιανικής να μοιράζονται τα δεδομένα τους με τους προμηθευτές τους με ένα δομημένο και κατανοητό τρόπο. Ανάλογα με το πλάνο, παρέχει χρήσιμες πληροφορίες (insights) για τον προμηθευτή, αλλά και τον ανταγωνισμό του, όπως τζίρο, τεμάχια, μέση τιμή προϊόντων, διείσδυση καλαθιού, διείσδυση πελατών, απόδοση κατηγορίας, προβολές σελίδας, ρυθμός αγοράς, κατάσταση ευκαιρίας, συνάφεια προϊόντων, συνάφεια κατηγοριών, συνάφεια εμπορικών επωνυμιών, ευκαιρίες διασταυρούμενης πώλησης, ανάλυση ημερομηνίας-ώρας, πληροφορίες καναλιών μάρκετινγκ και συσκευών, όγκος αναζητήσεων, μερίδιο πλοήγησης στην ιστοσελίδα, απόδοση προωθητικών ενεργειών, έκθεση σε μέσα ενημέρωσης και πολλά άλλα. Η λίστα με τις παρεχόμενες πληροφορίες και δυνατότητες συνεχώς εμπλουτίζεται.



Εικόνα 9-Οθόνη Sales Pulse, eRAM

Αυτή τη στιγμή ωστόσο η πλατφόρμα δεν παρέχει κάποιου είδους πρόβλεψη για μελλοντικές χρονικές περιόδους. Η ενδεχόμενη δυνατότητα παροχής τέτοιου είδους πληροφορίας σε προμηθευτές είναι το αντικείμενο της παρούσας εργασίας, και συγκεκριμένα η ενδεχόμενη δυνατότητα πρόβλεψης της μελλοντικής απώλειας πελατών (churn prediction).

3.3 Περιγραφή Διαθέσιμων Δεδομένων

3.3.1 Πηγές Δεδομένων

Η παροχή των δεδομένων στους πελάτες της πλατφόρμας επιτυγχάνεται με τη λήψη και επεξεργασία των διαθέσιμων δεδομένων του εκάστοτε ηλεκτρονικού (και όχι μόνο) καταστήματος (retailer). Η σύνδεση με τα δεδομένα γίνεται με διάφορους τρόπους, όπως: σύνδεση στα Google Analytics (GA4) ή το Google Big Query του retailer, σύνδεση μέσω Application Programing Interface - API που παρέχει ο retailer, ημερήσια αρχεία δεδομένων μέσω πρωτοκόλλου ασφαλούς μεταφοράς αρχείων (Secure File Transfer Protocol - SFTP). Κάθε retailer έχει ένα δικό του συνδυασμό των παραπάνω μεθόδων για την ενσωμάτωση των δεδομένων του στην πλατφόρμα.

Τα δεδομένα όλων των retailer αποθηκεύονται σε μία PostgreSQL βάση δεδομένων. Η πρόσβαση σε αυτά μπορεί να γίνει εύκολα μέσω SQL ερωτημάτων. Για την ευκολότερη διαχείριση ερωτημάτων προς τη βάση χρησιμοποιείται το Redash, μια πλατφόρμα επιχειρηματικής ευφυΐας (BI) και οπτικοποίησης δεδομένων ανοιχτού κώδικα που επιτρέπει στους χρήστες να συνδέονται με διάφορες πηγές δεδομένων, να δημιουργούν διαγράμματα και να πραγματοποιούν αναλύσεις. Εκεί εκτελούνται και τα ερωτήματα που έχουν δημιουργηθεί για την εύρεση και εξαγωγή των δεδομένων της εργασίας.

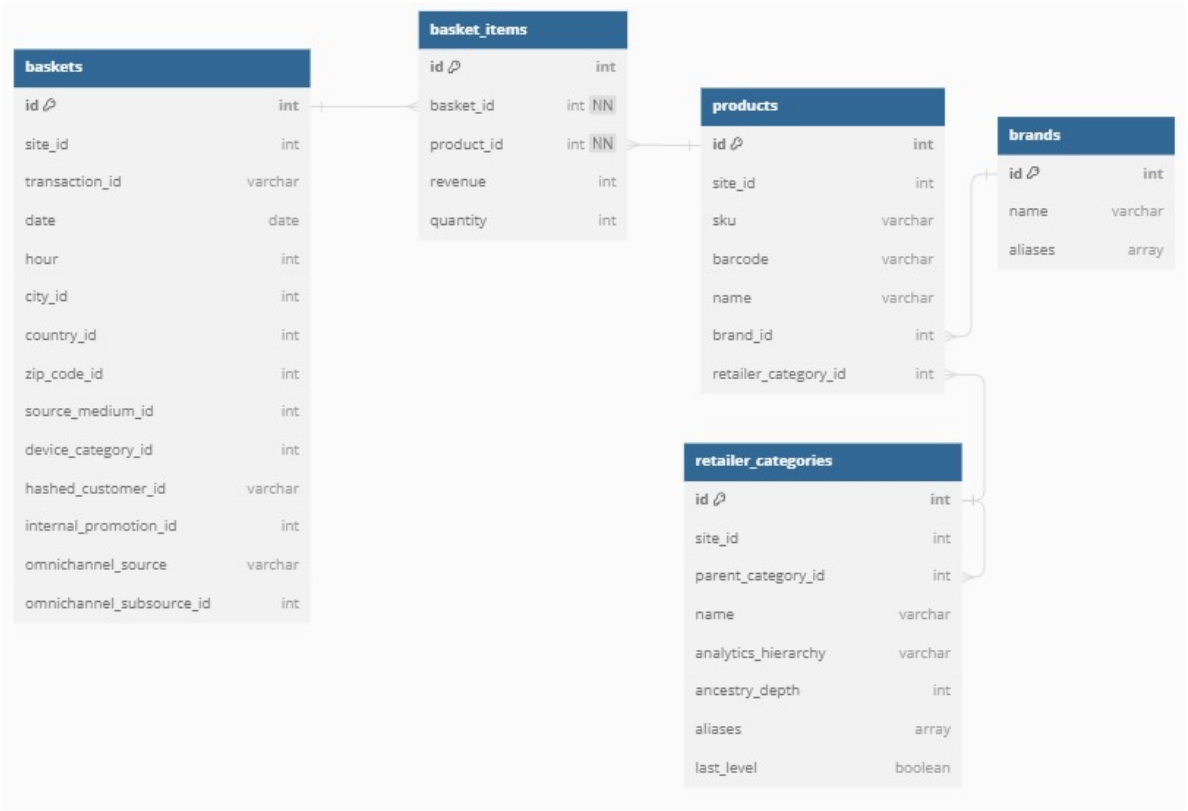
3.3.2 Δομή Βάσης Δεδομένων

Στο κεφάλαιο αυτό εστιάζουμε στο υποσύνολο της βάσης δεδομένων που αναφέρεται στα δεδομένα που λαμβάνουμε από τα ηλεκτρονικά καταστήματα. Αυτά μπορούν να χωριστούν σε 3 μεγάλους πυλώνες:

- **Δεδομένα Συναλλαγών**: αναφέρονται στις πωλήσεις του retailer. Για κάθε καλάθι που αγοράστηκε λαμβάνουμε ένα σύνολο από χαρακτηριστικά όπως: αναγνωριστικό καταστήματος (site_id), μοναδικός κωδικός συναλλαγής (transaction_id), ημερομηνία συναλλαγής, ώρα συναλλαγής, πόλη, χώρα, ταχυδρομικός κώδικας, κατηγορία συσκευής, κατακερματισμένο αναγνωριστικό πελάτη (hashed_customer_id), αναγνωριστικό προωθητικής ενέργειας, κανάλι αγοράς, κανάλι μάρκετινγκ (source_medium). Να σημειωθεί ότι δεν στέλνουν όλοι οι retailers όλες τις παραπάνω πληροφορίες. Ένα καλάθι μπορεί να αποθηκευτεί στη βάση, αρκεί να διαθέτει τις πληροφορίες site_id, transaction_id, date και hour.
- **Δεδομένα Σελίδων:** αναφέρονται στα διάφορα μονοπάτια (path) της ιστοσελίδας του retailer και στις προβολές (pageviews) που αυτά σημείωσαν ανά ημέρα. Κάθε ημέρα λαμβάνονται ουσιαστικά όλα τα σύνολα date-path-pageviews για κάθε retailer. Τα pageviews μιας σελίδας στα πλαίσια της ίδιας συνεδρίας του χρήστη δεν διπλομετρούνται.
- **Δεδομένα Όρων Αναζήτησης:** αναφέρονται στις λέξεις κλειδιά (search terms) που αναζητούν οι χρήστες στην ιστοσελίδα του retailer και στις προβολές (pageviews) που

αυτά σημείωσαν ανά ημέρα. Κάθε ημέρα λαμβάνονται ουσιαστικά όλα τα σύνολα date-search term-pageviews για κάθε retailer.

Τα δεδομένα που μας ενδιαφέρουν στην υπάρχουσα έρευνα είναι τα δεδομένα συναλλαγών. Για μεγαλύτερη κατανόηση παρέχονται παρακάτω σε διαγραμματική μορφή- χωρίς επέκταση σε πίνακες που δεν περιέχουν δεδομένα που έχουν χρησιμοποιηθεί.



Εικόνα 10-Δομή Χρησιμοποιούμενων Δεδομένων (χρήση https://dbdiagram.io)

3.4 Συνοπτική Περιγραφή Πειραματικής Μελέτης

Στην ενότητα 4 που ακολουθεί περιγράφεται λεπτομερώς όλη η πορεία που ακολουθείται προκειμένου να καταλήξουμε σε ένα – πρωταρχικό – μοντέλο πρόβλεψης απώλειας πελατών για μία εμπορική επωνυμία στα πλαίσια ενός ηλεκτρονικού καταστήματος.

Αρχικά περιγράφεται η πορεία που ακολουθείται για την επιλογή ενός συνόλου δεδομένων ανάμεσα σε όλα τα δεδομένα της πλατφόρμας. Αυτό που ιδανικά ψάχνουμε είναι μία εμπορική επωνυμία που εντός ενός καταστήματος έχει αγοραστεί πολλές φορές από πολλούς διαφορετικούς πελάτες. Όταν βρεθεί το ζευγάρι κατάστημα – εμπορική επωνυμία το οποίο καλύπτει με το βέλτιστο τρόπο τα παραπάνω κριτήρια εξάγεται από τη βάση δεδομένων του eRAM όλη η «χύμα» (raw) πληροφορία που το αφορά και είμαστε πλέον σε θέση να ξεκινήσουμε τη διαδικασία κατασκευής του τελικού συνόλου δεδομένων που θα αξιοποιηθεί για την εκπαίδευση μοντέλων μηχανικής μάθησης. Μέχρι εδώ χρησιμοποιείται το Redash και η γλώσσα SQL που αναφέρθηκαν στο 3.3.1. Τα ίδια εργαλεία χρησιμοποιούνται και για την εύρεση του διαστήματος που ορίζουμε ότι πρέπει να μεσολαβήσει από την τελευταία αγορά της μάρκας από τον πελάτη προκειμένου να θεωρηθεί χαμένος.

Έχοντας τη χύμα πληροφορία χρησιμοποιείται στη συνέχεια η γλώσσα Python για την κατασκευή του τελικού συνόλου δεδομένων. Το τελικό αυτό σύνολο περιλαμβάνει 22

χαρακτηριστικά για 12.595 διαφορετικούς πελάτες ενός καταστήματος που έχουν αγοράσει συγκεκριμένη μάρκα προϊόντων τουλάχιστον 3 φορές μέσα σε ορισμένο χρονικό διάστημα. Περιλαμβάνει επιπλέον την κατηγορία κάθε πελάτη (churner/non churner). Περιγράφεται στη συνέχεια όλη η προεπεξεργασία που γίνεται στο σύνολο δεδομένων καθώς και η διαδικασία της εκπαίδευσης μοντέλων μηχανικής μάθησης που ακολουθεί.

Για την εκπαίδευση μοντέλων χρησιμοποιούνται 9 διαφορετικοί αλγόριθμοι και συγκεκριμένα: λογιστική παλινδρόμηση, δέντρο απόφασης, naïve bayes, τυχαία δάση, μηχανές διανυσμάτων υποστήριξης, XGBoost, LightGBM, AdaBoost και voting classifier (ο οποίος ουσιαστικά συνδυάζει όλους τους προηγούμενους). Επιπλέον γίνεται προσπάθεια για βελτιστοποίηση παραμέτρων, όπου δίνεται η δυνατότητα.

Κάθε ένας από τους 9 παραπάνω αλγορίθμους εκπαιδεύεται με 6 διαφορετικούς τρόπους, ανάλογα με τα δεδομένα εκπαίδευσης που παίρνουν ως είσοδο. Και στις 6 περιπτώσεις τα δεδομένα διαχωρίζονται σε δεδομένα εκπαίδευσης και δεδομένα δοκιμής, κατά ποσοστά 80% και 20% αντίστοιχα. Κάθε τρόπος από τους 6 χαρακτηρίζεται από ένα διαφορετικό τρόπο επιλογής χαρακτηριστικών, και έτσι έχουμε τους κάτωθι:

1. Χρήση όλων των χαρακτηριστικών
2. Εκπαίδευση ύστερα από μετασχηματισμό των αρχικών δεδομένων με ανάλυση κυρίων συνιστωσών (PCA)
3. Επιλογή χαρακτηριστικών βάσει συσχέτισης (correlation-based feature selection)
4. Επιλογή χαρακτηριστικών βάσει δέντρου (tree-based feature importance)
5. Επιλογή χαρακτηριστικών με σταδιακό αποκλεισμό (recursive feature elimination)
6. Επιλογή χαρακτηριστικών με κανονικοποίηση LASSO

Κεφάλαιο 4 – Πειραματική Μελέτη

4.1 Επιλογή Συνόλου Δεδομένων

Στόχος είναι να μπορέσουμε να προβλέψουμε, για μία συγκεκριμένη εμπορική επωνυμία (brand) σε ένα συγκεκριμένο ηλεκτρονικό κατάστημα (retailer), πόσοι από τους πελάτες που το αγόρασαν σε ορισμένη χρονική περίοδο υπάρχει κίνδυνος να πάψουν να το αγοράζουν. Η ιδιαιτερότητα είναι πως δεν έχουμε καθόλου στοιχεία για τους πελάτες του εκάστοτε καταστήματος, στα πλαίσια της διαφύλαξης προσωπικών δεδομένων. Τα χαρακτηριστικά για τη διαδικασία της μηχανικής μάθησης θα πρέπει να παραχθούν έμμεσα από τα δεδομένα συναλλαγών του κάθε πελάτη. Η προσπάθεια για την πρόβλεψη αυτή μπορεί φυσικά να γίνει μόνο σε retailers που μας παρέχουν την πληροφορία hashed_customer_id στα καλάθια αγορών, άρα retailers που μοιράζονται με τους προμηθευτές τους την αναφορά διείσδυσης πελατών (customer penetration report).

Ανάμεσα στους retailers που παρέχουν την παραπάνω πληροφορία επιχειρείται μέσω ενός ερωτήματος SQL (query1) να βρεθεί το brand εκείνο με τους περισσότερους πελάτες που το έχουν αγοράσει σε τουλάχιστον 3 συναλλαγές. Μέσω αυτού εντοπίζουμε πως στον retailer με id=101 υπάρχουν για το brand_id=158 14.077 hashed_customer_ids που το έχουν συμπεριλάβει σε τουλάχιστον 3 συναλλαγές τους στο κατάστημα. Στη μελέτη αυτή θα χρησιμοποιηθεί επομένως το ζευγάρι αυτό ώστε να έχουμε επαρκή αριθμό από παρατηρήσεις.

Πρέπει επιπλέον να οριστεί το διάστημα που πρέπει να μεσολαβήσει από την τελευταία αγορά του brand από ένα πελάτη ώστε αυτός να θεωρηθεί χαμένος. Για να οριστεί το διάστημα αυτό ακολουθείται η εξής διαδικασία: με ένα νέο SQL ερώτημα (query2) υπολογίζεται, για τους πελάτες που έχουν αγοράσει το brand_id=158 τουλάχιστον 2 φορές, το μέσο διάστημα σε ημέρες μεταξύ 2 συναλλαγών τους που περιλαμβάνουν το brand. Το αποτέλεσμα είναι 120 ημέρες ≈ 4 μήνες. Πρόκειται για πολύ λογικό διάστημα, δεδομένου πως πρόκειται για brand με καταναλωτικά προϊόντα στο χώρο του καλλυντικού με μέσο διάστημα χρήσης που προσεγγίζει τους 4-6 μήνες. Η παραπάνω διαδικασία για τον ορισμό χρονικού διαστήματος συναντάται και στη βιβλιογραφία, όπως για παράδειγμα στο [6].

Με βάση τα παραπάνω το σύνολο δεδομένων που θα δημιουργηθεί θα έχει χαρακτηριστικά που θα παραχθούν από τα δεδομένα συναλλαγών των πελατών που αγόρασαν το brand_id=158 στον retailer με id=101 τουλάχιστον 3 φορές κατά το διάστημα 01-01-2022 έως 30-06-2023 – άρα από παρελθοντικά δεδομένα συναλλαγών 1μιση χρόνου. Οι ετικέτες για κάθε πελάτη (churned/ not churned) θα προκύψουν από το αν ο πελάτης αγόρασε το brand στους επόμενους 4 μήνες – άρα στο διάστημα 01-07-2023 έως 31-10-2023.

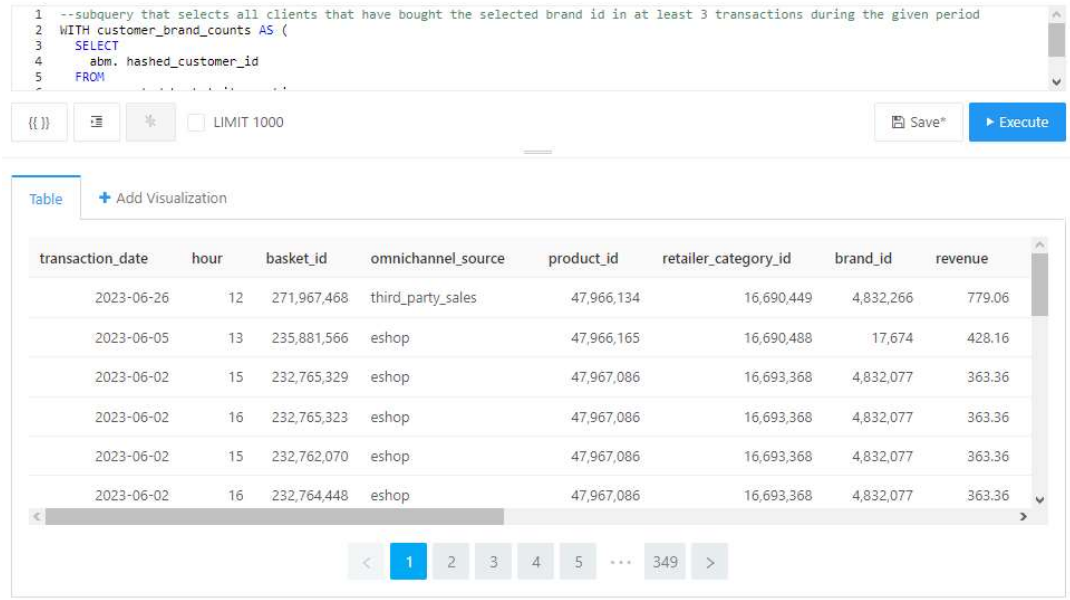
Εδώ πρέπει να σημειωθούν 2 περιορισμοί

1. Δε μπορούμε να υπολογίσουμε τις συναλλαγές ενός πελάτη μεταξύ διαφορετικών ηλεκτρονικών καταστημάτων στην πλατφόρμα, εφόσον κάθε ένα από αυτά στέλνει τα δικά του μοναδικά hashed_customer_ids. Αποχώρηση από 1 brand σε συγκεκριμένο retailer δε σημαίνει απαραίτητα ότι ο πελάτης δεν ξαναγοράζει ποτέ το brand αλλά ενδέχεται να το αγοράσει από διαφορετικό κατάστημα. Η έρευνα επικεντρώνεται επομένως αποκλειστικά στην αποχώρηση εντός συγκεκριμένου retailer.

2. Ενδέχεται να περιέχονται και συναλλαγές χονδρικής ανάμεσα στα διαθέσιμα δεδομένα, τις οποίες όμως δε μπορούμε να ταυτοποιήσουμε με ασφάλεια.

4.2 Κατασκευή Συνόλου Δεδομένων

Η κατασκευή του συνόλου δεδομένων ξεκινά με τη δημιουργία και εκτέλεση στο Redash ενός νέου ερωτήματος SQL (query3) με το οποίο ανακτώνται όλα τα δεδομένα συναλλαγών, για το διάστημα 01-01-2022 έως 30-06-2023, όλων των πελατών που στο ίδιο διάστημα έχουν συμπεριλάβει το brand 158 σε τουλάχιστον 3 συναλλαγές τους. Ουσιαστικά το σύνολο δεδομένων που ανακτούμε από το ερώτημα αυτό είναι ένα csv αρχείο στο οποίο κάθε γραμμή αντιπροσωπεύει ένα είδος καλαθιού (basket item) αγορασμένο από κάποιον από τους παραπάνω πελάτες. Κάθε basket item αντιπροσωπεύει ένα διαφορετικό κωδικό προϊόντος. Τα basket items που μοιράζονται το ίδιο basket id αγοράστηκαν κατά την ίδια συναλλαγή.



Εικόνα 11-Εκτέλεση SQL Query στο redash

Το csv αρχείο με τα "raw" δεδομένα που προκύπτει έχει τις ακόλουθες, για κάθε 1 από τα 632.415 basket items που περιέχει, πληροφορίες που παρουσιάζονται στον παρακάτω πίνακα:

a/a	Μεταβλητή	Τύπος	Μοναδικές Τιμές	Περιγραφή
1	hashed_customer_id	Κατηγορική	12.596	κατακερματισμένο αναγνωριστικό πελάτη
2	transaction_date	Κατηγορική	547	ημερομηνία συναλλαγής
3	hour	Συνεχής	24	ώρα συναλλαγής
4	basket_id	Κατηγορική	119.079	μοναδικός κωδικός καλαθιού
5	omnichannel_source	Κατηγορική	2	κανάλι αγοράς (τιμές eshop/third_party_sales)
6	product_id	Κατηγορική	36.037	κωδικός προϊόντος
7	retailer_category_id	Κατηγορική	726	κωδικός κατηγορίας προϊόντος
8	brand_id	Κατηγορική	1.468	κωδικός εμπορικής επωνυμίας

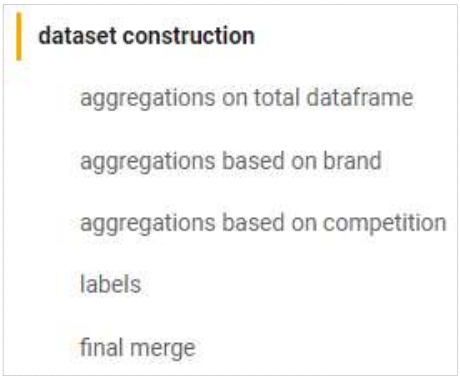
9	revenue	Συνεχής	10.422	τζίρος από την αγορά του προϊόντος
10	quantity	Συνεχής	104	αριθμός τεμαχίων του προϊόντος που αγοράστηκαν
11	source_medium	Κατηγορική	17	κανάλι μάρκετινγκ - περιέχει τιμές όπως Direct, SEO, SEA, mailing, flyerQRcode, search cpc, unknown, shopping cpc
12	device_category	Κατηγορική	2	κατηγορία συσκευής (desktop/mobile)
13	internal_promotion_id	Κατηγορική	1308	αναγνωριστικό προωθητικής ενέργειας - όταν υπάρχει τιμή σημαίνει πως στο καλάθι οδήγησε κάποια προωθητική ενέργεια του ηλεκτρονικού καταστήματος

Πίνακας 1-Λίστα Raw Δεδομένων

Όπως είναι φυσικό γραμμές με κοινό basket_id, που συνεπώς ανήκουν στην ίδια συναλλαγή, έχουν κοινές τιμές σε όλες τις στήλες εκτός των product_id, retailer_category_id, brand_id, revenue και quantity. Τα μοναδικά hashed_customer_id του αρχείου είναι 12.595, επομένως αυτό θα είναι τελικά και το μέγεθος του δείγματος.

Με ένα τελευταίο ερώτημα SQL (query4) εξάγουμε για τα 12.595 μοναδικά hashed_customer_id που περιέχονται στο raw_data.csv την τιμή TRUE ή FALSE, ανάλογα με το αν είναι churners ή όχι. Churner, όπως ορίστηκε στο προηγούμενο κεφάλαιο θεωρούμε τον πελάτη αν δεν έχει ξαναγοράσει έστω μία φορά κάποιο προϊόν του brand κατά το διάστημα 01-07-2023 έως 31-10-2023. Οι ετικέτες αυτές αποθηκεύονται στο αρχείο labels.csv.

Με την εξαγωγή του csv με τα raw δεδομένα μπορεί να ξεκινήσει η διαδικασία του μετασχηματισμού ώστε να παραχθεί το τελικό σύνολο δεδομένων που θα χρησιμοποιηθεί στη μηχανική μάθηση. Τα είδη των χαρακτηριστικών που κατασκευάζονται έχουν ως βάση την πηγή [6] που επίσης ασχολείται με το ηλεκτρονικό εμπόριο, αλλά σε επίπεδο ολόκληρου retailer. Η κατασκευή του τελικού συνόλου δεδομένων γίνεται με χρήση της γλώσσας python και της βιβλιοθήκης pandas. Το σύνολο του κώδικα python που εκτελείται στα πλαίσια αυτής της εργασίας έχει οργανωθεί σε ένα βιβλίο εργασίας του collab (προσβάσιμο μέσω του συνδέσμου: https://shorturl.at/tAJ48). Όλα τα αρχεία δεδομένων αποθηκεύονται στο Google Drive για ευκολότερη πρόσβαση από το βιβλίο εργασίας. Ο κώδικας για την κατασκευή του τελικού συνόλου δεδομένων βρίσκεται στην ενότητα “dataset construction”. Αρχικά φορτώνεται το raw_data.csv από το Google Drive σε ένα πλαίσιο δεδομένων (dataframe) της βιβλιοθήκης pandas.



Εικόνα 12-Στάδια Κατασκευής Dataset

Τα πρώτα χαρακτηριστικά του τελικού dataset παράγονται με συναθροίσεις ανά πελάτη επί του συνολικού dataframe. Τα χαρακτηριστικά αυτά ανά πελάτη (διάστημα 01-01-2022 έως 30-06-2023) είναι τα ακόλουθα:

a/a	Μεταβλητή	Τύπος	Περιγραφή
1	total_revenue	Συνεχής	συνολικός τζίρος στον retailer
2	total_units	Συνεχής	συνολικά τεμάχια στον retailer
3	total_transactions	Συνεχής	συνολικός αριθμός συναλλαγών
4	unique_categories	Συνεχής	συνολικός αριθμός μοναδικών κατηγοριών προϊόντων που έχει αγοράσει
5	unique_product_codes	Συνεχής	συνολικός αριθμός μοναδικών κωδικών προϊόντος που έχει αγοράσει
6	days_btw_1st-last_tid	Συνεχής	ημέρες μεταξύ της 1ης και της τελευταίας καταγεγραμμένης συναλλαγής στον retailer
7	mean_days_btw_tids	Συνεχής	μέσο διάστημα σε ημέρες μεταξύ διαδοχικών συναλλαγών στον retailer
8	top_source_medium	Κατηγορική	κανάλι μάρκετινγκ μέσω του οποίου έχει κάνει τις περισσότερες συναλλαγές
9	baskets_eshop	Συνεχής	αριθμός συναλλαγών που έγιναν μέσω του ηλεκτρονικού καταστήματος
10	baskets_third_party	Συνεχής	αριθμός συναλλαγών που έγιναν μέσω τρίτων παρόχων (third_party_sales)
11	baskets_desktop	Συνεχής	αριθμός συναλλαγών που έγιναν μέσω υπολογιστή
12	baskets_mobile	Συνεχής	αριθμός συναλλαγών που έγιναν μέσω κινητού τηλεφώνου

Πίνακας 2-Χαρακτηριστικά από Συναθροίσεις επί Ολόκληρου του Dataset

Στη συνέχεια γίνεται φιλτράρισμα του αρχικού dataframe ώστε να κρατήσουμε μόνο τα basket items που αφορούν προϊόντα με brand_id=158. Τα επιπλέον χαρακτηριστικά ανά πελάτη που παράγονται από αυτό το υποσύνολο του dataframe είναι:

a/a	Μεταβλητή	Τύπος	Περιγραφή
13	revenue_on_brand	Συνεχής	συνολικός τζίρος σε προϊόντα του brand
14	units_on_brand	Συνεχής	συνολικά τεμάχια προϊόντων του brand
15	transactions_incl_brand	Συνεχής	συνολικός αριθμός συναλλαγών που περιλαμβάνουν προϊόντα του brand
16	unique_categories_of_brand	Συνεχής	συνολικός αριθμός μοναδικών κατηγοριών προϊόντων του brand που έχει αγοράσει

17	unique_product_codes_of_brand	Συνεχής	συνολικός αριθμός μοναδικών κωδικών προϊόντος του brand που έχει αγοράσει
18	days_btw_1st-last_tid_on_brand	Συνεχής	ημέρες μεταξύ της 1ης και της τελευταίας καταγεγραμμένης συναλλαγής που περιείχε προϊόντα του brand
19	mean_days_btw_tids_on_brand	Συνεχής	μέσο διάστημα σε ημέρες μεταξύ διαδοχικών συναλλαγών που περιείχαν προϊόντα του brand
20	brand_tids_from_promotion	Συνεχής	συνολικός αριθμός συναλλαγών που περιλαμβάνουν προϊόντα του brand και προήλθαν από κάποια προωθητική ενέργεια (άρα έχουν promotion_id != null)

Πίνακας 3- Χαρακτηριστικά από Συναθροίσεις στα Basket Items που αφορούν το Brand Ενδιαφέροντος

Τέλος γίνεται ένα επιπλέον φιλτράρισμα του αρχικού dataframe ώστε να κρατήσουμε τα basket items με κατηγορία στην οποία έχει παρουσία το brand αλλά όπου brand_id != 158. Ουσιαστικά πρόκειται για τις αγορές ανταγωνιστικών προϊόντων. Από το νέο αυτό dataframe έχουμε 2 επιπλέον χαρακτηριστικά ανά πελάτη:

a/a	Μεταβλητή	Τύπος	Περιγραφή
21	revenue_on_competition	Συνεχής	συνολικός τζίρος σε ανταγωνιστικά brand
22	quantity_on_competition	Συνεχής	συνολικά τεμάχια προϊόντων ανταγωνιστικών brand

Πίνακας 4- Χαρακτηριστικά από Συναθροίσεις στα Basket Items που αφορούν Ανταγωνιστικά Προϊόντα

Στο τελικό στάδιο κατασκευής φορτώνεται το αρχείο με τις ετικέτες σε ένα ακόμα dataframe. Όλα τα dataframes που περιέχουν τα χαρακτηριστικά που αναφέρθηκαν παραπάνω καθώς και το dataframe με τις ετικέτες συγχωνεύονται σε ένα ενιαίο dataframe που πλέον περιλαμβάνει 1 γραμμή ανά πελάτη του brand και κάθε στήλη αντιπροσωπεύει ένα χαρακτηριστικό. Το dataframe αυτό εγγράφεται, τέλος, σε ένα αρχείο csv (data.csv) στο Google Drive και έτσι είναι απευθείας προσβάσιμο χωρίς να χρειάζεται να ξανατρέξει ο κώδικας της ενότητας “dataset construction”.

4.3 Επεξεργασία Συνόλου Δεδομένων

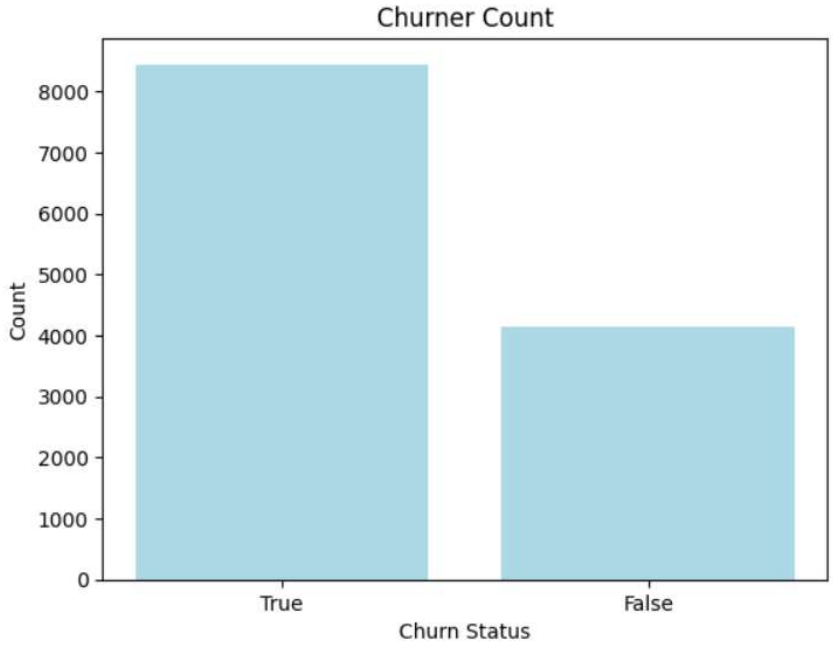
Πριν ξεκινήσει η κατασκευή μοντέλων μηχανικής μάθησης χρειάζεται μια στοιχειώδεις προεπεξεργασία του συνόλου δεδομένων. Αυτή γίνεται στην ενότητα “data pre-processing” του κώδικα στο collab. Οι αλλαγές που γίνονται είναι οι ακόλουθες:

1) διαχείριση ελλιπών τιμών: οι στήλες στις οποίες λείπουν τιμές είναι οι brand_tids_from_promotion (συνολικός αριθμός συναλλαγών που περιλαμβάνουν προϊόντα του brand και προήλθαν από κάποια προωθητική ενέργεια), revenue_on_competition (συνολικός τζίρος σε ανταγωνιστικά brand) και quantity_on_competition (συνολικά τεμάχια προϊόντων ανταγωνιστικών brand). Και

στις 3 περιπτώσεις η μη ύπαρξη κάποιας τιμής ουσιαστικά ισοδυναμεί με το να είχαμε την τιμή μηδέν. Συνεπώς και στις 3 στήλες οι null τιμές αντικαθίστανται με μηδενικά.

2) διαχείριση κατηγορικών μεταβλητών: στη στήλη top_source_medium (κανάλι μάρκετινγκ μέσω του οποίου ο πελάτης έχει κάνει τις περισσότερες συναλλαγές) έχουμε κατηγορικές τιμές (πχ Direct, Mailing). Οι τιμές αυτές θα μετατραπούν σε δυαδικές μέσω κωδικοποίησης one-hot, όπως αυτό αναλύθηκε στο Κεφάλαιο 2. Ουσιαστικά η στήλη top_source_medium καταργείται και τη θέση της παίρνουν 6 νέες στήλες, όσες οι μοναδικές τιμές της αρχικής στήλης.
3) μετατροπή των ετικετών σε δυαδικές τιμές (1/0 αντί TRUE/FALSE)
4) αφαίρεση της στήλης hashed_customer_id: ουσιαστικά δεν πρόκειται για χαρακτηριστικό αλλά για ένα είδος ταυτότητας κάθε γραμμής. Εφόσον η επεξεργασία λαμβάνει τέλος δεν είναι πλέον απαραίτητη.

Το τελικό σύνολο δεδομένων, ύστερα από όλα τα παραπάνω, έχει 28 στήλες – 27 χαρακτηριστικά και τη στήλη με την ετικέτα κάθε γραμμής. Οι παρατηρήσεις είναι σχετικά ομοιόμορφα κατανεμημένες ανάμεσα στις 2 κλάσεις με περίπου τα 2/3 (churn) στη μία και το 1/3 στην άλλη (no churn).



Εικόνα 13-Κατανομή Μεταβλητής Στόχου (is_churner)

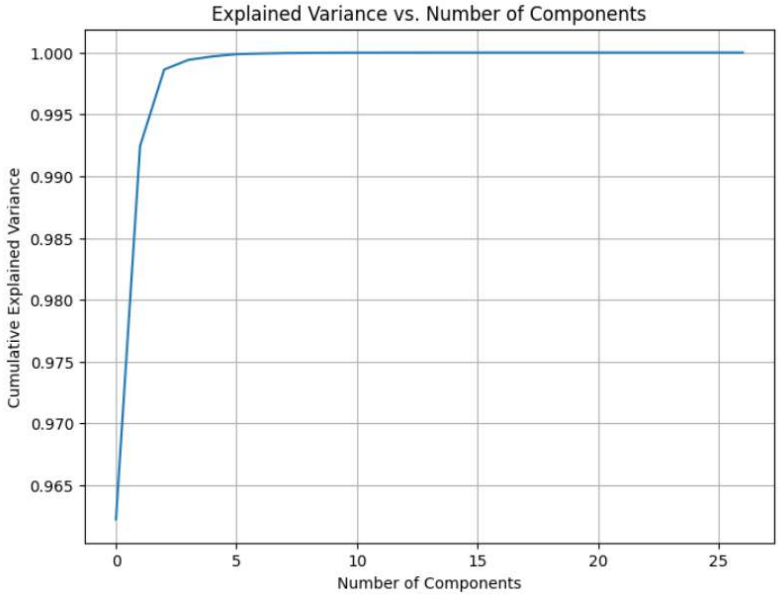
4.4 Χρήση Τεχνικών Μηχανικής Μάθησης για την Πρόβλεψη Απώλειας Πελατών

Για την προσπάθεια επίλυσης του προβλήματος χρησιμοποιούνται 9 διαφορετικά μοντέλα ταξινόμησης, τα οποία στη συνέχεια προσπαθούμε να αξιολογήσουμε. Για την κατασκευή των μοντέλων χρησιμοποιείται κυρίως η βιβλιοθήκη scikit-learn της python καθώς και οι βιβλιοθήκες xgboost & lightgbm (για την κατασκευή των αντίστοιχων boosting μοντέλων), matplotlib & seaborn (για οπτικοποίηση) και joblib (για αποθήκευση των μοντέλων για αργότερα ώστε να μην απαιτείται κάθε φορά η κατασκευή τους εκ νέου).

Οι 9 διαφορετικοί αλγόριθμοι που χρησιμοποιούνται για την εκπαίδευση μοντέλων είναι οι ακόλουθοι: λογιστική παλινδρόμηση, δέντρο απόφασης (με grid search για την επιλογή των παραμέτρων max_depth και min_sample_split), naïve bayes, τυχαία δάση (με grid search για

την επιλογή των παραμέτρων n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features, bootstrap), μηχανές διανυσμάτων υποστήριξης, xgboost (με grid search για την επιλογή των παραμέτρων n_estimators, max_depth, learning_rate), lightgbm (με grid search για την επιλογή των παραμέτρων n_estimators, num_leaves, learning_rate), adaboost (με grid search για την επιλογή των παραμέτρων n_estimators, base_estimator, learning_rate) & voting classifier (όπου χρησιμοποιούνται τα προηγούμενα 8 μοντέλα ταξινόμησης). Κάθε ένας από τους 9 παραπάνω αλγορίθμους εκπαιδεύεται με 6 διαφορετικούς τρόπους, ανάλογα με τα δεδομένα εκπαίδευσης που παίρνουν ως είσοδο. Και στις 6 περιπτώσεις τα δεδομένα διαχωρίζονται σε δεδομένα εκπαίδευσης «X_train» και δεδομένα δοκιμής «X_test», κατά ποσοστά 80% και 20% αντίστοιχα.

Στην πρώτη περίπτωση χρησιμοποιούνται ως είσοδος όλα τα δεδομένα εκπαίδευσης, χωρίς να αφαιρεθεί κανένα από τα 27 χαρακτηριστικά τους. Στο 2ο «γύρο» εκπαίδευσης των 9 διαφορετικών ταξινομητών επιχειρούμε το μετασχηματισμό των αρχικών δεδομένων με ανάλυση κυρίων συνιστωσών (PCA). Από το διάγραμμα αθροιστικής μεταβλητότητας που προκύπτει συμπεραίνεται ότι το 99% της μεταβλητότητας καλύπτεται μόλις από τις 2 πρώτες κύριες συνιστώσες. Επομένως παράγουμε 2 νέα σύνολα, τα «X_train_pca» και «X_test_pca», χρησιμοποιώντας PCA με 2 συνιστώσες. Τα νέα αυτά σύνολα θα χρησιμοποιηθούν για την εκπαίδευση των νέων μοντέλων.



Εικόνα 14-Διάγραμμα Αθροιστικής Μεταβλητότητας

Οι επόμενες 4 «απόπειρες» εκπαίδευσης των μοντέλων γίνονται μετά από εφαρμογή τεσσάρων διαφορετικών τεχνικών επιλογής χαρακτηριστικών στα δεδομένα εκπαίδευσης. Με τη χρήση correlation-based feature selection και ορίζοντας ως στόχο την επιλογή 10 από το σύνολο των 27 χαρακτηριστικών καταλήγουμε στα νέα «X_train_corr» και «X_test_corr» που περιλαμβάνουν τα χαρακτηριστικά: days_btw_1st-last_tid_on_brand, days_btw_1st-last_tid, unique_product_codes_of_brand, unique_categories_of_brand, unique_categories, transactions_incl_brand, mean_days_btw_tids, unique_product_codes, brand_tids_from_ promotion, baskets_third_party.

Η χρήση tree-based feature importance, πάλι με στόχο την επιλογή των 10 σημαντικότερων χαρακτηριστικών, καταλήγει στα «X_train_rf» και «X_test_rf» που περιλαμβάνουν τα χαρακτηριστικά: revenue_on_brand, days_btw_1st-last_tid_on_brand, days_btw_1st-last_tid, mean_days_btw_tids_on_brand, total_revenue, revenue_on_

competition, mean_days_btw_tids, total_units, unique_product_codes, quantity_on_competition.

Αντίστοιχα, με χρήση της τεχνικής recursive feature elimination καταλήγουμε στα εξής 10 σημαντικότερα χαρακτηριστικά για τα «X_train_rfe» και «X_test_rfe»: total_transactions, baskets_eshop, baskets_third_party, baskets_mobile, transactions_incl_brand, unique_categories_of_brand, sm_Mailing, sm_SEA, sm_SEO, sm_Unknown. Στη συγκεκριμένη τεχνική μπορούμε να ανακτήσουμε και το βαθμό σημαντικότητας των επιλεγμένων χαρακτηριστικών. Έτσι βλέπουμε πως τα total_transactions, unique_ categories_of_brand και transactions_incl_brand θεωρούνται τα 3 σημαντικότερα.

Η χρήση κανονικοποίησης LASSO οδηγεί, τέλος, στα «X_train_lasso» και «X_test_lasso» με χαρακτηριστικά τα total_revenue, total_units, days_btw_1st-last_tid, mean_days_btw_tids, baskets_third_party, revenue_on_brand, days_btw_1st-last_tid_on_brand, mean_days_btw _tids_on_brand, revenue_on_competition, quantity_on_competition. Παρατηρώντας τους συντελεστές που το μοντέλο αποδίδει στα χαρακτηριστικά μπορούμε να συμπεράνουμε ότι το total_units αναδεικνύεται ως το σημαντικότερο.

Παρατηρούμε ότι ανάλογα με την τεχνική επιλογής χαρακτηριστικών τα σύνολα δεδομένων εκπαίδευσης διαφέρουν πολύ ως προς τα χαρακτηριστικά που περιέχουν. Μετά την παραπάνω διαδικασία έχουμε 6 παραλλαγές του αρχικού dataset για εκπαίδευση των 9 διαφορετικών μοντέλων. Καταλήγουμε έτσι στην εκπαίδευση 9 X 6 = 54 διαφορετικών μοντέλων ταξινόμησης. Όλα βρίσκονται αποθηκευμένα και έτοιμα προς χρήση στο Google Drive, με τη βοήθεια της βιβλιοθήκης joblib. Τα μοντέλα και οι τεχνικές λεπτομέρειες του καθενός συνοψίζονται, για ευκολότερη κατανόηση στον παρακάτω πίνακα:

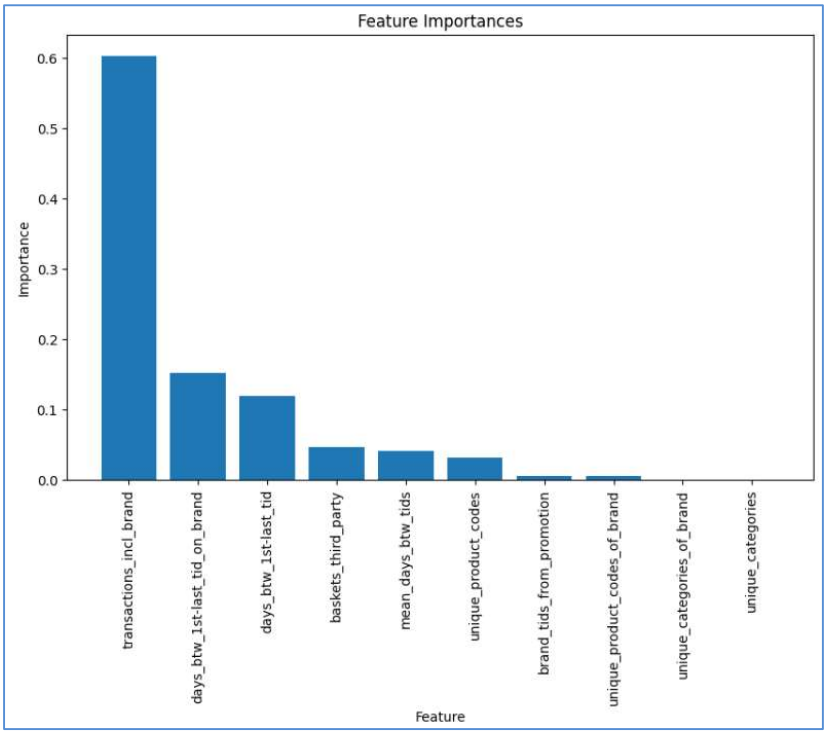
α/α	Όνομα	Μείωση Διαστάσεων	Δεδομένα Train	Δεδομένα Test	#Features	Παράμετροι (εκτός των default)
Logistic Regression						
1	lg	-	X_train	X_test	27	-
2	lg_pca	PCA	X_train_pca	X_test_pca	2	
3	lg_corr	CFS	X_train_corr	X_test_corr	10	
4	lg_rfe	RFE	X_train_rfe	X_test_rfe	10	
5	lg_rf	Tree Based	X_train_rf	X_test_rf	10	
6	lg_lasso	LASSO	X_train_lasso	X_test_lasso	10	
Naïve Bayes						
7	nb	-	X_train	X_test	27	-
8	nb_pca	PCA	X_train_pca	X_test_pca	2	
9	nb_corr	CFS	X_train_corr	X_test_corr	10	
10	nb_rfe	RFE	X_train_rfe	X_test_rfe	10	
11	nb_rf	Tree Based	X_train_rf	X_test_rf	10	
12	nb_lasso	LASSO	X_train_lasso	X_test_lasso	10	
Support Vector Machine						
13	svm	-	X_train	X_test	27	probability: true
14	svm_pca	PCA	X_train_pca	X_test_pca	2	
15	svm_corr	CFS	X_train_corr	X_test_corr	10	
16	svm_rfe	RFE	X_train_rfe	X_test_rfe	10	

17	svm_rf	Tree Based	X_train_rf	X_test_rf	10	
18	svm_lasso	LASSO	X_train_lasso	X_test_lasso	10	
Decision Tree						
19	dt	-	X_train	X_test	27	max_depth: 5, min_sample_split: 5
20	dt_pca	PCA	X_train_pca	X_test_pca	2	max_depth: 5, min_sample_split: 2
21	dt_corr	CFS	X_train_corr	X_test_corr	10	max_depth: 5, min_sample_split: 5
22	dt_rfe	RFE	X_train_rfe	X_test_rfe	10	max_depth: 5, min_sample_split: 20
23	dt_rf	Tree Based	X_train_rf	X_test_rf	10	max_depth: 5, min_sample_split: 2
24	dt_lasso	LASSO	X_train_lasso	X_test_lasso	10	max_depth: 5, min_sample_split: 5
Random Forest						
25	rf	-	X_train	X_test	27	n_estimators: 100, max_depth: 10, min_samples_split: 10, min_samples_leaf: 2
26	rf_pca	PCA	X_train_pca	X_test_pca	2	n_estimators: 100, max_depth: 10, min_samples_split: 10, min_samples_leaf: 4
27	rf_corr	CFS	X_train_corr	X_test_corr	10	n_estimators: 100, max_depth: 10, min_samples_split: 10, min_samples_leaf: 4
28	rf_rfe	RFE	X_train_rfe	X_test_rfe	10	n_estimators: 100, max_depth: 10, min_samples_split: 2, min_samples_leaf: 4
29	rf_rf	Tree Based	X_train_rf	X_test_rf	10	n_estimators: 100, max_depth: 10, min_samples_split: 10, min_samples_leaf: 4
30	rf_lasso	LASSO	X_train_lasso	X_test_lasso	10	n_estimators: 100, max_depth: 10, min_samples_split: 2, min_samples_leaf: 2
XGBoost						
31	xg	-	X_train	X_test	27	n_estimators: 300, max_depth: 4, learning_rate: 0.01
32	xg_pca	PCA	X_train_pca	X_test_pca	2	n_estimators: 100, max_depth: 5, learning_rate: 0.01
33	xg_corr	CFS	X_train_corr	X_test_corr	10	n_estimators: 200, max_depth: 3, learning_rate: 0.01
34	xg_rfe	RFE	X_train_rfe	X_test_rfe	10	n_estimators: 300, max_depth: 3, learning_rate: 0.01
35	xg_rf	Tree Based	X_train_rf	X_test_rf	10	n_estimators: 100, max_depth: 3, learning_rate:0.2
36	xg_lasso	LASSO	X_train_lasso	X_test_lasso	10	n_estimators: 300, max_depth: 5, learning_rate: 0.01
LightGBM						
37	lgb	-	X_train	X_test	27	n_estimators: 50, learning_rate: 0.05, num_leaves: 31
38	lgb_pca	PCA	X_train_pca	X_test_pca	2	n_estimators: 100, learning_rate: 0.01, num_leaves: 31
39	lgb_corr	CFS	X_train_corr	X_test_corr	10	n_estimators: 200, learning_rate: 0.01, num_leaves: 20

40	lgb_rfe	RFE	X_train_rfe	X_test_rfe	10	n_estimators: 200, learning_rate: 0.01, num_leaves: 20
41	lgb_rf	Tree Based	X_train_rf	X_test_rf	10	n_estimators: 50, learning_rate: 0.05, num_leaves: 20
42	lgb_lasso	LASSO	X_train_lasso	X_test_lasso	10	n_estimators: 50, learning_rate: 0.05, num_leaves: 40
Adaboost						
43	ada	-	X_train	X_test	27	n_estimators: 200, learning_rate: 0.5, base_estimator: decision tree
44	ada_pca	PCA	X_train_pca	X_test_pca	2	n_estimators: 50, learning_rate: 0.1, base_estimator: decision tree
45	ada_corr	CFS	X_train_corr	X_test_corr	10	n_estimators: 50, learning_rate: 0.1, base_estimator: decision tree
46	ada_rfe	RFE	X_train_rfe	X_test_rfe	10	n_estimators: 50, learning_rate: 0.1, base_estimator: decision tree
47	ada_rf	Tree Based	X_train_rf	X_test_rf	10	n_estimators: 200, learning_rate: 0.5, base_estimator: decision tree
48	ada_lasso	LASSO	X_train_lasso	X_test_lasso	10	n_estimators: 50, learning_rate: 0.1, base_estimator: decision tree
Voting Classifier						
49	vc	-	X_train	X_test	27	voting: 'soft' & estimators: all models trained with X_train
50	vc_pca	PCA	X_train_pca	X_test_pca	2	voting: 'soft' & estimators: all models trained with X_train_pca
51	vc_corr	CFS	X_train_corr	X_test_corr	10	voting: 'soft' & estimators: all models trained with X_train_corr
52	vc_rfe	RFE	X_train_rfe	X_test_rfe	10	voting: 'soft' & estimators: all models trained with X_train_rfe
53	vc_rf	Tree Based	X_train_rf	X_test_rf	10	voting: 'soft' & estimators: all models trained with X_train_rf
54	vc_lasso	LASSO	X_train_lasso	X_test_lasso	10	voting: 'soft' & estimators: all models trained with X_train_lasso

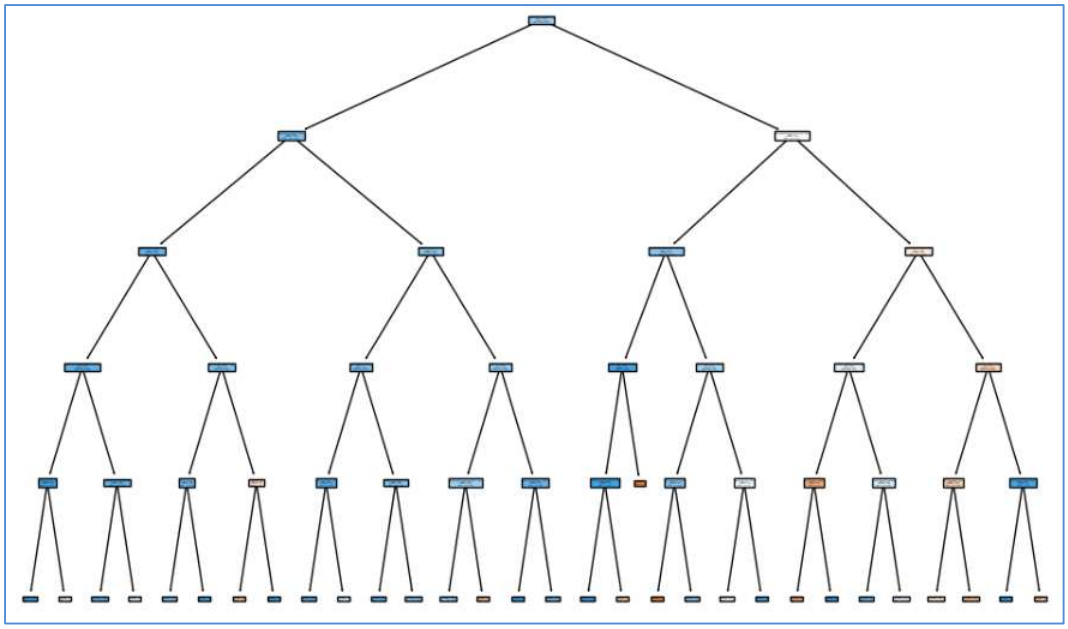
Πίνακας 5-Λίστα Μοντέλων Ταξινόμησης

Κατά τη διαδικασία της εκπαίδευσης των παραπάνω μοντέλων μας παρέχεται πρόσβαση σε κάποιες οπτικοποιήσεις που βοηθούν στην καλύτερη κατανόηση της διαδικασίας. Για παράδειγμα, κατά την κατασκευή των διαφόρων δέντρων απόφασης βλέπουμε ένα διάγραμμα στο οποίο κατατάσσονται τα χαρακτηριστικά του συνόλου εκπαίδευσης ανάλογα με τη σημαντικότητά τους για το μοντέλο. Στην εικόνα 15 βλέπουμε συγκεκριμένα το αντίστοιχο διάγραμμα που παράχθηκε κατά την εκπαίδευση του μοντέλου «dt_corr» (δέντρο απόφασης με τη χρήση του συνόλου εκπαίδευσης «X_train_corr». Εδώ τα 3 χαρακτηριστικά που αναδεικνύονται ως τα σημαντικότερα για την κατάταξη ενός δείγματος σε μία από τις 2 κατηγορίες είναι – κατά φθίνουσα σημαντικότητα - τα: transactions_incl_brand, days_btw_1st-last_tid_on_brand και days_btw_1st-last_tid.



Εικόνα 15-Σημαντικότητα Χαρακτηριστικών στην Εκπαίδευση Δέντρου Απόφασης με το «X_train_corr»

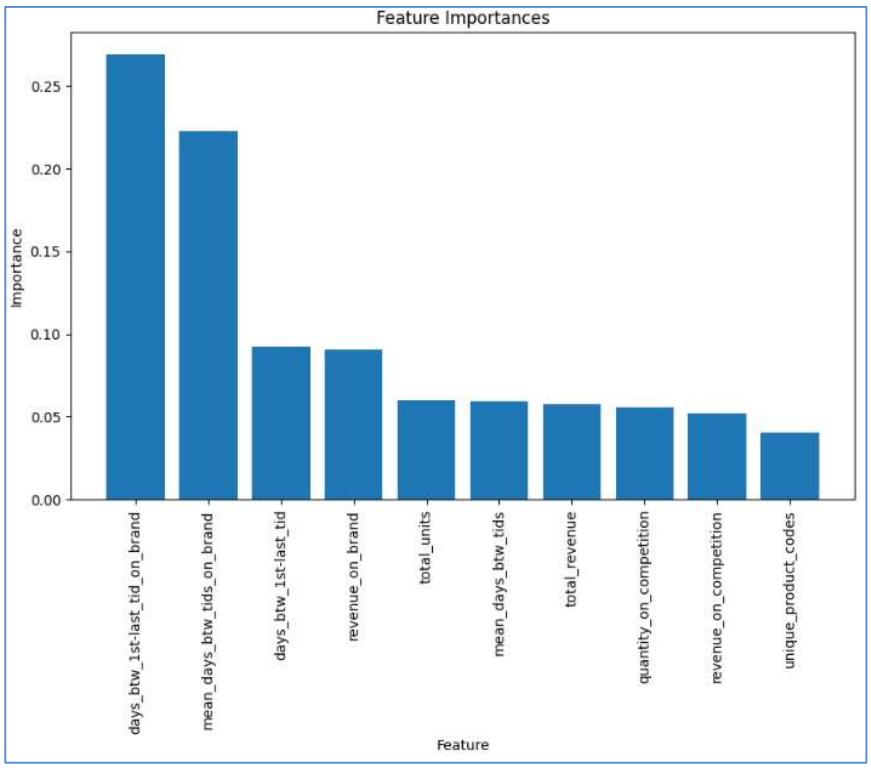
Κατά την κατασκευή των δέντρων απόφασης μπορούμε επιπλέον να δούμε οπτικοποιημένο ολόκληρο το τελικό δέντρο. Για παράδειγμα στην εικόνα 16 μπορούμε να δούμε την οπτική αναπαράσταση του μοντέλου «rf» (δέντρο απόφασης με τη χρήση του συνόλου εκπαίδευσης «X_train».



Εικόνα 16-Οπτική Αναπαράσταση Δέντρου Απόφασης με Χρήση του «X_train»

Με παρόμοιο τρόπο, μπορούμε να δούμε την κατάταξη των χαρακτηριστικών του συνόλου εκπαίδευσης κατά φθίνουσα σημαντικότητα κατά την εκπαίδευση ενός μοντέλου με τον αλγόριθμο XGBoost. Στην εικόνα 17 βλέπουμε συγκεκριμένα το αντίστοιχο διάγραμμα που παράχθηκε κατά την εκπαίδευση του μοντέλου «xg_rf» (μοντέλο XGBoost με τη χρήση του συνόλου εκπαίδευσης «X_train_rf». Εδώ τα 3 χαρακτηριστικά που αναδεικνύονται ως τα σημαντικότερα για την κατάταξη ενός δείγματος σε μία από τις 2 κατηγορίες είναι – κατά

φθίνουσα σημαντικότητα - τα: days_btw_1st-last_tid_on_brand, mean_days_btw_tids_on_brand, και days_btw_1st-last_tid.



Εικόνα 17- Σημαντικότητα Χαρακτηριστικών στην Εκπαίδευση Μοντέλου XGBoost με το «X_train_rf»

Κεφάλαιο 5 - Συμπεράσματα

5.1 Αποτελέσματα Μηχανικής Μάθησης

Για την αξιολόγηση των αποτελεσμάτων των μοντέλων θα πορευτούμε με την παραδοχή ότι μας ενδιαφέρει εξίσου η σωστή πρόβλεψη και των 2 κλάσεων (churners/ non churners). Δεδομένου αυτού, ο δείκτης F1-score μας ενδιαφέρει περισσότερο καθώς ισορροπεί μεταξύ ακρίβειας και ανάκλησης.

Ωστόσο κάποια άλλη μετρική θα μπορούσε να έχει μεγαλύτερη βαρύτητα, ανάλογα με τις συγκεκριμένες ανάγκες μιας επιχείρησης. Για παράδειγμα, εάν θέλουμε να δώσουμε προτεραιότητα στην ελαχιστοποίηση των ψευδώς θετικών προβλέψεων (πρόβλεψη ότι ένας πελάτης θα αποχωρήσει όταν δεν θα το κάνει), θα προτιμήσουμε μοντέλα με υψηλότερη ειδικότητα (specificity).

Έχοντας κατά νου τη βαρύτητα του f1-score, ακολουθεί η αποτίμηση των 54 μοντέλων που δημιουργήθηκαν, χωρισμένα ανά επιλογή χαρακτηριστικών για το σύνολο δεδομένων εκπαίδευσης. Σε κάθε πίνακα σημειώνονται με bold οι 2 καλύτερες τιμές ανά μετρική, η καλύτερη από τις 2 με το σκουρότερο χρώμα. Τα μοντέλα με ειδικότητα 0 ή 1 απορρίπτονται εξ αρχής καθώς από τους πίνακες σύγχυσής τους διαπιστώνεται ότι οι τιμές οφείλονται στην κατάταξη όλων των δειγμάτων από το μοντέλο σε μία και μόνο κλάση.

5.1.1 Αποτελέσματα με Χρήση Όλων των Χαρακτηριστικών

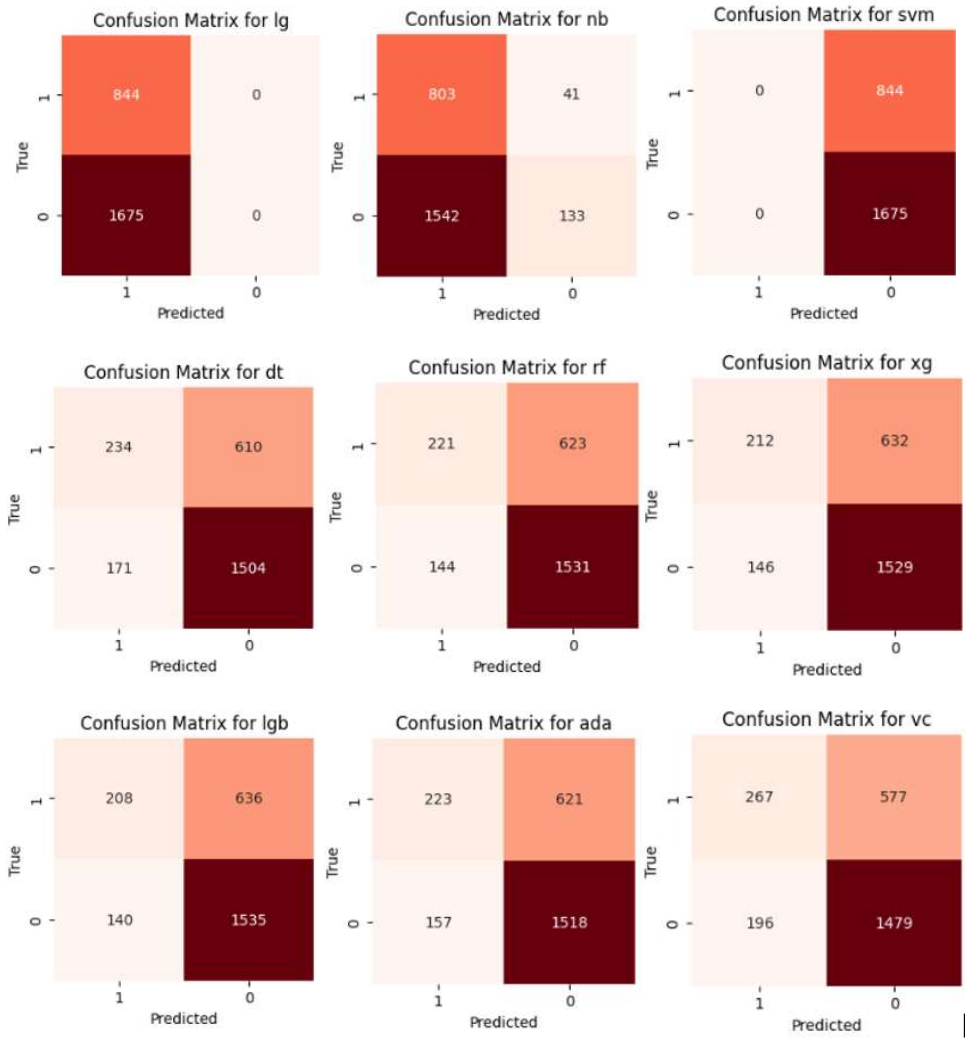
Όνομα	Accuracy	Precision	Recall	Specificity	F1-score	MCC	AUC
lg	0.335	0.112	0.335	~~1.000~~	0.168	0.000	0.64
nb	0.372	0.623	0.372	**0.951**	0.264	0.057	0.63
svm	0.665	0.442	0.665	~~0.000~~	0.531	0.000	0.66
dt	0.690	0.667	0.690	0.277	0.653	0.225	0.65
rf	**0.696**	**0.676**	**0.696**	0.269	**0.654**	**0.236**	**0.67**
xg	0.691	0.669	0.691	0.251	0.648	0.222	**0.68**
lgb	0.692	0.670	0.692	0.246	0.648	0.223	**0.67**
ada	0.691	0.669	0.691	0.264	0.651	0.225	0.66
vc	**0.693**	**0.672**	**0.693**	**0.316**	**0.664**	**0.243**	**0.68**

Πίνακας 6-Μετρικές Αξιολόγησης Μοντέλων που εκπαιδεύτηκαν με το «X_train»

Αναφορικά με τα μοντέλα που εκπαιδεύτηκαν χρησιμοποιώντας το σύνολο των χαρακτηριστικών του συνόλου δεδομένων βλέπουμε πως οι καλύτερες μετρικές απόδοσης προκύπτουν από το μοντέλο «rf» και από το «vc». Συγκεκριμένα το πρώτο επιδεικνύει καλύτερη συνολική ακρίβεια πρόβλεψης ενώ το δεύτερο δίνει ελαφρώς πιο ισορροπημένες προβλέψεις καθώς έχει λίγο υψηλότερο f1-score.

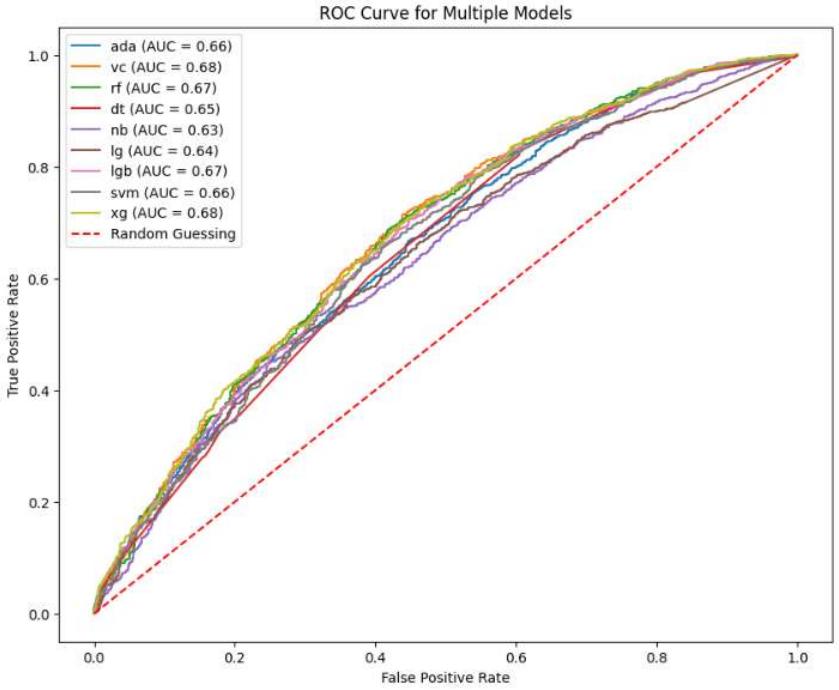
Παρατίθενται επιπλέον οι πίνακες σύγχυσης και οι καμπύλες ROC των μοντέλων που κατασκευάστηκαν. Από τους πίνακες σύγχυσης βλέπουμε, όπως αναμένεται, ότι για το μοντέλο «rf», το οποίο επιδεικνύει τη μεγαλύτερη ακρίβεια, το άθροισμα των αληθώς θετικών και των αληθώς αρνητικών προβλέψεων (TP+TN) είναι το μεγαλύτερο. Αντίστοιχα παρατηρούμε για το μοντέλο «vc», το οποίο επιδεικνύει το καλύτερο f1-score, ότι το άθροισμα των ψευδώς θετικών και των ψευδώς αρνητικών προβλέψεων (FP+FN) είναι το μικρότερο.

Βλέπουμε επιπλέον για το μοντέλο «lg» που απορρίφτηκε ότι δεν έχει επιδείξει ούτε μία πρόβλεψη για την κλάση 0. Τέλος, βλέπουμε για το μοντέλο «svm», που επίσης απορρίφτηκε, ότι δεν έχει επιδείξει ούτε μία πρόβλεψη για την κλάση 1.



Εικόνα 18-Πίνακες Σύγχυσης Μοντέλων που εκπαιδεύτηκαν με το «X_train»

Παρατηρώντας το διάγραμμα με τις καμπύλες ROC βλέπουμε πως οι καμπύλες των ταξινομητών με τις καλύτερες μετρικές αξιολόγησης είναι και οι 2 πιο μετατοπισμένες προς την πάνω αριστερή γωνία του τετραγώνου, δηλαδή το σημείο (0,1), ένδειξη πως πρόκειται για τα 2 καλύτερα διαθέσιμα μοντέλα. Αντίθετα όσο χαμηλότερες μετρικές αξιολόγησης έχουν τα υπόλοιπα μοντέλα, τόσο πιο κοντά βρίσκονται στην ευθεία που υποδηλώνει την τυχαία πρόβλεψη.



Εικόνα 19-Καμπύλη ROC Μοντέλων που εκπαιδεύτηκαν με το «X_train»

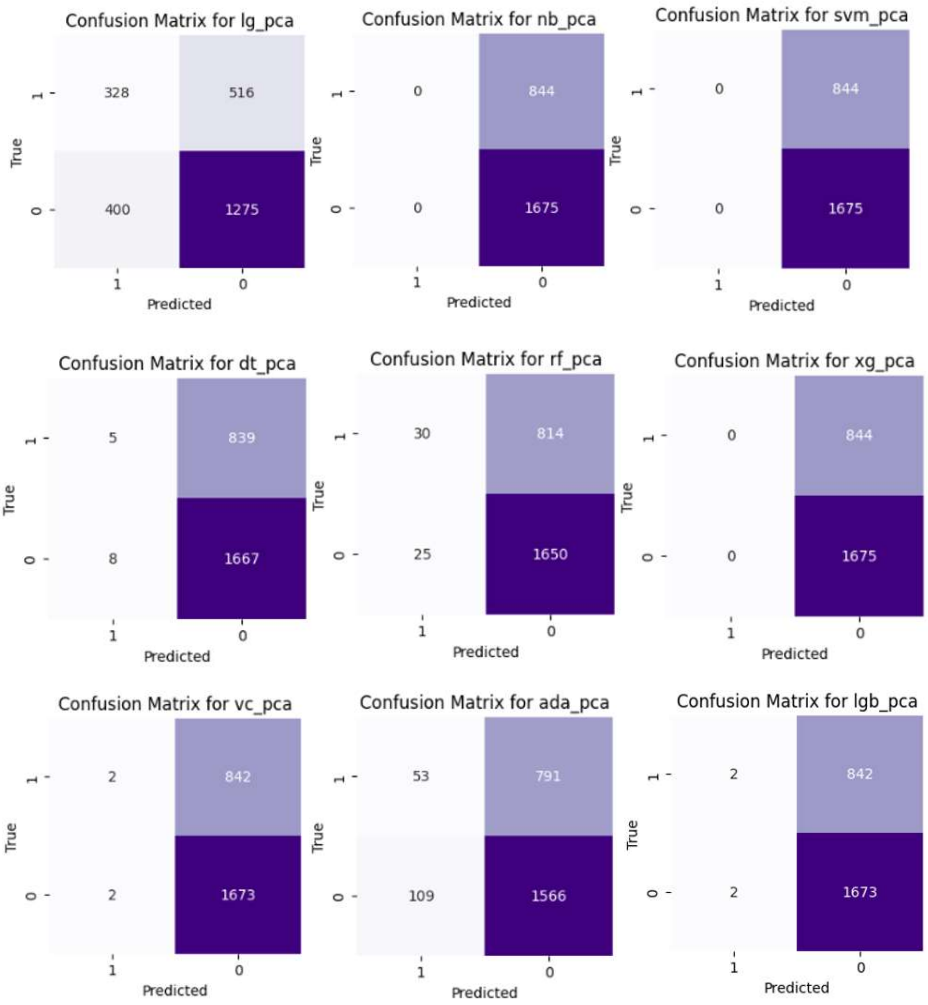
5.1.2 Αποτελέσματα μετά από Ανάλυση Κυρίων Συνιστωσών

Όνομα	Accuracy	Precision	Recall	Specificity	F1-score	MCC	AUC
lg_pca	0.636	**0.624**	0.636	**0.389**	**0.629**	**0.156**	**0.60**
nb_pca	**0.665**	0.442	**0.665**	~~0.000~~	0.531	0.000	0.53
svm_pca	**0.665**	0.442	**0.665**	~~0.000~~	0.531	0.000	0.50
dt_pca	0.664	0.571	0.664	0.006	0.534	0.008	**0.61**
rf_pca	**0.667**	**0.628**	**0.667**	0.036	**0.553**	**0.067**	**0.60**
xg_pca	**0.665**	0.442	**0.665**	~~0.000~~	0.531	0.000	**0.61**
lgb_pca	**0.665**	0.610	**0.665**	0.002	0.533	0.014	**0.60**
ada_pca	0.643	0.551	0.643	**0.063**	0.552	-0.004	0.58
vc_pca	**0.665**	0.610	**0.665**	0.002	0.533	0.014	**0.61**

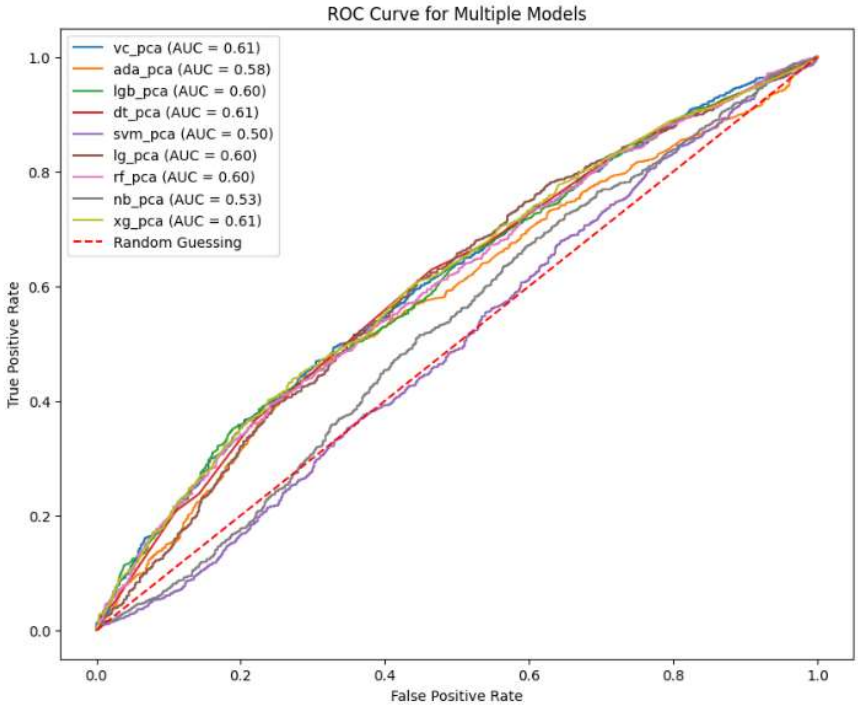
Πίνακας 7-Μετρικές Αξιολόγησης Μοντέλων που εκπαιδεύτηκαν με το «X_train_pca»

Αναφορικά με τα μοντέλα που εκπαιδεύτηκαν ύστερα από την ανάλυση κυρίων συνιστωσών βλέπουμε πως οι καλύτερες μετρικές απόδοσης προκύπτουν από το μοντέλο «rf_pca» και από το «lg_pca». Συγκεκριμένα το πρώτο επιδεικνύει καλύτερη συνολική ακρίβεια πρόβλεψης ενώ το δεύτερο δίνει ελαφρώς πιο ισορροπημένες προβλέψεις καθώς έχει λίγο υψηλότερο f1-score.

Παρατίθενται επιπλέον οι πίνακες σύγχυσης και οι καμπύλες ROC των μοντέλων που κατασκευάστηκαν. Από τους πίνακες σύγχυσης βλέπουμε, όπως αναμένεται, ότι για το μοντέλο «rf_pca», το οποίο επιδεικνύει τη μεγαλύτερη ακρίβεια, το άθροισμα των αληθώς θετικών και των αληθώς αρνητικών προβλέψεων (TP+TN) είναι το μεγαλύτερο. Αντίστοιχα παρατηρούμε για το μοντέλο «lg_pca», το οποίο επιδεικνύει το καλύτερο f1-score, ότι το άθροισμα των ψευδώς θετικών και των ψευδώς αρνητικών προβλέψεων (FP+FN) είναι το μικρότερο. Βλέπουμε επιπλέον ότι και τα 3 μοντέλα που απορρίφτηκαν δεν κατάφεραν να κάνουν ούτε μία πρόβλεψη για την κλάση 1.



Εικόνα 20-Πίνακες Σύγχυσης Μοντέλων που εκπαιδεύτηκαν με το «X_train_pca»



Εικόνα 21-Καμπύλη ROC Μοντέλων που εκπαιδεύτηκαν με το «X_train_pca»

Παρατηρώντας το διάγραμμα με τις καμπύλες ROC βλέπουμε πως η καμπύλη του μοντέλου λογιστικής παλινδρόμησης είναι η πιο μετατοπισμένη προς την πάνω αριστερή γωνία του τετραγώνου, δηλαδή το σημείο (0,1), ένδειξη πως πρόκειται για το καλύτερο διαθέσιμο μοντέλο. Αντίθετα όσο χαμηλότερες μετρικές αξιολόγησης έχουν τα υπόλοιπα μοντέλα, τόσο πιο κοντά βρίσκονται στην ευθεία που υποδηλώνει την τυχαία πρόβλεψη. Μάλιστα σε κάποιες περιπτώσεις βρίσκονται ακόμα και κάτω από αυτή.

5.1.3 Αποτελέσματα μετά από Επιλογή Χαρακτηριστικών

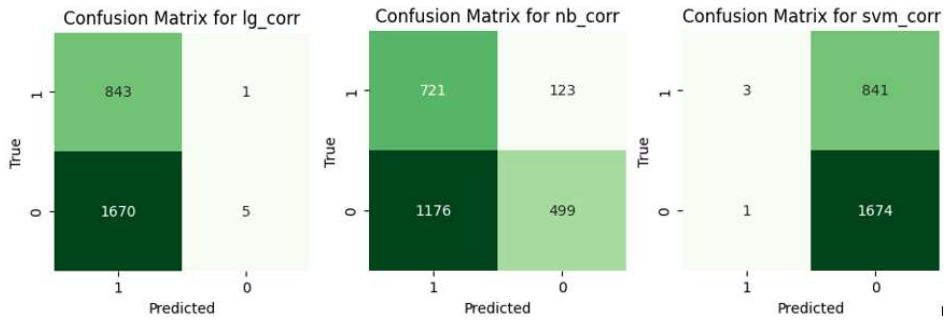
Επιλογή Χαρακτηριστικών Βάσει Συσχέτισης

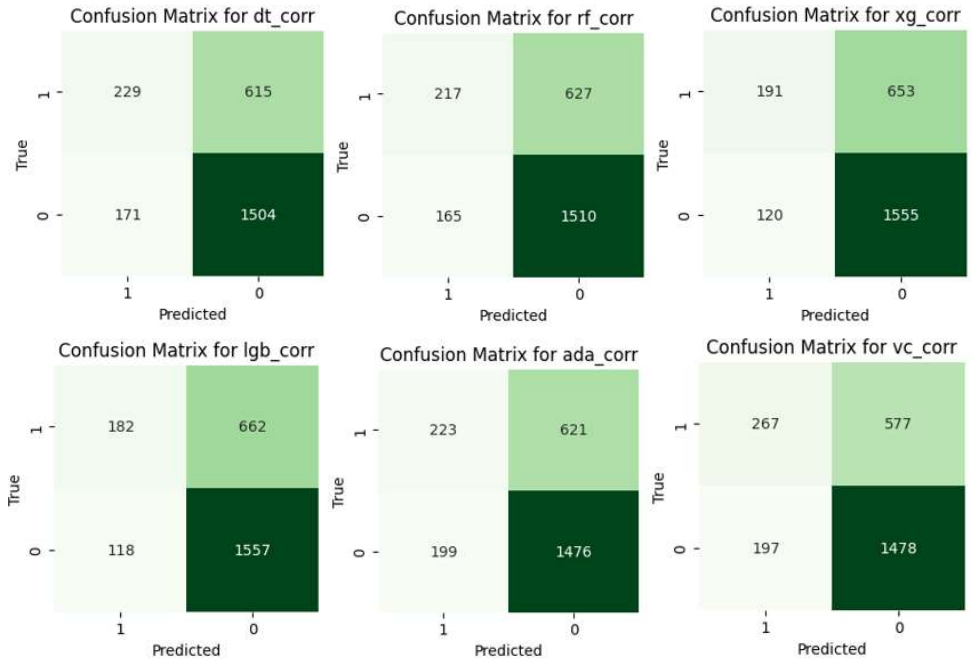
Όνομα	Accuracy	Precision	Recall	Specificity	F1-score	MCC	AUC
lg_corr	0.337	0.667	0.337	**0.999**	0.172	0.017	0.62
nb_corr	0.484	0.661	0.484	**0.854**	0.465	0.167	0.64
svm_corr	0.666	**0.694**	0.666	0.004	0.534	0.035	0.64
dt_corr	0.688	0.664	0.688	0.271	**0.651**	0.219	0.65
rf_corr	0.686	0.660	0.686	0.257	0.645	0.209	**0.66**
xg_corr	**0.693**	**0.674**	**0.693**	0.226	0.643	**0.222**	**0.67**
lgb_corr	**0.690**	0.670	**0.690**	0.216	0.638	0.212	**0.67**
ada_corr	0.675	0.645	0.675	0.264	0.638	0.184	0.64
vc_corr	**0.693**	0.671	**0.693**	0.316	**0.664**	**0.242**	**0.67**

Πίνακας 8-Μετρικές Αξιολόγησης Μοντέλων που εκπαιδεύτηκαν με το «X_train_corr»

Ανάμεσα στα μοντέλα που εκπαιδεύτηκαν μετά από την επιλογή χαρακτηριστικών βάσει συσχέτισης (correlation-based feature selection) βλέπουμε πως οι καλύτερες μετρικές απόδοσης προκύπτουν από το «dt_corr», το «xg_corr» και το «vc_corr». Συγκεκριμένα τα 2 πρώτα υπερισχύουν ως προς τη συνολική ακρίβεια πρόβλεψης ενώ το τρίτο υπερισχύει και ως προς τη μετρική f1-score.

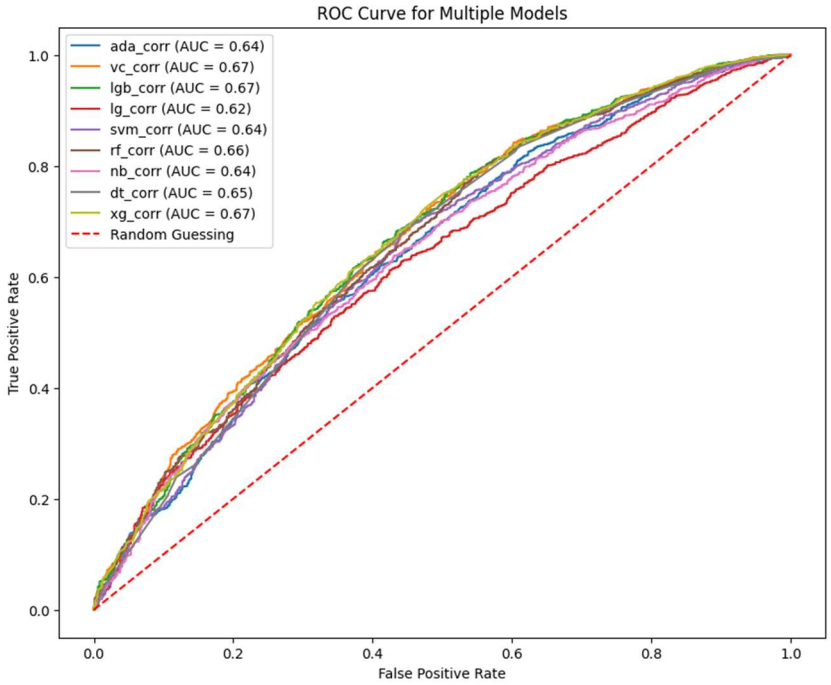
Παρατίθενται επιπλέον οι πίνακες σύγχυσης και οι καμπύλες ROC των μοντέλων που κατασκευάστηκαν. Από τους πίνακες σύγχυσης βλέπουμε, όπως αναμένεται, ότι για τα μοντέλα «dt_corr» και «xg_corr», τα οποία επιδεικνύουν τη μεγαλύτερη ακρίβεια, τα αθροίσματα των αληθώς θετικών και των αληθώς αρνητικών προβλέψεων (TP+TN) είναι τα μεγαλύτερα. Αντίστοιχα παρατηρούμε για το μοντέλο «vc_corr», το οποίο επιδεικνύει το καλύτερο f1-score, ότι το άθροισμα των ψευδώς θετικών και των ψευδώς αρνητικών προβλέψεων (FP+FN) είναι το μικρότερο.





Εικόνα 22- Πίνακες Σύγχυσης Μοντέλων που εκπαιδεύτηκαν με το «X_train_corr»

Παρατηρώντας το διάγραμμα με τις καμπύλες ROC βλέπουμε και πάλι πως οι καμπύλες των μοντέλων με τις καλύτερες μετρικές είναι πιο μετατοπισμένες προς την πάνω αριστερή γωνία του τετραγώνου, δηλαδή το σημείο (0,1), πράγμα που επιβεβαιώνει την υπεροχή τους.



Εικόνα 23-Καμπύλη ROC Μοντέλων που εκπαιδεύτηκαν με το «X_train_corr»

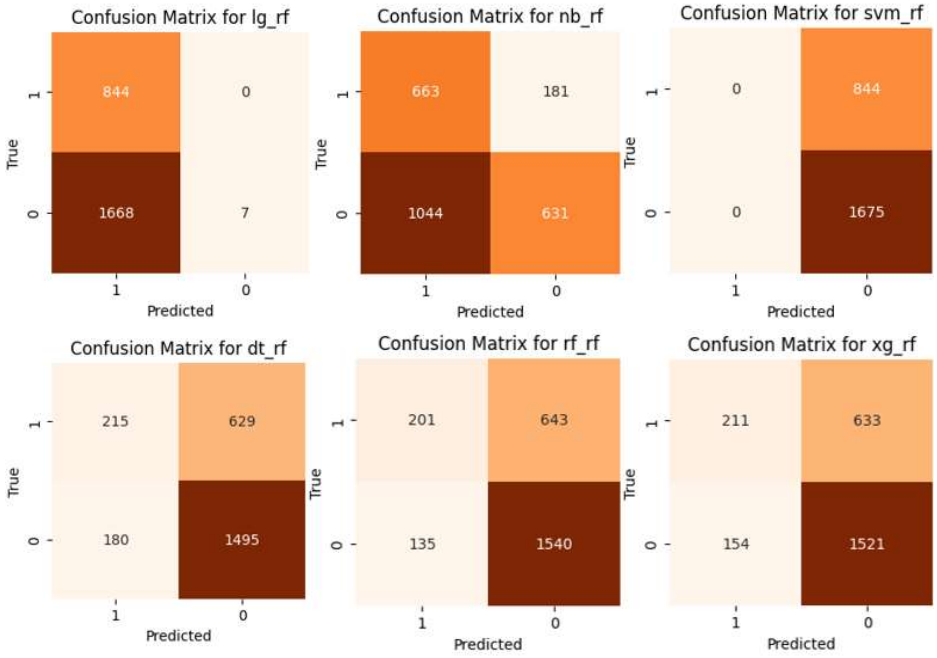
Επιλογή Χαρακτηριστικών Βάσει Δέντρου

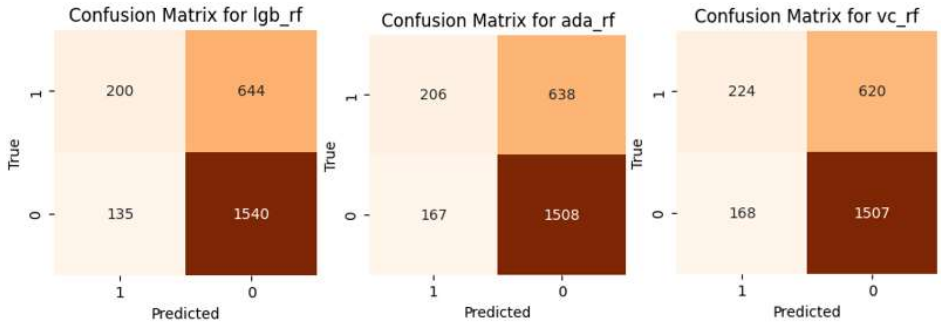
Όνομα	Accuracy	Precision	Recall	Specificity	F1-score	MCC	AUC
lg_rf	0.338	**0.777**	0.338	~~1.000~~	0.174	0.038	0.63
nb_rf	0.514	0.647	0.514	**0.786**	0.512	0.164	0.63
svm_rf	0.665	0.442	0.665	~~0.000~~	0.531	0.000	0.66
dt_rf	0.679	0.650	0.679	0.255	0.640	0.191	0.65
rf_rf	**0.691**	**0.670**	**0.691**	0.238	**0.645**	**0.219**	**0.68**
xg_rf	**0.688**	0.663	**0.688**	0.250	**0.645**	0.212	**0.67**
lgb_rf	**0.691**	0.669	**0.691**	0.237	0.644	**0.217**	**0.67**
ada_rf	0.680	0.652	0.680	0.244	0.638	0.192	0.65
vc_rf	0.687	0.663	0.687	**0.265**	**0.649**	0.215	**0.67**

Πίνακας 9-Μετρικές Αξιολόγησης Μοντέλων που εκπαιδεύτηκαν με το «X_train_rf»

Μεταξύ των μοντέλων που εκπαιδεύτηκαν μετά από την επιλογή χαρακτηριστικών βάσει δέντρου (tree based feature importance) βλέπουμε πως οι καλύτερες μετρικές απόδοσης προκύπτουν από τα μοντέλα «rf_rf», «lgb_rf» και «vc_rf». Συγκεκριμένα τα 2 πρώτα υπερισχύουν ως προς τη συνολική ακρίβεια πρόβλεψης ενώ η το τρίτο δίνει ελαφρώς πιο ισορροπημένες προβλέψεις καθώς έχει λίγο υψηλότερο f1-score.

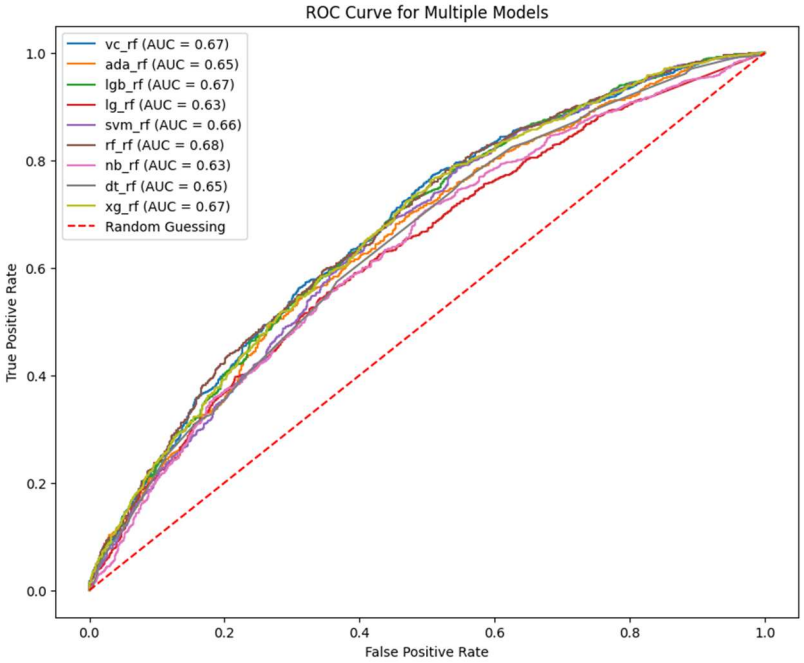
Παρατίθενται επιπλέον οι πίνακες σύγχυσης και οι καμπύλες ROC των μοντέλων που κατασκευάστηκαν. Από τους πίνακες σύγχυσης βλέπουμε, όπως αναμένεται, ότι για τα μοντέλα «rf_rf» και «lgb_rf», τα οποία επιδεικνύουν τη μεγαλύτερη ακρίβεια, τα αθροίσματα των αληθώς θετικών και των αληθώς αρνητικών προβλέψεων (TP+TN) είναι τα μεγαλύτερα. Αντίστοιχα παρατηρούμε για το μοντέλο «vc_rf», το οποίο επιδεικνύει το καλύτερο f1-score, ότι το άθροισμα των ψευδώς θετικών και των ψευδώς αρνητικών προβλέψεων (FP+FN) είναι το μικρότερο. Βλέπουμε επιπλέον ότι το μοντέλο «svm_rf» που απορρίφτηκε δεν κατάφερε να κάνει ούτε μία πρόβλεψη για την κλάση 1.





Εικόνα 24- Πίνακες Σύγχυσης Μοντέλων που εκπαιδεύτηκαν με το «X_train_rf»

Παρατηρώντας το διάγραμμα με τις καμπύλες ROC βλέπουμε για άλλη μία φορά πως οι καμπύλες των μοντέλων με τις καλύτερες μετρικές είναι πιο μετατοπισμένες προς την πάνω αριστερή γωνία του τετραγώνου, δηλαδή το σημείο (0,1), πράγμα που επιβεβαιώνει την υπεροχή τους.



Εικόνα 25-Καμπύλη ROC Μοντέλων που εκπαιδεύτηκαν με το «X_train_rf»

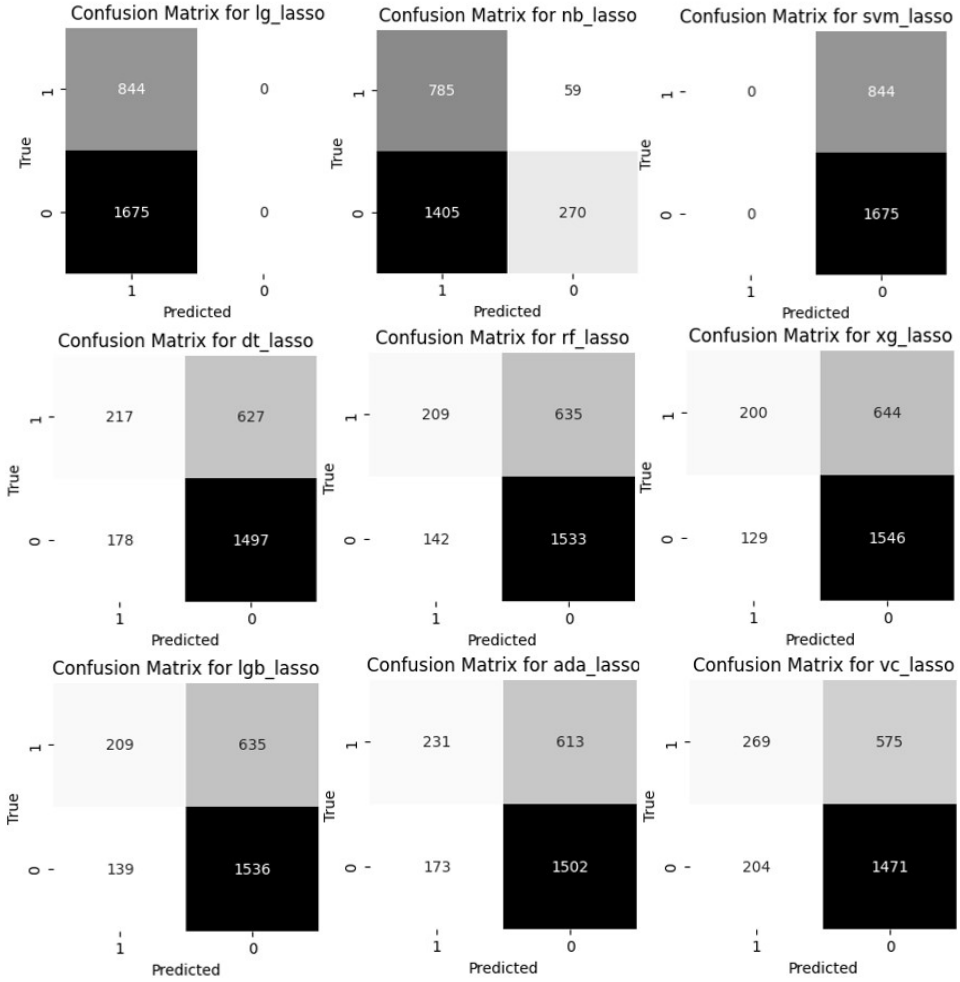
Κανονικοποίηση LASSO

Όνομα	Accuracy	Precision	Recall	Specificity	F1-score	MCC	AUC
lg_lasso	0.335	0.112	0.335	~~1.000~~	0.168	0.000	0.63
nb_lasso	0.419	0.666	0.419	**0.930**	0.353	0.128	0.65
svm_lasso	0.665	0.442	0.665	~~0.000~~	0.531	0.000	0.66
dt_lasso	0.680	0.653	0.680	0.257	0.641	0.196	0.65
rf_lasso	0.692	0.670	0.692	0.248	0.648	0.222	**0.67**
xg_lasso	**0.693**	**0.673**	**0.693**	0.237	0.646	0.224	**0.67**
lgb_lasso	**0.693**	0.672	**0.693**	0.248	0.649	**0.225**	**0.67**
ada_lasso	0.688	0.664	0.688	0.274	0.651	0.219	0.64
vc_lasso	0.691	0.669	0.691	**0.319**	**0.663**	0.238	**0.67**

Πίνακας 10-Μετρικές Αξιολόγησης Μοντέλων που εκπαιδεύτηκαν με το «X_train_lasso»

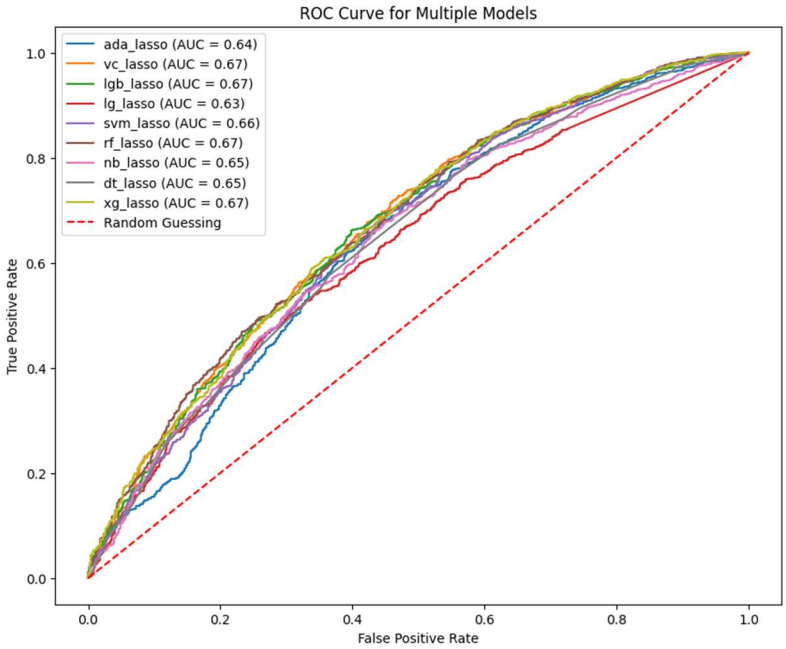
Όσον αφορά τα μοντέλα που εκπαιδεύτηκαν μετά από την επιλογή χαρακτηριστικών με τη χρήση κανονικοποίησης lasso υπερισχύουν τα μοντέλα «xg_lasso», «lgb_lasso» και «vc_lasso». Συγκεκριμένα τα 2 πρώτα δίνουν καλύτερη συνολική ακρίβεια πρόβλεψης ενώ η το τρίτο δίνει ελαφρώς πιο ισορροπημένες προβλέψεις καθώς έχει λίγο υψηλότερο f1-score.

Παρατίθενται επιπλέον οι πίνακες σύγχυσης και οι καμπύλες ROC των μοντέλων που κατασκευάστηκαν. Από τους πίνακες σύγχυσης βλέπουμε, όπως αναμένεται, ότι για τα μοντέλα «xg_lasso» και «lgb_lasso», τα οποία επιδεικνύουν τη μεγαλύτερη ακρίβεια, τα αθροίσματα των αληθώς θετικών και των αληθώς αρνητικών προβλέψεων (TP+TN) είναι τα μεγαλύτερα. Αντίστοιχα παρατηρούμε για το μοντέλο «vc_lasso», το οποίο επιδεικνύει το καλύτερο f1-score, ότι το άθροισμα των ψευδώς θετικών και των ψευδώς αρνητικών προβλέψεων (FP+FN) είναι το μικρότερο. Βλέπουμε επιπλέον τα 2 μοντέλα που απορρίφτηκαν αποτυγχάνουν να κάνουν έστω και μία πρόβλεψη για μία από της 2 κλάσεις – το «lg_lasso» για τη 0 και το «svm_lasso» για την 1 αντιστοίχως.



Εικόνα 26- Πίνακες Σύγχυσης Μοντέλων που εκπαιδεύτηκαν με το «X_train_lasso»

Παρατηρώντας το διάγραμμα με τις καμπύλες ROC βλέπουμε για άλλη μία φορά πως οι καμπύλες των μοντέλων με τις καλύτερες μετρικές είναι πιο μετατοπισμένες προς την πάνω αριστερή γωνία του τετραγώνου, δηλαδή το σημείο (0,1), πράγμα που επιβεβαιώνει ότι πρόκειται για τα καλύτερα διαθέσιμα μοντέλα.



Εικόνα 27-Καμπύλη ROC Μοντέλων που εκπαιδεύτηκαν με το «X_train_lasso»

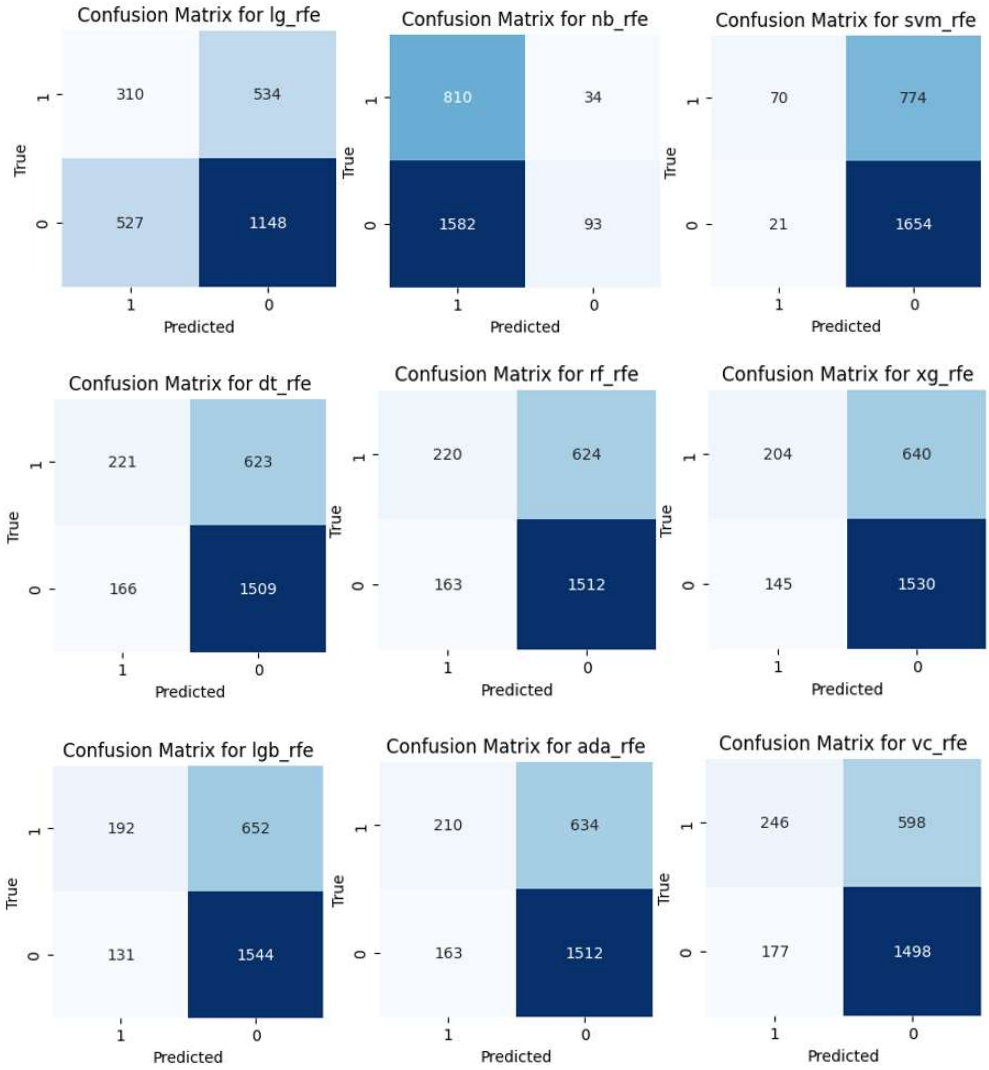
Σταδιακός Αποκλεισμός Χαρακτηριστικών

Όνομα	Accuracy	Precision	Recall	Specificity	F1-score	MCC	AUC
lg_rfe	0.579	0.578	0.579	0.367	0.578	0.053	0.54
nb_rfe	0.359	0.600	0.359	**0.960**	0.236	0.033	0.56
svm_rfe	0.684	**0.711**	0.684	0.083	0.586	0.178	0.65
dt_rfe	0.687	0.662	0.687	0.262	0.647	0.213	0.65
rf_rfe	0.688	0.663	0.688	0.261	0.648	0.215	**0.66**
xg_rfe	0.688	0.665	0.688	0.242	0.644	0.212	**0.66**
lgb_rfe	0.689	0.667	0.689	0.228	0.641	0.211	**0.66**
ada_rfe	0.684	0.657	0.684	0.249	0.642	0.201	0.64
vc_rfe	**0.692**	0.670	**0.692**	0.292	**0.658**	**0.235**	**0.66**

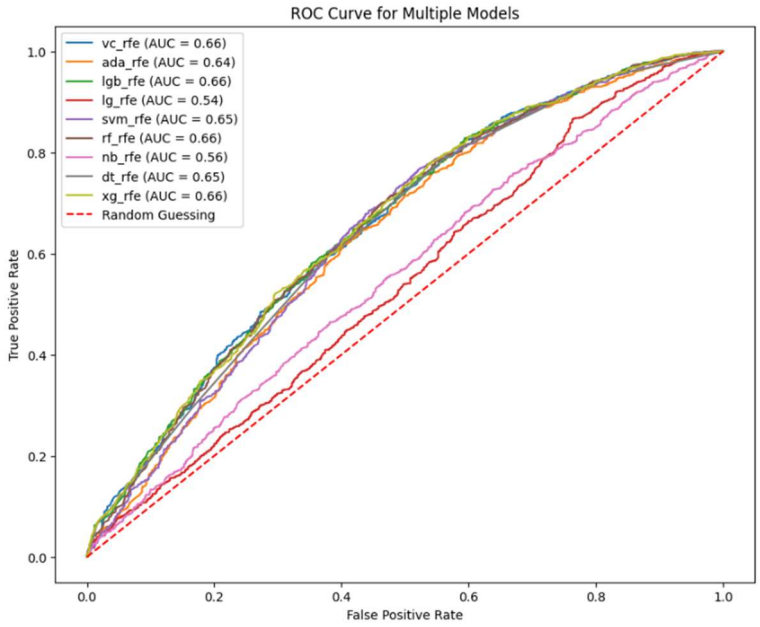
Πίνακας 11-Μετρικές Αξιολόγησης Μοντέλων που εκπαιδεύτηκαν με το «X_train_rfe»

Τέλος, ανάμεσα στα μοντέλα που εκπαιδεύτηκαν μετά από την επιλογή χαρακτηριστικών με σταδιακό αποκλεισμό (recursive feature elimination), ξεχωρίζουν τα «lgb_rfe» και «vc_rfe». Το «vc_rfe» υπερέχει, συγκεκριμένα τόσο ως προς τη συνολική ακρίβεια πρόβλεψης όσο και ως προς τη μετρική f1-score.

Αναμενόμενα, κοιτώντας τους πίνακες σύγχυσης που παρατίθενται, βλέπουμε πως για το μοντέλο «vc_rfe» το άθροισμα των αληθώς θετικών και των αληθώς αρνητικών προβλέψεων (TP+TN) είναι το μεγαλύτερο ενώ το άθροισμα των ψευδώς θετικών και των ψευδώς αρνητικών προβλέψεων (FP+FN) είναι το μικρότερο.



Εικόνα 28- Πίνακες Σύγχυσης Μοντέλων που εκπαιδεύτηκαν με το «X_train_rfe»



Εικόνα 29-Καμπύλη ROC Μοντέλων που εκπαιδεύτηκαν με το «X_train_rfe»

Η καμπύλη ROC του «vc_rfe» φαίνεται να είναι και η πιο μετατοπισμένη προς την πάνω αριστερή γωνία του τετραγώνου στο σχετικό διάγραμμα, πράγμα που δείχνει πως πρόκειται για τον πιο δυνατό ταξινομητή.

5.1.4 Σύγκριση Αποτελεσμάτων

Εφόσον έχουμε ορίσει τη μετρική f1-score ως τη σημαντικότερη για την αξιολόγηση των μοντέλων ταξινόμησης, συγκεντρώνουμε σε αυτό το σημείο τα μοντέλα που υπερέχουν ως προς τη μετρική αυτή σε κάθε μία από τις προηγούμενες υποενότητες. Συγκεντρώνονται έτσι στον παρακάτω πίνακα τα 6 καλύτερα μοντέλα που εκπαιδεύτηκαν, το καλύτερο ανά τρόπο επιλογής χαρακτηριστικών για τα δεδομένα εκπαίδευσης.

Όνομα	Accuracy	Precision	Recall	Specificity	F1-score	MCC	AUC
vc	**0.693**	**0.672**	**0.693**	0.316	**0.664**	**0.243**	**0.68**
lg_pca	0.636	0.624	0.636	**0.389**	0.629	0.156	0.60
vc_corr	**0.693**	0.671	**0.693**	0.316	**0.664**	0.242	0.67
vc_rf	0.687	0.663	0.687	0.265	0.649	0.215	0.67
vc_lasso	0.691	0.669	0.691	0.319	0.663	0.238	0.67
vc_rfe	0.692	0.670	0.692	0.292	0.658	0.235	0.66

Πίνακας 12-Μετρικές Αξιολόγησης των 6 καλύτερων Μοντέλων (βάσει f1-score)

Παρατηρούμε ότι σε όλες τις περιπτώσεις, εκτός από τη χρήση ανάλυσης κυρίων συνιστωσών, λαμβάνουμε πάντα καλύτερα αποτελέσματα από το συνδυασμό πολλών ταξινομητών, μέσω του μοντέλου voting classifier, παρά από κάποιο μεμονωμένο ταξινομητή. Κάτι ακόμα που παρατηρούμε είναι πως παρά τη χρήση τεχνικών μείωσης διαστάσεων και επιλογής χαρακτηριστικών έχουμε, τελικά, καλύτερα αποτελέσματα όταν χρησιμοποιούμε το σύνολο δεδομένων εκπαίδευσης στην αρχική του μορφή, χωρίς να αφαιρέσουμε χαρακτηριστικά. Αυτό δε σημαίνει ότι απορρίπτουμε τις τεχνικές αυτές. Οι χρήση τους μπορεί να είναι ιδιαιτέρως χρήσιμη σε περιπτώσεις όπου τα δεδομένα εκπαίδευσης είναι πολύ περισσότερα ή, ακόμα περισσότερο, σε περιπτώσεις όπου έχουν πολύ μεγαλύτερο αριθμό χαρακτηριστικών.

5.2 Προοπτικές Αξιοποίησης σε Παραγωγικό Περιβάλλον – Μελλοντική Εργασία

Βάσει της ανάλυσης που πραγματοποιήθηκε στα προηγούμενα κεφάλαια, και δεδομένου ότι κατασκευάστηκαν μοντέλα με ακρίβεια πρόβλεψης που πλησιάζει το 70%, μπορούμε να συμπεράνουμε με σχετική ασφάλεια πως υπάρχουν προοπτικές για την αξιοποίηση εργαλείων μηχανικής μάθησης για τη δημιουργία προβλέψεων στα πλαίσια της πλατφόρμας eRetail Audit Marketplace.

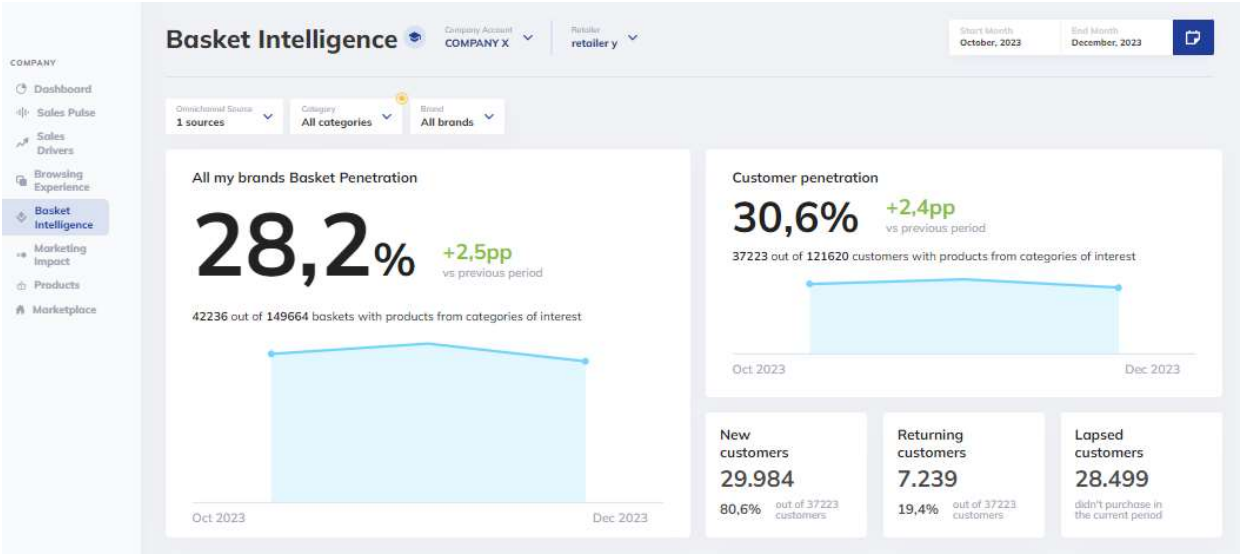
Η ενσωμάτωση αυτής της λογικής θα απαιτήσει, κατά πάσα πιθανότητα, για κάθε ηλεκτρονικό κατάστημα που διαθέτει hashed customer ids να κατασκευαστούν αντίστοιχα μοντέλα για κάθε brand που το κατάστημα πουλάει. Θα πρέπει επιπλέον αυτά τα μοντέλα να ανανεώνονται ανά τακτά χρονικά διαστήματα ώστε να ενσωματώνουν επιπλέον πληροφορίες από μελλοντικές συναλλαγές.

Πέραν αυτών θα είχε νόημα να επιχειρηθούν τα παρακάτω ώστε να καταλήξουμε, ενδεχομένως, σε μοντέλα με μεγαλύτερη ακόμα ακρίβεια πρόβλεψης:

- Πειραματισμός με περισσότερες παραμέτρους και περισσότερες τιμές για κάθε μία από αυτές, κατά την εκπαίδευση των μοντέλων. Εδώ υπήρχε περιορισμός λόγω

μηχανημάτων αλλά σε επιχειρησιακό επίπεδο υπάρχουν περισσότεροι πόροι ώστε κάτι τέτοιο να είναι δυνατό.

- Ενδεχόμενη συμπερίληψη όλων των πελατών στο μοντέλο προκειμένου να εντοπίσουμε πιθανούς νέους πελάτες ή πελάτες με 1-2 συναλλαγές επί του brand, αλλά με προοπτικές να γίνουν πιστοί πελάτες.
- Χρήση περισσότερων δεδομένων (με μεθόδους augmentation ή με τη χρήση δεδομένων για μεγαλύτερο χρονικό διάστημα.
- Απόκτηση και χρήση, με τις ανάλογες δικλείδες προστασίας προσωπικών δεδομένων, περισσότερων δεδομένων αναφορικά με τους πελάτες (πχ δημογραφικά, γεωγραφικά, δεδομένα πλοήγησης)



Εικόνα 30-Οθόνη Basket Intelligence, eRAM

Ουσιαστικά η προβλέψεις σχετικά με το customer churn θα μπορούσαν να ενσωματωθούν στην οθόνη "Basket Intelligence" που παρέχεται στους πελάτες της πλατφόρμας. Αυτή τη στιγμή περιέχει ένα κομμάτι customer report στο οποίο αναφέρεται, για την ορισμένη χρονική περίοδο, πόσοι από εκείνους που αγόρασαν τα brand του πελάτη-προμηθευτή είναι εντελώς καινούργιοι πελάτες (New Customers) και πόσοι έχουν αγοράσει ξανά (Returning Customers). Επιπλέον δείχνει πόσοι δεν αγόρασαν τη συγκεκριμένη χρονική περίοδο, παρόλο που έχουν αγοράσει στο παρελθόν (Lapsed Customers). Περιλαμβάνει, τέλος, και το Customer Penetration Report.

Η χρήση των μοντέλων πρόβλεψης μελλοντικής απώλειας πελατών μπορεί να οδηγήσει στην προσθήκη μίας ακόμα καρτέλας όπου θα αναφέρεται πόσοι από τους «πιστούς» πελάτες του brand κινδυνεύουν να αποχωρήσουν.

Εικόνα 31-Πρόταση Ενσωμάτωσης Μοντέλου στο Basket Intelligence

Σε μία μελλοντική έκδοση των μοντέλων το πρόβλημα θα μπορούσε να μετατραπεί από δυαδικό σε πρόβλημα 3 κλάσεων. Συγκεκριμένα θα είχε ίσως νόημα να churners να

χωριστούν σε 2 υποκατηγορίες, δηλαδή σε εκείνους που αποχωρούν συνολικά από το ηλεκτρονικό κατάστημα και σε αυτούς που παραμένουν σε αυτό αλλά στρέφονται σε ανταγωνιστικά brands. Η δεύτερη κατηγορία είναι πιθανά πιο σημαντική γιατί οι πελάτες αυτοί θα μπορούσαν με διάφορες ενέργειες να ανακτηθούν, ενώ αποχώρηση σε επίπεδο καταστήματος δεν σημαίνει απαραίτητα αποχώρηση από το brand παρά αγορά του από άλλο κατάστημα με ευνοϊκότερους όρους (τιμή, προσφορές, συγκέντρωση πόντων κτλ.).

Πηγαίνοντας, τέλος, ένα βήμα παραπέρα, θα μπορούσαμε ανάμεσα στους churners αυτής τη κατηγορίας να εντοπίζουμε αυτούς των οποίων η αποχώρηση είναι πιο επιζήμια, ενσωματώνοντας στα μοντέλα την έννοια του κέρδους όπως συναντάμε συχνά στη βιβλιογραφία.

Βιβλιογραφία

[1] T. T. Frederick F. Reichheld, *The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value*. Harvard Business School Press, 2001.

[2] A. De Caigny, K. Coussement, and K. W. De Bock, “A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees,” *Eur J Oper Res*, vol. 269, no. 2, pp. 760–772, Sep. 2018, doi: 10.1016/j.ejor.2018.02.009.

[3] S. H. Chen, “The gamma CUSUM chart method for online customer churn prediction,” *Electron Commer Res Appl*, vol. 17, pp. 99–111, May 2016, doi: 10.1016/j.elerap.2016.04.003.

[4] K. Coussement, D. F. Benoit, and D. Van den Poel, “Preventing Customers from Running Away! Exploring Generalized Additive Models for Customer Churn Prediction,” 2015, pp. 238–238. doi: 10.1007/978-3-319-10873-5_134.

[5] P. Berger and M. Kompan, “User Modeling for Churn Prediction in E-Commerce,” *IEEE Intell Syst*, vol. 34, no. 2, pp. 44–52, Mar. 2019, doi: 10.1109/MIS.2019.2895788.

[6] S. J, C. Gangadhar, R. K. Arora, P. N. Renjith, J. Bamini, and Y. devidas Chincholkar, “E-commerce customer churn prevention using machine learning-based business intelligence strategy,” *Measurement: Sensors*, vol. 27, Jun. 2023, doi: 10.1016/j.measen.2023.100728.

[7] Felix Langfeld, “Churn Prediction for News Media — predicting Churn with Google Analytics and Python.” Accessed: Sep. 16, 2023. [Online]. Available: https://medium.com/@felixlangfeld/churn-prediction-for-news-media-dc1b2fa3b49f

[8] F. Shirazi and M. Mohammadi, “A big data analytics model for customer churn prediction in the retiree segment,” *Int J Inf Manage*, vol. 48, pp. 238–253, Oct. 2019, doi: 10.1016/j.ijinfomgt.2018.10.005.

[9] F. Safinejad, E. A. Z. Noughabi, and B. H. Far, “A Fuzzy Dynamic Model for Customer Churn Prediction in Retail Banking Industry,” in *Applications of Data Management and Analysis*, 2018, pp. 85–101. doi: 10.1007/978-3-319-95810-1_7.

[10] M. Szmydt, “Predicting Customer Churn in Electronic Banking,” in *Business Information Systems Workshops,* Springer, 2019, pp. 687–696. doi: 10.1007/978-3-030-04849-5_58.

[11] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, “A comparison of machine learning techniques for customer churn prediction,” *Simul Model Pract Theory*, vol. 55, pp. 1–9, Jun. 2015, doi: 10.1016/j.simpat.2015.03.003.

[12] M. Usman, W. Ahmad, and A. Fong, “Design and Implementation of a System for Comparative Analysis of Learning Architectures for Churn Prediction,” *IEEE Communications Magazine*, vol. 59, no. 9, pp. 86–90, Sep. 2021, doi: 10.1109/MCOM.110.2100145.

[13] M. Li, C. Yan, W. Liu, and X. Liu, “An early warning model for customer churn prediction in telecommunication sector based on improved bat algorithm to optimize ELM,” *International Journal of Intelligent Systems*, vol. 36, no. 7, pp. 3401–3428, Jul. 2021, doi: 10.1002/int.22421.

[14] S. Höppner, E. Stripling, B. Baesens, S. vanden Broucke, and T. Verdonck, “Profit driven decision trees for churn prediction,” *Eur J Oper Res*, vol. 284, no. 3, pp. 920–933, Aug. 2020, doi: 10.1016/j.ejor.2018.11.072.

[15] E. Stripling, S. vanden Broucke, K. Antonio, B. Baesens, and M. Snoeck, “Profit maximizing logistic model for customer churn prediction using genetic algorithms,” *Swarm Evol Comput*, vol. 40, pp. 116–130, Jun. 2018, doi: 10.1016/j.swevo.2017.10.010.

[16] S. Tavassoli and H. Koosha, “Hybrid ensemble learning approaches to customer churn prediction,” *Kybernetes*, vol. 51, no. 3, pp. 1062–1088, Feb. 2022, doi: 10.1108/K-04-2020-0214.

[17] M. N. Haque, N. J. de Vries, and P. Moscato, “A Multi-objective Meta-Analytic Method for Customer Churn Prediction,” in *Business and Consumer Analytics: New Ideas*, Cham: Springer International Publishing, 2019, pp. 781–813. doi: 10.1007/978-3-030-06222-4_20.

[18] J. Vijaya, E. Sivasankar, and S. Gayathri, “Fuzzy Clustering with Ensemble Classification Techniques to Improve the Customer Churn Prediction in Telecommunication Sector,” in *Recent*

Developments in Machine Learning and Data Analytics, 2019, pp. 261–274. doi: 10.1007/978-981-13-1280-9_25.

[19] W. Bi, M. Cai, M. Liu, and G. Li, "A Big Data Clustering Algorithm for Mitigating the Risk of Customer Churn," *IEEE Trans Industr Inform*, vol. 12, no. 3, pp. 1270–1281, Jun. 2016, doi: 10.1109/TII.2016.2547584.

[20] M. U. Tariq, M. Babar, M. Poulin, and A. S. Khattak, "Distributed model for customer churn prediction using convolutional neural network," *Journal of Modelling in Management*, vol. 17, no. 3, pp. 853–863, Aug. 2022, doi: 10.1108/JM2-01-2021-0032.

[21] "Telco Customer Churn." Accessed: Nov. 19, 2023. [Online]. Available: https://www.kaggle.com/datasets/blastchar/telco-customer-churn

[22] A. Amin *et al.*, "Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods," *Int J Inf Manage*, vol. 46, pp. 304–319, Jun. 2019, doi: 10.1016/j.ijinfomgt.2018.08.015.

[23] I. V. Pustokhina *et al.*, "Dynamic customer churn prediction strategy for business intelligence using text analytics with evolutionary optimization algorithms," *Inf Process Manag*, vol. 58, no. 6, Nov. 2021, doi: 10.1016/j.ipm.2021.102706.

[24] W. Verbeke, D. Martens, and B. Baesens, "Social network analysis for customer churn prediction," *Applied Soft Computing Journal*, vol. 14, no. PART C, pp. 431–446, 2014, doi: 10.1016/j.asoc.2013.09.017.

[25] A. Barfar, B. Padmanabhan, and A. Hevner, "Applying behavioral economics in predictive analytics for B2B churn: Findings from service quality data," *Decis Support Syst*, vol. 101, pp. 115–127, Sep. 2017, doi: 10.1016/j.dss.2017.06.006.

[26] A. De Caigny, K. Coussement, W. Verbeke, K. Idbenjra, and M. Phan, "Uplift modeling and its implications for B2B customer churn prediction: A segmentation-based modeling approach," *Industrial Marketing Management*, vol. 99, pp. 28–39, Nov. 2021, doi: 10.1016/j.indmarman.2021.10.001.

[27] E. Zdravevski, P. Lameski, C. Apanowicz, and D. Ślęzak, "From Big Data to business analytics: The case study of churn prediction," *Applied Soft Computing Journal*, vol. 90, May 2020, doi: 10.1016/j.asoc.2020.106164.

[28] E. Lee, B. Kim, S. Kang, B. Kang, Y. Jang, and H. K. Kim, "Profit optimizing churn prediction for long-term loyal customers in online games," *IEEE Trans Games*, vol. 12, no. 1, pp. 41–53, Mar. 2020, doi: 10.1109/TG.2018.2871215.

[29] Y. Lai and J. Zeng, "Analysis of customer churn behavior in digital libraries," *Program*, vol. 48, no. 4, pp. 370–382, Aug. 2014, doi: 10.1108/PROG-08-2011-0035.

[30] T. A. Maier and S. Prusty, "Managing Customer Retention in Private Clubs Using Churn Analysis: Some Empirical Findings," *Journal of Hospitality Marketing & Management*, vol. 25, no. 7, pp. 797–819, Oct. 2016, doi: 10.1080/19368623.2016.1113904.

[31] K. Coussement and D. Van den Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," *Expert Syst Appl*, vol. 34, no. 1, pp. 313–327, Jan. 2008, doi: 10.1016/j.eswa.2006.09.038.

[32] S. Maldonado, G. Domínguez, D. Olaya, and W. Verbeke, "Profit-driven churn prediction for the mutual fund industry: A multisegment approach," *Omega (United Kingdom)*, vol. 100, Apr. 2021, doi: 10.1016/j.omega.2020.102380.

[33] D. R. Cox, "The Regression Analysis of Binary Sequences," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 215–232, Jul. 1958, doi: 10.1111/j.2517-6161.1958.tb00292.x.

[34] S. Menard, *Applied Logistic Regression Analysis*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc., 2002. doi: 10.4135/9781412983433.

[35] J. R. Quinlan, "Induction of decision trees," *Mach Learn*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[36] Nidhi, “Decision Trees: A Powerful Tool in Machine Learning.” Accessed: Feb. 17, 2024. [Online]. Available: https://medium.com/@nidhigh/decision-trees-a-powerful-tool-in-machine-learning-dd0724dad4b6

[37] C. Cortes and V. Vapnik, “Support-Vector Networks,” *Mach Learn*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1023/A:1022627411411.

[38] Vlahavas I, Kefalas P, Bassiliades N, Kokkoras F, and Sakellariou I, *Artificial Intelligence*, 4th ed. Giourdas, 2020.

[39] L. Breiman, “Random Forests,” *Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[40] Abhishek Sharma, “Random Forest vs Decision Tree | Which Is Right for You?” Accessed: Dec. 16, 2023. [Online]. Available: https://www.analyticsvidhya.com/ blog/2020/05/decision-tree-vs-random-forest-algorithm/

[41] Y. Freund and R. E. Schapire, “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting,” *J Comput Syst Sci*, vol. 55, no. 1, pp. 119–139, Aug. 1997, doi: 10.1006/jcss.1997.1504.

[42] J. H. Friedman, “Stochastic gradient boosting,” *Comput Stat Data Anal*, vol. 38, no. 4, pp. 367–378, Feb. 2002, doi: 10.1016/S0167-9473(01)00065-2.

[43] T. Chen and C. Guestrin, “XGBoost,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[44] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” *Adv Neural Inf Process Syst*, vol. 30, 2017, [Online]. Available: https://github.com/ Microsoft/ LightGBM.

[45] George Novack, “Building a One Hot Encoding Layer with TensorFlow.” Accessed: Jan. 14, 2024. [Online]. Available: https://towardsdatascience.com/building-a-one-hot-encoding-layer-with-tensorflow-f907d686bf39

[46] H. Abdi and L. J. Williams, “Principal component analysis,” *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[47] I. Jolliffe, “Principal Component Analysis,” in *Wiley StatsRef: Statistics Reference Online*, Wiley, 2014. doi: 10.1002/9781118445112.stat06472.

[48] Zakaria Jaadi, “A Step-by-Step Explanation of Principal Component Analysis (PCA).” Accessed: Dec. 21, 2023. [Online]. Available: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[49] Hanchuan Peng, Fuhui Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Trans Pattern Anal Mach Intell*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[50] R. Tibshirani, “Regression Shrinkage and Selection Via the Lasso,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, Jan. 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[51] I. Prapas, “Applied Dimensionality Reduction — 3 Techniques using Python.” Accessed: Dec. 26, 2023. [Online]. Available: https://www.learndatasci.com /tutorials/ applied-dimensionality-reduction-techniques-using-python/

[52] Mohd Zuhaib, “Demystifying the Confusion Matrix Using a Business Example.” Accessed: Dec. 20, 2023. [Online]. Available: https://towardsdatascience.com/ demystifying-confusion-matrix-29f3037b0cfa

[53] “Receiver operating characteristic.” Accessed: Dec. 21, 2023. [Online]. Available: https://en.wikipedia.org/wiki/Receiver_operating_characteristic

[54] “Convert Group.” Accessed: Nov. 11, 2023. [Online]. Available: https://convertgroup.com/ company/