ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ & ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Μεθοδολογία εκτίμησης πιθανότητας απώλειας πελατών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μάριος Α. Μητρόπουλος

Επιβλέπων : Βασίλειος Ασημακόπουλος
Καθηγητής Ε.Μ.Π.

Υπεύθυνος : Αρτέμιος-Ανάργυρος Σεμένογλου
Υποψήφιος Διδάκτωρ Ε.Μ.Π.

Αθήνα, Οκτώβριος 2022

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ & ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

Μεθοδολογία εκτίμησης πιθανότητας απώλειας πελατών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Μάριος Α. Μητρόπουλος

Επιβλέπων : Βασίλειος Ασημακόπουλος
Καθηγητής Ε.Μ.Π.

Υπεύθυνος : Αρτέμιος-Ανάργυρος Σεμένογλου
Υποψήφιος Διδάκτωρ Ε.Μ.Π.

Εγκρίθηκε από την τριμελή επιτροπή την 12/10/2022

.....	.....	.....
Βασίλειος Ασημακόπουλος Καθηγητής Ε.Μ.Π.	Δούκας Χρυσόστομος Αναπληρωτής Καθηγητής Ε.Μ.Π.	Δημήτριος Ασκούνης Καθηγητής Ε.Μ.Π.

Αθήνα, Οκτώβριος 2022

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΙΚΩΝ ΒΙΟΜΗΧΑΝΙΚΩΝ ΔΙΑΤΑΞΕΩΝ & ΣΥΣΤΗΜΑΤΩΝ ΑΠΟΦΑΣΕΩΝ

.....
Μάριος Μητρόπουλος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Ηλεκτρονικών Υπολογιστών

Περίληψη

Η παρούσα διπλωματική εργασία έχει σαν στόχο την ανάπτυξη μιας γενικότερης μεθοδολογίας για την εκτίμηση της πιθανότητας απώλειας ενός πελάτη μιας εταιρείας ή ενός οργανισμού, καθώς και την ερμηνεία και αξιολόγηση των αποτελεσμάτων αυτών. Το ποσοστό απωλειών πελατών αποτελεί έναν από τους πιο σημαντικούς δείκτες που μετράει τη δυνατότητα μιας επιχείρησης να διατηρεί το πελατολόγιό της. Συνεπώς, μια μέθοδος πρόβλεψης της πιθανότητας απώλειας κάποιου πελάτη αποτελεί ένα ιδιαίτερα σημαντικό εργαλείο, το οποίο μπορεί να χρησιμοποιηθεί για την ανανέωση των παρεχόμενων υπηρεσιών ή τη χάραξη στοχευμένων προωθητικών ενεργειών. Ιδιαίτερα τα τελευταία χρόνια που ο όγκος των διαθέσιμων δεδομένων έχει αυξηθεί, η χρήση εργαλείων μηχανικής μάθησης έχει προσφέρει νέες δυνατότητες.

Η βασική ιδέα της προτεινόμενης μεθοδολογίας είναι η αυτοματοποίηση της διαδικασίας παραγωγής προβλέψεων από σύνολα δεδομένων που αφορούν την απώλεια πελατών. Το προτεινόμενο πλαίσιο δέχεται ως είσοδο ένα σύνολο δεδομένων και παράγει ως έξοδο τις προβλέψεις για τα εισαγόμενα δεδομένα, καθώς και γραφικές παραστάσεις και μετρικές για την αξιολόγησή τους. Η μεθοδολογία αυτή περιλαμβάνει πέντε βήματα. Το πρώτο περιλαμβάνει την προεπεξεργασία των δεδομένων εισόδου για την καλύτερη εκπαίδευση των μοντέλων πρόβλεψης, η οποία πραγματοποιείται συγχρόνως με τη βελτιστοποίηση των υπερπαραμέτρων των μοντέλων στο δεύτερο βήμα. Στο τρίτο στάδιο, δοκιμάζονται οι συνδυασμοί των μοντέλων που αναπτύχθηκαν και επιλέγεται ως βέλτιστο το μοντέλο με τα καλύτερα αποτελέσματα. Στο τέταρτο βήμα, στο καλύτερο μοντέλο του προηγούμενου βήματος δοκιμάζονται διάφορα κατώφλια προβλέψεων με στόχο την επιλογή του βέλτιστου. Τέλος, στο πέμπτο αξιολογείται το συνολικό μοντέλο που προέκυψε με χρήση των παραγόμενων γραφικών παραστάσεων και μετρικών.

Για την ανάπτυξη της μεθοδολογίας αυτής διεξάχθηκε μια εκτενής πειραματική διαδικασία, εξετάζοντας κάθε φορά διαφορετικές παραμέτρους του συστήματος. Για τα πειράματα αυτά χρησιμοποιήθηκε ένα σύνολο δεδομένων παρεχόμενο από την εταιρεία KKBox Inc., που περιλαμβάνει τα δεδομένα των χρηστών της υπηρεσίας που προσφέρει η εταιρεία, μια επιγραμμική υπηρεσία αναπαραγωγής μουσικής. Στο πλαίσιο της παρούσας διπλωματικής εργασίας περιγράφεται αναλυτικά αυτή η πειραματική διαδικασία, καθώς και τα συμπεράσματα στα οποία καταλήξαμε μέσα από την ανάλυση.

Λέξεις κλειδιά: Μηχανική μάθηση, Απώλεια πελατών, Πρόβλεψη πιθανοτήτων

Abstract

The aim of this diploma thesis is the development of a general methodology for estimating the customer churn probability of a company or an organization, as well as the interpretation and evaluation of the results. The customer churn rate is one the most important indicators that measures the ability of a company to retain its customer base. Therefore, a method of predicting the probability of churn of a customer is a particularly important tool, which can be used for the renewal of the supplied services or the engraving of targeted promotional actions. Especially in recent years where the volume of available data has increased, the use of machine learning tools has offered new possibilities.

The fundamental idea of the proposed methodology is the automation of the process of generating predictions from data sets relating to customer churn. The proposed framework receives as input a data set and outputs the predictions for the input data, as well as graphs and metrics used for their evaluation. This methodology consists of five steps. The first step contains the preprocessing of the input data for the better training of the prediction models, which is carried out in the second step along with the hyperparameter optimization of the models. In the third stage, the combinations of the developed models are tested, of which the one with the best results is selected as optimal. In the fourth step, multiple prediction thresholds are tested using the best model of the previous step, and the optimal one is selected. Finally, in the fifth stage the resulting complete model is evaluated using the generated graphs and metrics.

For the development of this methodology an extensive experimental procedure was conducted, examining different parameters of the system each time. For these experiments a data set supplied from KKBox Inc. was used, containing the user data of the service provided by the company, an online music streaming service. As part of this diploma thesis, this experimental procedure, as well as the conclusions stemming from this analysis, is described thoroughly.

Keywords: Machine learning, Customer churn, Probability prediction

Ευχαριστίες

Η διπλωματική αυτή εργασία εκπονήθηκε στα πλαίσια των ερευνητικών δραστηριοτήτων της Μονάδας Προβλέψεων και Στρατηγικής κατά το ακαδημαϊκό έτος 2021 – 2022. Η μονάδα υπάγεται στον Τομέα Βιομηχανικών Διατάξεων και Συστημάτων Αποφάσεων της Σχολής Ηλεκτρολόγων Μηχανικών & Μηχανικών Η/Υ, του Εθνικού Μετσόβιου Πολυτεχνείου.

Αρχικά, θα ήθελα να ευχαριστήσω τον Καθηγητή κ. Βασίλειο Ασημακόπουλο για την ευκαιρία που μου έδωσε να ασχοληθώ σε βάθος με το πρόβλημα ανάπτυξης μεθοδολογίας εκτίμησης πιθανότητας απώλειας πελατών με χρήση μηχανικής μάθησης. Θα ήθελα, ακόμα, να ευχαριστήσω τον Καθηγητή κ. Δούκα Χρυσόστομο και τον Καθηγητή κ. Δημήτριο Ασκούνη για τη συμμετοχή τους στην επιτροπή εξέτασης της εργασίας.

Επιπρόσθετα, θα ήθελα να ευχαριστήσω θερμά τον Υποψήφιο Διδάκτορα κ. Αρτέμιο-Ανάργυρο Σεμένογλου για τη συνεχή του βοήθεια και καθοδήγηση στην εκπόνηση αυτής της εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου για την υποστήριξή τους καθ’ όλη τη διάρκεια των σπουδών μου.

Μάριος Μητρόπουλος,

Αθήνα, Οκτώβριος 2022

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κεφάλαιο 1. Εισαγωγή

1.1 Αντικείμενο της εργασίας

Οι σύγχρονες επιχειρήσεις αντιμετωπίζουν συνεχώς την πρόκληση της εγκαθίδρυσής τους στη διαρκώς εναλλασσόμενη παγκόσμια αγορά. Για να την επιτύχουν, απαιτείται η ανάπτυξη και διατήρηση μιας πελατειακής βάσης. Παραδοσιακά, οι περισσότερες επιχειρήσεις επενδύουν περισσότερα χρήματα για την απόκτηση νέων πελατών, θεωρώντας πως είναι ένας γρήγορος τρόπος αύξησης του τζίρου. Η απομάκρυνση ενός πελάτη της εταιρείας συνεπάγεται την απώλεια των μελλοντικών συναλλαγών και αγορών που θα απέφεραν κέρδος στην εταιρεία, καθώς και τη μερική υποτίμηση της αρχικής επένδυσης που κατατέθηκε για την απόκτησή του. Ταυτόχρονα, η επένδυση για τη διατήρηση ενός ήδη υπάρχοντος πελάτη είναι έως και πέντε φορές μικρότερη κατά μέσο όρο. Έτσι, η διατήρηση της πελατειακής βάσης μιας εταιρείας αποτελεί αδιαμφισβήτητα σημαντικό στοιχείο της ανάπτυξης και εξέλιξης της.

Για να επιτύχει αυτή τη διατήρηση, η εταιρεία καλείται να χρησιμοποιήσει εργαλεία ανάλυσης και πρόβλεψης απώλειας πελατών. Συνεπώς, η ανάλυση αυτή είναι κρίσιμη για την επιτυχία μιας επιχείρησης σε μια παγκόσμια αγορά που επηρεάζεται σημαντικά από τη ραγδαία ανάπτυξη της τεχνολογίας και τον ανταγωνισμό. Μια μέθοδος πρόβλεψης της πιθανότητας απώλειας κάποιου πελάτη αποτελεί ένα ιδιαίτερα σημαντικό εργαλείο, το οποίο μπορεί να χρησιμοποιηθεί για την ανανέωση των παρεχόμενων υπηρεσιών ή τη χάραξη στοχευμένων προωθητικών ενεργειών. Έτσι, δημιουργούνται νέα κίνητρα παραμονής στην υπηρεσία για τους πελάτες, οι οποίοι είναι λιγότερο πιθανό να κάνουν churn. Παράλληλα, μέσα από την ανάλυση της απώλειας πελατών, μια εταιρεία μπορεί να κατανοήσει τους λόγους για τους οποίους οι πελάτες οδηγούνται στη διακοπή χρήσης της υπηρεσίας που παρέχει. Με τον τρόπο αυτό, η εταιρεία μπορεί να βελτιώσει την υπηρεσία και τα προϊόντα που προσφέρει, κερδίζοντας ένα πλεονέκτημα απέναντι στον ανταγωνισμό και ικανοποιώντας τους συγκεκριμένους, αλλά και νέους πελάτες, εκ των προτέρων.

Η διαδικασία ανάλυσης και πρόβλεψης της απώλειας πελατών βασίζεται στη συλλογή επαρκών δεδομένων που αφορούν τους χρήστες της προσφερόμενης υπηρεσίας. Τα δεδομένα αυτά περιλαμβάνουν δημογραφικά δεδομένα, δεδομένα συναλλαγών μεταξύ χρήστη και εταιρείας, και δεδομένα καταγραφής δραστηριότητας και χρήσης της υπηρεσίας. Γενικά, χρησιμοποιούνταν μέθοδοι όπως η γραμμική παλινδρόμηση ή η χρήση Μπεϋζιανής θεωρίας για τη διεξαγωγή της ανάλυσης και πρόβλεψης. Ωστόσο, ο συνεχώς αυξανόμενος όγκος των διαθέσιμων δεδομένων σε συνδυασμό με τις διαρκώς διογκούμενες απαιτήσεις των καταναλωτών, που προκύπτουν ως παραπροϊόν του έντονου ανταγωνισμού της αγοράς, υπαγορεύουν τη χρήση αποτελεσματικότερων μεθόδων για την ταχύτερη και πιο αποδοτική εκτέλεση του έργου που πρέπει να επιτελεστεί. Η τεχνολογική ανάπτυξη των τελευταίων χρόνων επιτρέπει πλέον τη χρήση της μηχανικής μάθησης σε εφαρμογές όπως αναγνώριση προτύπων, υπολογισμός συναρτήσεων, βελτιστοποίηση, αυτόματος έλεγχος, καθώς και

σε προβλήματα ταξινόμησης και πρόβλεψης, που είναι και ο τύπος προβλήματος που περιγράφουμε.

Η μηχανική μάθηση είναι το πεδίο της επιστημονικής έρευνας στο οποίο μελετώνται και σχεδιάζονται συστήματα και μέθοδοι που «μαθαίνουν», δηλαδή που αξιοποιούν νέα δεδομένα για να βελτιώσουν την απόδοσή τους στη συγκεκριμένη εργασία. Το πεδίο αυτό αποτελεί κλάδο της τεχνητής νοημοσύνης. Οι αλγόριθμοι που χρησιμοποιούνται δημιουργούν ένα μοντέλο, το οποίο εκπαιδεύεται με ένα σύνολο δεδομένων που έχει προκύψει από παρατηρήσεις και καταγραφές, για την παραγωγή προβλέψεων ή/και αποφάσεων χωρίς να έχουν προγραμματιστεί ρητά για την κάθε εργασία. Στο πλαίσιο της παρούσας διπλωματικής εργασίας, χρησιμοποιήθηκαν κυρίως αλγόριθμοι βασισμένοι στα δέντρα αποφάσεων, δηλαδή ένα μοντέλο με σχήμα δέντρου που κάθε κόμβος αντιστοιχεί σε ένα χαρακτηριστικό των δεδομένων και η ταξινόμηση πραγματοποιείται ανάλογα με την τιμή που λαμβάνει το εισαγόμενο παράδειγμα στο εκάστοτε χαρακτηριστικό.

Στόχος της εργασίας είναι η ανάπτυξη μιας γενικότερης μεθοδολογίας για την εκτίμηση της πιθανότητας απώλειας ενός πελάτη μιας εταιρείας ή ενός οργανισμού, καθώς και η ερμηνεία και αξιολόγηση των αποτελεσμάτων αυτών. Η προτεινόμενη μεθοδολογία λαμβάνει ως είσοδο ένα σύνολο δεδομένων που αφορά το churn, το οποίο δέχεται μια καθορισμένη προεπεξεργασία πριν την εισαγωγή των δεδομένων στα μοντέλα πρόβλεψης. Τα μοντέλα αυτά εκπαιδεύονται χρησιμοποιώντας τα δεδομένα και, ύστερα, επιλέγεται ο συνδυασμός αυτών που παράγει τα καλύτερα αποτελέσματα, καθώς και το βέλτιστο κατώφλι πρόβλεψης. Τέλος, η μεθοδολογία παράγει ως έξοδο τις προβλέψεις για τα εισαγόμενα δεδομένα, καθώς και γραφικές παραστάσεις και μετρικές για την αξιολόγησή τους.

Η προτεινόμενη μεθοδολογία προέκυψε μέσα από τη διεξαγωγή μιας εκτενούς πειραματικής διαδικασίας, τα συμπεράσματα της οποίας χρησιμοποιήθηκαν για την ανάπτυξη του πλαισίου. Περιλαμβάνει μια εκτενή προεπεξεργασία με χρήση στατιστικής ανάλυσης για την εξέταση των χαρακτηριστικών του συνόλου δεδομένων που χρησιμοποιήθηκε, τα δεδομένα χρήσης της επιγραμμικής υπηρεσίας αναπαραγωγής μουσικής που παρέχει η εταιρεία KKBox Inc., καθώς και πολλαπλά πειράματα που αφορούν τη βελτιστοποίηση των εισαγόμενων δεδομένων και τη διαδικασία εκπαίδευσης των μοντέλων πρόβλεψης.

1.2 Οργάνωση της εργασίας

Στο δεύτερο κεφάλαιο της παρούσας διπλωματικής εργασίας γίνεται μια εισαγωγή στην απώλεια πελατών. Αρχικά, δίνεται ο ορισμός του churn γενικότερα και αναλύεται ο λόγος για τον οποίο απασχολεί τις σύγχρονες επιχειρήσεις. Στη συνέχεια, γίνεται μια εισαγωγή στην πρόβλεψη του churn και περιγράφεται ο τρόπος με τον οποίο εντάσσεται στο ευρύτερο πλαίσιο της ανάλυσης της συμπεριφοράς των πελατών και η αξία της. Ύστερα, παρουσιάζονται διαφορετικές μέθοδοι για την πρόβλεψη της

απώλειας πελατών, καθώς και μερικές μετρικές για την αξιολόγηση των προβλέψεων αυτών.

Το τρίτο κεφάλαιο αποτελεί εισαγωγή στη μηχανική μάθηση και τα νευρωνικά δίκτυα. Γίνεται, αρχικά, μια ανασκόπηση των κατηγοριών της μηχανικής μάθησης και, έπειτα, παρουσιάζεται το μοντέλο του τεχνητού νευρώνα στο οποίο βασίζονται οι διάφορες αρχιτεκτονικές των νευρωνικών δικτύων. Τέλος, επεξηγείται η λειτουργία των μοντέλων πρόβλεψης που χρησιμοποιούνται στην παρούσα εργασία, δηλαδή των ταξινομητών Logistic Regression και Gaussian Naive Bayes, καθώς και μοντέλων που βασίζονται στα δέντρα αποφάσεων, των οποίων η δομή και η μέθοδος αναλύεται εκτενώς.

Στο τέταρτο κεφάλαιο παρουσιάζεται αναλυτικά η προτεινόμενη μεθοδολογία αυτοματοποίησης της διαδικασίας παραγωγής προβλέψεων για την απώλεια πελατών μιας εταιρείας ή ενός οργανισμού. Η μεθοδολογία περιλαμβάνει την προεπεξεργασία των δεδομένων εισόδου, την ανάπτυξη των μοντέλων πρόβλεψης, την επιλογή του συνδυασμού των μοντέλων και του κατωφλίου πρόβλεψης που δίνουν τα βέλτιστα αποτελέσματα, και την εξαγωγή των προβλέψεων της πιθανότητας churn για το εισαγόμενο σύνολο δεδομένων, καθώς και γραφικών παραστάσεων και μετρικών για την αξιολόγησή τους.

Το πέμπτο κεφάλαιο είναι αφιερωμένο στην αναλυτική περιγραφή της πειραματικής διαδικασίας που ακολουθήθηκε για την ανάπτυξη της προτεινόμενης μεθοδολογίας. Αρχικά, παρουσιάζεται το σύνολο δεδομένων που χρησιμοποιήθηκε και η προεπεξεργασία που εκτελέστηκε σε αυτό για τη διεξαγωγή των διαφόρων πειραμάτων. Έπειτα, επισημαίνονται οι δείκτες με τους οποίους αξιολογούνται τα αποτελέσματα των πειραμάτων που παρουσιάζονται στη συνέχεια του κεφαλαίου και αφορούν διαφορετικές παραμέτρους του συστήματος. Τέλος, αναδεικνύονται τα αποτελέσματα των πειραμάτων που εκτελέστηκαν στην προτεινόμενη μεθοδολογία χρησιμοποιώντας το αρχικό σύνολο δεδομένων, καθώς και ενός επιπλέον.

Στο έκτο κεφάλαιο εξάγονται τα συμπεράσματα και διερευνώνται τρόποι επέκτασης της διεξαχθείσας μελέτης.

Κεφάλαιο 2. Απώλεια Πελατών (Customer Churn)

2.1 Εισαγωγή

Η συνεχόμενη, ραγδαία ανάπτυξη της τεχνολογίας και του ανταγωνισμού μεταξύ εταιρειών έχει επιφέρει σημαντικές αλλαγές στην παγκόσμια αγορά. Παραδοσιακά, οι περισσότερες επιχειρήσεις επενδύουν περισσότερα χρήματα για την απόκτηση νέων πελατών, θεωρώντας πως είναι ένας γρήγορος τρόπος αύξησης του τζίρου. Ωστόσο, ένα από τα σημαντικότερα στοιχεία για την εξέλιξη και την ανάπτυξη μιας επιχείρησης είναι η διατήρηση των ήδη υπαρχόντων πελατών της, καθώς οι πελάτες που αποδεικνύονται πιο κερδοφόροι για μια επιχείρηση δεν είναι εκείνοι που εκτελούν μερικές συναλλαγές, αλλά εκείνοι με τους οποίους χτίζονται δυνατές σχέσεις που οδηγούν σε επαναλαμβανόμενες αγορές. Συνεπώς, με το καταναλωτικό κοινό να αναζητά συνεχώς καλύτερες προσφορές για τις υπηρεσίες και τα προϊόντα που αγοράζει, η κατανόηση των λόγων απώλειας πελατών και η πρόβλεψή της αποτελούν κύρια ζητήματα για την ορθή λειτουργία μιας επιχείρησης στη σύγχρονη εποχή [1, 38, 39].

2.2 Ορισμός του churn

Με τον όρο churn περιγράφουμε το φαινόμενο κατά το οποίο ένας καταναλωτής σταματά να χρησιμοποιεί την υπηρεσία που παρέχεται από μια εταιρεία [2]. Ενώ στο παρελθόν ήταν αποτέλεσμα της λύσης της αντίστοιχης σύμβασης, στις σύγχρονες λιανικές χρηματοπιστωτικές υπηρεσίες μέσω διαδικτύου οι παρατεταμένες περίοδοι αδράνειας των καταναλωτών αποτελούν customer churn. Έτσι, μπορούμε να διαχωρίσουμε το churn σε δύο κατηγορίες: συμβατικό (contractual) και μη συμβατικό (non-contractual).

Το contractual churn αναφέρεται στην περίπτωση κατά την οποία ο πελάτης δεν παρατείνει τη ισχύουσα σύμβαση για την αντίστοιχη υπηρεσία που χρησιμοποιεί. Πιθανοί λόγοι αποτελούν η πλέον μη κάλυψη των αναγκών του καταναλωτή ή η αλλαγή σε αντίστοιχη υπηρεσία που παρέχεται από ανταγωνιστή. Αυτό το είδος churn παρατηρείται σε υπηρεσίες όπως τραπεζικοί λογαριασμοί, κινητή τηλεφωνία, και συνδρομητικών διαδικτυακών υπηρεσιών αναπαραγωγής μουσικής ή ταινιών.

Σε μια σχέση μεταξύ εταιρείας και καταναλωτή που δεν περιλαμβάνει τη σύναψη κάποιας σύμβασης, ο πελάτης μπορεί να σταματήσει να χρησιμοποιεί την υπηρεσία χωρίς κάποιο χρονικό περιορισμό. Έτσι, σε αυτό το non-contractual περιβάλλον, για να οριστεί το churn των πελατών πρέπει πρώτα να οριστεί κάποιο χρονικό κριτήριο, βάσει του οποίου θα μπορεί να μετρηθεί το churn. Έτσι, όταν αυτή η περίοδος αδράνειας που έχει οριστεί ως threshold ξεπεραστεί, ο πελάτης θεωρείται churned. Με τον τρόπο αυτό, μπορεί να υπολογιστεί η πιθανότητα του customer churn εντός της συγκεκριμένης χρονικής περιόδου. Παράδειγμα για αυτό το είδους churn αποτελεί η αγορά ενός προϊόντος ή η παραγγελία φαγητού μέσω διαδικτύου, αφού συνήθως σε αντίστοιχες

υπηρεσίες δε διαγράφονται οι λογαριασμοί των χρηστών, αλλά οι ίδιοι μπορεί να τις χρησιμοποιούν περιστασιακά και αραιά [3].

Αξίζει να σημειωθεί ότι ο όρος churn χρησιμοποιείται και για την περιγραφή της αποχώρησης υπαλλήλων από μια εταιρεία (λόγω εργασιακού περιβάλλοντος ή μιας καλύτερης προσφοράς από άλλη εταιρεία), φαινόμενο που είναι γνωστό ως employee churn.

2.3 Ορισμός της πρόβλεψης απώλειας πελατών

Η πρόβλεψη απώλειας πελατών (customer churn prediction) είναι η διαδικασία κατά την οποία, χρησιμοποιώντας τα διαθέσιμα δεδομένα και κάποια μεθοδολογία, γίνεται η πρόβλεψη βάσει πιθανοτήτων για το ποιοι πελάτες θα αποχωρήσουν στο άμεσο μέλλον. Η τεχνική αυτή αποτελεί σημαντικό κομμάτι εργασίας των data analytics στις σύγχρονες Customer Relationship Management ομάδες των εταιρειών [4, 40].

Στον τομέα των τηλεπικοινωνιών λέγεται ότι το κόστος απόκτησης ενός νέου πελάτη είναι 5 ή περισσότερες φορές μεγαλύτερο από εκείνο που απαιτείται για τη διατήρηση ενός πελάτη. Είναι, συνεπώς, φανερό ότι η ανάλυση του churn των πελατών μιας εταιρείας είναι κρίσιμη για την κατάστρωση αποτελεσματικών στρατηγικών μάρκετινγκ που έχουν ως στόχο την ανάπτυξη της εταιρείας αυτής. Μέρος της ανάλυσης αυτής είναι ο καθορισμός της αξίας ενός πελάτη, καθώς πολλές φορές μπορεί να μην αποφέρει κέρδος η διατήρηση ενός πελάτη που δε χρησιμοποιεί ενεργά την εκάστοτε υπηρεσία. Έτσι, ορίζουμε το Customer Lifetime Value ως το συνολικό εισόδημα της εταιρείας από τον πελάτη καθ' όλη τη διάρκεια συνεργασίας. Ο υπολογισμός της μετρικής αυτής είναι απλός όταν η πελατειακή σχέση έχει λήξει. Ωστόσο, για τη μεγιστοποίηση του κέρδους είναι σημαντικός ο ορισμός της αξίας αυτής κατά τη διάρκεια, ή ακόμα και πριν, το ενεργό στάδιο της πελατειακής σχέσης. Για τον υπολογισμό του Customer Lifetime Value χρησιμοποιούνται τα εξής στοιχεία [5, 41]:

- Η τρέχουσα αξία του πελάτη μέχρι στιγμής (έσοδα και έξοδα)
- Το ποσοστό παρακράτησης ή η διάρκεια χρήσης της υπηρεσίας
- Ο συντελεστής προεξόφλησης

Η ανάλυση του churn έχει ως κύριους στόχους την πρόβλεψη του πιθανού churn ενός πελάτη, και την κατανόηση των λόγων για τους οποίους ένας πελάτης θα αποφάσιζε να κάνει churn. Με την πρόβλεψη των πελατών που είναι πιο πιθανό να τερματίσουν τη χρήση της υπηρεσίας, μια εταιρεία μπορεί να μειώσει το ποσοστό churn παρέχοντας νέα κίνητρα στους πελάτες αυτούς, όπως για παράδειγμα κάποια ειδική έκπτωση. Με την κατανόηση των λόγων του customer churn μια εταιρεία μπορεί να βελτιώσει την υπηρεσία που παρέχει, ικανοποιώντας τους αντίστοιχους αλλά και νέους πελάτες εκ των προτέρων [2].

Συμπερασματικά, το churn αποτελεί ένα αναπόφευκτο πρόβλημα για κάθε εταιρεία στη σημερινή αγορά, όπου οι καταναλωτές μπορούν να επιλέξουν ανάμεσα από

πολλά ωφέλιμα και ανταγωνιστικά πακέτα υπηρεσιών και η εναλλαγή μεταξύ παρόχων είναι πιο εύκολη από ποτέ. Επίσης, πελάτες μπορεί να επηρεαστούν από το churn άλλων πελατών της εταιρείας και να την οδηγήσουν σε περεταίρω churn. Επομένως, η ανάλυση και η πρόβλεψη της απώλειας πελατών κρίνεται αναγκαία για μια εταιρεία στη σύγχρονη εποχή, αφού καλείται να αντιμετωπίσει το βέβαιο πρόβλημα του churn και να δημιουργήσει σχέσεις εμπιστοσύνης μεταξύ εκείνης και των πελατών της.

2.4 Μέθοδοι πρόβλεψης απώλειας πελατών

Για τη διεξαγωγή του churn analysis, και κατ' επέκταση της πρόβλεψης απώλειας πελατών, απαιτείται η χρήση τεχνικών data mining, η οποία είναι μέθοδος εντοπισμού και μοντελοποίησης των σχέσεων μεγάλου όγκου δεδομένων. Η διαδικασία του data mining αποτελείται από τέσσερα βήματα [2, 42]:

1. Εντοπισμός και συλλογή δεδομένων
2. Προεπεξεργασία των δεδομένων
3. Ανάπτυξη μοντέλου πρόβλεψης
4. Ερμηνεία και σχεδιασμός ενεργειών (υποδείγματα, τάσεις)

Τα δεδομένα που συλλέγονται για την πραγματοποίηση αυτής της ανάλυσης περιλαμβάνουν:

- Δημογραφικά δεδομένα (π.χ. ηλικία, φύλο, χώρα/πόλη)
- Ιστορικό συναλλαγών του χρήστη με την εταιρεία (π.χ. ποσό, τρόπος πληρωμής)
- Αρχεία καταγραφής της δραστηριότητας του χρήστη της υπηρεσίας (π.χ. διάρκεια/αριθμός κλήσεων αν πρόκειται για τηλεφωνική εταιρεία)
- Το πλάνο που είναι εγγεγραμμένος ο πελάτης (παρόν και παρελθόν)
- Αρχεία καταγραφής πιθανών παραπόνων του χρήστη

Για την πρόβλεψη του churn, μερικές από τις πιο δημοφιλείς μεθόδους που χρησιμοποιούνται παρουσιάζονται παρακάτω.

Regression Analysis

Το μοντέλο δυαδικής λογιστικής παλινδρόμησης είναι ένα είδους regression model που χρησιμοποιείται στην περίπτωση που η εξαρτημένη μεταβλητή, η οποία στην περίπτωσή μας είναι το churn, δεν είναι συνεχής, αλλά περιγράφει μια κατάσταση που θα πραγματοποιηθεί ή όχι. Έτσι, χρησιμοποιείται για να προβλέψει ένα διακριτό αποτέλεσμα, δηλαδή 0 ή 1 που αντιστοιχεί στη συγκεκριμένη περίπτωση στο ότι θα παραμείνει πελάτης ή όχι, βάσει συνεχών ή/και κατηγορικών μεταβλητών. Το τυπικό linear regression model είναι το εξής:

$$P(X) = \alpha + \beta X, \text{όπου } X = (x_1, x_2, \dots, x_n).$$

Για τον περιορισμό του αποτελέσματος στις τιμές 0 και 1 που επιθυμούμε, πρέπει να αναθέσουμε περίπλοκες ιδιότητες στο θόρυβο ε. Για να περιορίσουμε την πιθανότητα εντός του διαστήματος $[0, 1]$ και να διαφοροποιείται μονοτονικά βάσει του X, απαιτείται η χρήση συναρτήσεων πέρα των γραμμικών. Μια συνάρτηση που πληροί τις προϋποθέσεις αυτές είναι η logistic function:

$$P(X) = \frac{e^{-(\alpha+\beta X)}}{1 + e^{-(\alpha+\beta X)}}$$

και άρα,

$$Q(X) = 1 - P(X) = \frac{1}{1+e^{-(\alpha+\beta X)}}.$$

Στο τυπικό linear regression model με n εισόδους, η έξοδος προκύπτει ως εξής:

$$P(X) = \alpha + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

και ο αντίστοιχος τύπος στο logistic regression μοντέλο είναι:

$$Q(X) = 1 - P(X) = \frac{1}{1+e^{-(\alpha+b_1 x_1+b_2 x_2+\cdots+b_n x_n)}}.$$

Η συνεχής έξοδος που προκύπτει από τους παραπάνω τύπους διαχωρίζεται στις δύο κατηγορίες που επιθυμούμε, churners και μη, χρησιμοποιώντας κάποιο threshold, συνήθως το 0.5 που είναι το μέσο του διαστήματος [5].

K-means Clustering

Με τη μέθοδο αυτή τα δεδομένα ομαδοποιούνται βάσει την απόστασή τους από τα κέντρα των δύο κατηγοριών. Για τον υπολογισμό των αποστάσεων χρησιμοποιείται κάποια ευριστική συνάρτηση (ευκλείδεια απόσταση ή Manhattan). Ο αλγόριθμος ξεκινά επιλέγοντας k σημεία από τα δεδομένα. Το k είναι ο αριθμός των clusters που θα δημιουργήσει ο αλγόριθμος. Έτσι, στη συγκεκριμένη περίπτωση το k είναι 2. Βάσει την απόστασή τους από τα δύο αρχικά σημεία, κατηγοριοποιούνται όλα τα υπόλοιπα. Ύστερα, ορίζουμε ως νέο κέντρο το μέσο όρο των σημείων του κάθε cluster και κατηγοριοποιούμε ξανά τα σημεία βάσει απόστασης. Η διαδικασία αυτή συνεχίζεται μέχρι η τελευταία εκτέλεση του αλγορίθμου να μην αλλάξει την κατηγοριοποίηση κάποιου στοιχείου, κι άρα να μην έχουμε αλλαγή του κέντρου των clusters [6].

Naive Bayes

Το μοντέλο Naive Bayes είναι βασισμένο το θεώρημα του Bayes για τις υπό συνθήκη πιθανότητες [49]:

$$p(C_k|x) = \frac{p(C_k)\, p(x|C_k)}{p(x)},$$

όπου C_k οι κατηγορίες και x ανεξάρτητη μεταβλητή εισόδου. Για n μεταβλητές εισόδου έχουμε:

$$p(C_k, x_1, x_2, \dots, x_n) = p(x_1|x_2, \dots, x_n, C_k)\, p(x_2|x_3, \dots, x_n, C_k) \dots p(x_{n-1}|x_n, C_k)\, p(x_n|C_k)\, p(C_k).$$

Η naive υπόθεση του μοντέλου είναι ότι οι ανεξάρτητες μεταβλητές είναι και αμοιβαία ανεξάρτητες. Έτσι, έχουμε:

$$p(x_i|x_{i+1}, \dots, x_n, C_k) = p(x_i|C_k)$$

και τελικά προκύπτει:

$$p(C_k|x_1, x_2, \dots, x_n) = \frac{1}{Z} p(C_k) \prod_{i=1}^{n} p(x_i|C_k),$$

όπου $Z = p(x) = \sum_k p(C_k)\, p(x|C_k)$ είναι παράγοντας κλιμάκωσης που εξαρτάται μόνο από τα $x_1, x_2, \dots, x_n$ και είναι σταθερά αν οι τιμές τους είναι γνωστές [6, 49].

Νευρωνικά Δίκτυα και Μηχανική Μάθηση

Η πιο διαδεδομένη μέθοδος πρόβλεψης του churn είναι η χρήση κάποιου μοντέλου μηχανικής μάθησης. Από τις πιο συνηθισμένες επιλογές είναι τα νευρωνικά δίκτυα, δηλαδή ένα σύνολο κόμβων συνδεδεμένα με ακμές κατάλληλα για να προσομοιάσουν τη λειτουργία των συνάψεων στο μυαλό των ανθρώπων. Κάθε κόμβος καλείται νευρώνας και, βάσει το άθροισμα των εισόδων του συμπεριλαμβανομένου ενός βάρους για να θέσουμε το κατάλληλο threshold κάθε φορά, προκύπτει η έξοδος του νευρώνα. Αυτή μπορεί να χρησιμοποιηθεί στη συνέχεια ως είσοδο για επόμενο επίπεδο νευρώνων.

Επόμενη ευρέως χρησιμοποιημένη τεχνική μηχανικής μάθησης για την πρόβλεψη απώλειας πελατών είναι τα δέντρα απόφασης. Οι κόμβοι του δέντρου δημιουργούνται βάσει τις τιμές και τη σημασία των χαρακτηριστικών των δεδομένων εισόδου. Ακολουθώντας τη ροή του δέντρου ανάλογα με τις τιμές στα αντίστοιχα χαρακτηριστικά καταλήγουμε σε κάποιο φύλλο του, καθένα από τα οποία αποτελούν τις κλάσεις που ταξινομούνται τα δεδομένα.

Τέλος, άλλη τεχνική που χρησιμοποιείται είναι τα Support Vector Machines. Η τεχνική αυτή βασίζεται στη θεωρία της στατιστικής μάθησης και είναι ικανή να διαχωρίσει δεδομένα σε δύο κλάσεις, στην περίπτωσή μας churners και πελάτες που θα παραμείνουν, μέσα από την αντιστοίχιση σημείων σε σημεία του χώρου και δημιουργώντας μια ευθεία διαχωρισμού με τη μέγιστη δυνατή απόσταση μεταξύ των δύο κλάσεων σημείων [6, 43].

Τα Νευρωνικά Δίκτυα και η Μηχανική Μάθηση, καθώς και τα διάφορα μοντέλα που χρησιμοποιούνται, αναλύονται εκτενώς στο επόμενο κεφάλαιο της εργασίας.

2.5 Μετρικές για την αξιολόγηση των προβλέψεων

Για την κατανόηση και την αξιολόγηση των αποτελεσμάτων των μοντέλων που χρησιμοποιούνται για την πρόβλεψη απώλειας πελατών απαιτείται ο ορισμός μετρικών. Οι μετρικές αυτές βασίζονται στα true/false positive/negative, τα οποία ορίζονται ως εξής [44]:

- True Positive (TP): το μοντέλο προβλέπει ορθώς ότι ο χρήστης είναι churner
- True Negative (TN): το μοντέλο προβλέπει ορθώς ότι ο χρήστης δεν είναι churner
- False Positive (FP): το μοντέλο προβλέπει λανθασμένα ότι ο χρήστης είναι churner
- False Negative (FN): το μοντέλο προβλέπει λανθασμένα ότι ο χρήστης δεν είναι churner

Έτσι, οι μετρικές που χρησιμοποιούνται περισσότερο για την αξιολόγηση των μοντέλων παρουσιάζονται παρακάτω [44]:

Confusion Matrix

Το confusion matrix παρέχει εποπτική αξιολόγηση του μοντέλου καθώς αναγράφει τα true/false positive/negative. Συνεπώς, δείχνει τη γενική επίδοση του μοντέλου [7].

	Κλάση 1 (predicted)	Κλάση 0 (predicted)
Κλάση 1 (actual)	True Positive	False Negative
Κλάση 0 (actual)	False Positive	True Negative

Πίνακας 1: 2x2 Confusion matrix

Accuracy

Το accuracy είναι μέτρο του πόσο κοντά ή μακριά είναι οι προβλέψεις από τις πραγματικές τιμές τους. Ορίζεται ως:

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

Αποτελεί καλή επιλογή για τη γενική επισκόπηση των αποτελεσμάτων, ωστόσο σε περιπτώσεις ανισορροπημένων δεδομένων μπορεί να προκύψει υψηλή τιμή του accuracy που να μην αντικατοπτρίζει τη συνολική εικόνα [7].

Precision

Το precision είναι το ποσοστό των συνολικών προβλέψεων της κλάσης 1 (churners) που ήταν ορθές. Δηλαδή, ορίζεται ως [7]:

$$Precision = \frac{TP}{TP + FP}.$$

Recall (Sensitivity)

Το recall είναι το ποσοστό των πραγματικών στιγμιότυπων της κλάσης 1 (churners) που προέβλεψε το μοντέλο. Δηλαδή, ορίζεται ως:

$$Recall = \frac{TP}{TP + FN}.$$

Η μετρική αυτή δίνει την πιθανότητα το μοντέλο να αναγνωρίσει ορθώς έναν churner, που αποτελεί το ζητούμενο του μοντέλου κι άρα η μετρική αυτή είναι σημαντική. Ωστόσο, δε μπορεί να αποτελέσει τη μόνη μετρική για την ερμηνεία των αποτελεσμάτων, καθώς η πρόβλεψη όλων των χρηστών ως churners πετυχαίνει 100% recall [7].

Specificity

Αντίστοιχα με το recall για την κλάση 0, το specificity είναι το ποσοστό των πραγματικών στιγμιότυπων της κλάσης 0 (non-churners) που προέβλεψε το μοντέλο. Δηλαδή, ορίζεται ως:

$$Recall = \frac{TN}{TN + FP}.$$

Η μετρική αυτή δίνει την πιθανότητα το μοντέλο να αναγνωρίσει ορθώς έναν non-churner.

F1-score

Το F1-score είναι ο αρμονικός μέσος μεταξύ precision και recall. Έτσι, ορίζεται ως:

$$F1 = \frac{2 \times Recall \times Precision}{Recall + Precision} = \frac{2TP}{2TP + FP + FN}.$$

Ο συνδυασμός των δύο παραπάνω μετρικών δίνει μια καλή γενική εικόνα των αποτελεσμάτων του μοντέλου. Γενικεύεται σε F_β-score για την πρόσθεση επιπλέον παράγοντα βάρους στο precision ή στο recall:

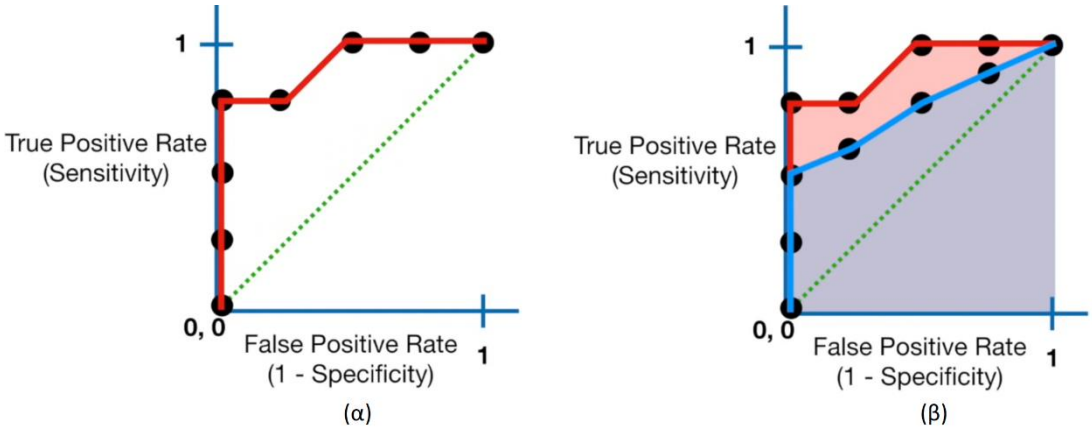
$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} = \frac{(1 + \beta^2) \times TP}{(1 + \beta^2) \times TP + \beta^2 \times FN + FP}$$

ROC curve, AUC, και Lift curve

Τα μοντέλα πρόβλεψης έχουν ως έξοδο την πιθανότητα του κάθε στιγμιότυπου να είναι churner ή όχι. Έτσι, ένα threshold απαιτείται για να μετατραπεί η πιθανότητα

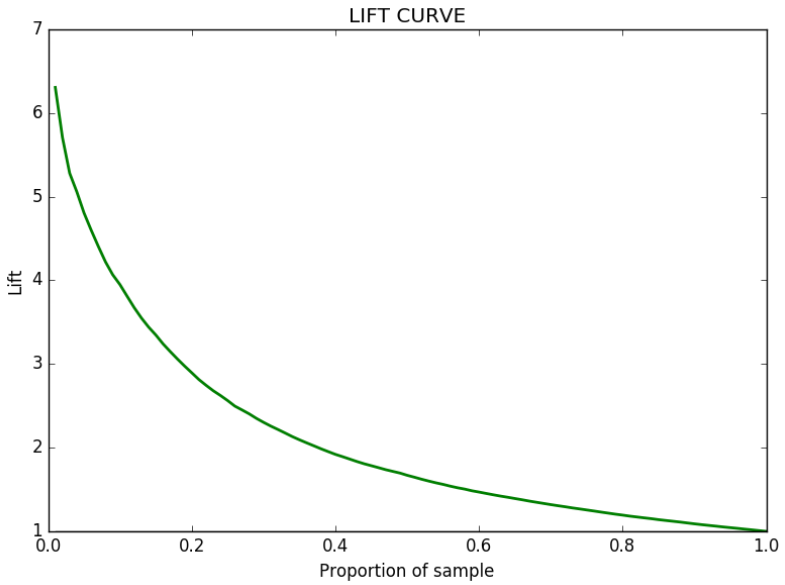
σε 0 ή 1 και να προκύψει το τελικό αποτέλεσμα. Το ROC (Receiver Operating Characteristic) γράφημα ουσιαστικά συνοψίζει όλα τα πιθανά confusion matrices που προκύπτουν από κάθε διαφορετικό threshold. Ο άξονας y είναι το recall του κάθε πίνακα, ενώ ο άξονας x ορίζεται ως το 1 – sensitivity. Με το γράφημα αυτό μπορούμε να αποφασίσουμε την τιμή του threshold που δίνει τα καλύτερα αποτελέσματα για το συγκεκριμένο μοντέλο.

Το AUC (Area Under the receiver operating characteristic Curve) είναι το εμβαδόν μεταξύ ενός ROC γραφήματος και του άξονα των x. Αποτελεί μετρική για τη σύγκριση μεταξύ μοντέλων, με το μεγαλύτερο εμβαδόν να υποδεικνύει καλύτερα αποτελέσματα. Δηλαδή, στο παρακάτω παράδειγμα το κόκκινο μοντέλο θεωρείται καλύτερο από το μπλε.



Σχήμα 1: Παράδειγμα (α) ROC curve (β) AUC

Το Lift είναι η μετρική επίδοσης που προκύπτει από την αντίστοιχη τιμή του ROC curve διαιρώντας με το baseline (πράσινη διακεκομμένη γραμμή στο Σχήμα 1). Το γράφημα δίνει το lift (άξονας y) σε σχέση με το δείγμα των συνολικών πελατών κάθε φορά (άξονας x) [5, 7].



Σχήμα 2: Παράδειγμα Lift curve, από [8]

Κεφάλαιο 3. Νευρωνικά Δίκτυα και Μηχανική Μάθηση

3.1 Εισαγωγή

Η μηχανική μάθηση είναι το πεδίο της επιστημονικής έρευνας στο οποίο μελετώνται και σχεδιάζονται συστήματα και μέθοδοι που «μαθαίνουν», δηλαδή που αξιοποιούν νέα δεδομένα για να βελτιώσουν την απόδοσή τους στη συγκεκριμένη εργασία. Το πεδίο αυτό αποτελεί κλάδο της τεχνητής νοημοσύνης. Οι αλγόριθμοι που χρησιμοποιούνται δημιουργούν ένα μοντέλο, το οποίο εκπαιδεύεται με ένα σύνολο δεδομένων που έχει προκύψει από παρατηρήσεις και καταγραφές, για την παραγωγή προβλέψεων ή/και αποφάσεων χωρίς να έχουν προγραμματιστεί ρητά για την κάθε εργασία. Για απλά προβλήματα που αναθέτονται στις μηχανές, είναι δυνατό ο προγραμματισμός του αλγορίθμου να περιλαμβάνει όλα τα βήματα για τη λύση του προβλήματος, ενώ για πιο πολύπλοκες εργασίες ίσως είναι ευκολότερο για τον άνθρωπο να βοηθήσει τον υπολογιστή να αναπτύξει το δικό του αλγόριθμο πάρα να το δημιουργήσει ο ίδιος. Έτσι, οι αλγόριθμοι μηχανικής μάθησης χρησιμοποιούνται σε μεγάλο εύρος εφαρμογών, όπου είναι δύσκολη ή ανέφικτη η ανάπτυξη συμβατικών αλγορίθμων για την εκτέλεση των εργασιών αυτών [9].

Από τις πιο διαδεδομένες μεθόδους στον τομέα της μηχανικής μάθησης είναι τα νευρωνικά δίκτυα, τα οποία εφαρμόζουν μηχανική μάθηση για να χρησιμοποιήσουν δεδομένα με τρόπο που προσομοιώνει τη λειτουργία του ανθρώπινου εγκεφάλου. Οι πρωτεργάτες στην υλοποίηση των πρώτων τεχνητών νευρωνικών δικτύων ήταν οι Warren S. McCulloch και Walter Pitts στις αρχές της δεκαετίας του 1940, οι οποίοι δημιούργησαν ένα υπολογιστικό μοντέλο για νευρωνικά δίκτυα [10]. Ύστερα από σχετικές επεκτάσεις και θεωρητικής ανάπτυξης από μεταγενέστερους επιστήμονες που ασχολήθηκαν με το συγκεκριμένο αντικείμενο, τα πρώτα λειτουργικά δίκτυα με πολλαπλά επίπεδα δημοσιεύθηκαν από τους Alexey G. Ivakhnenko και Valentin G. Lapa το 1965 [11]. Ωστόσο, η έκρηξη στην ανάπτυξη των νευρωνικών δικτύων και της μηχανικής μάθησης ήρθε με την ανάπτυξη της τεχνολογίας των τελευταίων χρόνων, η οποία προσέφερε λύσεις στα κύρια εμπόδια της έρευνας του αντικειμένου, δηλαδή την έλλειψη επαρκών υπολογιστικών πόρων και ανάπτυξη τεχνικών για την επίλυση προβλήματα μεγάλης κλίμακας.

Στις μέρες μας, τα νευρωνικά δίκτυα και η μηχανική μάθηση βρίσκονται στο προσκήνιο της τεχνολογικής ανάπτυξης. Οι εφαρμογές των νευρωνικών δικτύων και μηχανικής μάθησης που χρησιμοποιούνται έχουν προκύψει τα τελευταία λίγα χρόνια, όπως αναγνώριση προτύπων, υπολογισμός συναρτήσεων, βελτιστοποίηση, πρόβλεψη, αυτόματος έλεγχος, και όλο και περισσότερες από αυτές δημοσιεύονται ως έτοιμα προϊόντα στην αγορά και χρησιμοποιούνται ευρέως. Οι εφαρμογές αυτές αντιστοιχίζονται σε αμέτρητους κλάδους της καθημερινότητάς μας, από ένα απλό πρόβλημα ταξινόμησης μιας κατηγορίας ατόμων έως και την κατασκευή αυτοοδηγουμένων οχημάτων [12]. Είναι βέβαιο ότι τα επόμενα χρόνια ένας συνεχώς αυξανόμενος αριθμός θα ακολουθήσει.

3.2 Κατηγορίες μηχανικής μάθησης

Από την εποχή που εφευρέθηκαν οι ηλεκτρονικοί υπολογιστές, οι άνθρωποι αναρωτιόντουσαν εάν θα μπορούσαν να κατασκευάσουν υπολογιστές με τη δυνατότητα να μαθαίνουν. Μια επιτυχημένη κατανόηση του πώς να κάνουμε τους υπολογιστές να εκπαιδεύονται θα εξασφάλιζε νέα επίπεδα ευχέρειας και εξατομίκευσης. Πλέον, ακριβώς όπως υπάρχουν διαφορετικοί τρόποι με τους οποίους μαθαίνουν οι άνθρωποι από το περιβάλλον τους, το ίδιο ισχύει και για τα σύγχρονα νευρωνικά δίκτυα [13]. Οι αλγόριθμοι που έχουν αναπτυχθεί επιτρέπουν την αποτελεσματική αντιμετώπιση προβλημάτων σε διάφορους τομείς. Κύριο παράδειγμα αποτελεί το data mining για την εξαγωγή πληροφοριών από τεράστιες βάσεις δεδομένων νοσοκομειακών, τραπεζικών, και άλλων εγκαταστάσεων. Είναι φυσικό να υπάρχει πιθανότητα ανθρώπινου σφάλματος κατά τη διάρκεια της ανάλυσης, οπότε με την επιτυχημένη χρήση μηχανικής μάθησης να αυξηθεί δραματικά η απόδοση του συστήματος. Τα χαρακτηριστικά (features) σε αυτές τις βάσεις δεδομένων, και κατ' επέκταση σε κάθε dataset που χρησιμοποιείται από αλγορίθμους μηχανικής μάθησης, είναι τριών τύπων: συνεχή (continuous) που αναφέρονται σε νούμερα, κατηγορικά (categorical) που αναφέρονται σε δεδομένα συμβολοσειρών, και δυαδικά (binary) που αναφέρεται σε δεδομένα που παίρνουν τιμή 0 ή 1 [14].

Οι τύποι προβλημάτων που λύνουν οι αλγόριθμοι μηχανικής μάθησης είναι οι παρακάτω [45]:

- Παλινδρόμηση (regression), που αναφέρεται σε προβλήματα που απαιτούν την πρόβλεψη μιας συνεχής αριθμητικής τιμής, για παράδειγμα πρόβλεψη τιμής ενός σπιτιού.
- Ταξινόμηση (classification), που αναφέρεται σε προβλήματα που απαιτούν την ταξινόμηση των δεδομένων σε διαφορετικές κλάσεις. Αν οι κλάσεις είναι δύο, το πρόβλημα αποκαλείται δυαδική ταξινόμηση (binary classification), ενώ για περισσότερες πολυωνυμική ταξινόμηση (multi-nomial classification). Παράδειγμα αποτελεί και το αντικείμενο της διπλωματικής εργασίας που θα αναλυθεί στη συνέχεια, η ταξινόμηση των πελατών μιας εταιρείας ως churners ή μη.
- Συσταδοποίηση (clustering), όπου απαιτείται η εύρεση μιας δομής ή ενός μοτίβου σε μια συλλογή μη κατηγοριοποιημένων δεδομένων, για παράδειγμα ομαδοποίηση παρόμοιων ταινιών ή τραγουδιών για μια υπηρεσία streaming.
- Αναγνώριση ανωμαλιών (anomaly detection), στο οποίο ο αλγόριθμος καλείται να αναγνωρίσει outliers εντός ενός συνόλου δεδομένων, δηλαδή τιμές που δε φαίνονται λογικές και αποκλίνουν από τις υπόλοιπες. Ένα παράδειγμα αυτού του τύπου προβλήματος είναι η ανίχνευση ύποπτων συναλλαγών για απάτη σε πιστωτικές κάρτες.
- Μείωση διαστατικότητας (dimensionality reduction). Ο αριθμός των μεταβλητών εισόδου ή χαρακτηριστικών ενός συνόλου δεδομένων καλείται διαστατικότητα των δεδομένων αυτών, και, άρα, το πρόβλημα της μείωσης διαστατικότητας αναφέρεται σε τεχνικές που μειώνουν τον αριθμό των μεταβλητών εισόδου.

Οι διαδικασίες μάθησης κατηγοριοποιούνται βάσει ύπαρξης ή μη εκπαιδευτή κατά τη διάρκεια της μάθησης σε μάθηση με εκπαιδευτή, ή αλλιώς επιβλεπόμενη μάθηση, και μάθηση χωρίς εκπαιδευτή, αντίστοιχα.

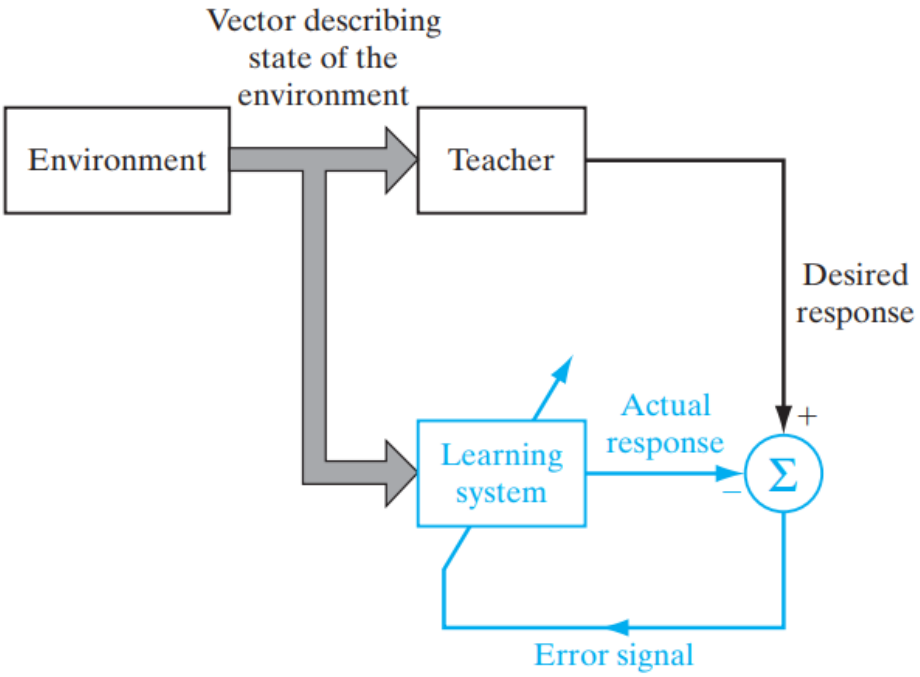
3.2.1 Μάθηση με εκπαιδευτή

Η μάθηση με εκπαιδευτή αναφέρεται και ως επιβλεπόμενη μάθηση. Αυτό το είδος μάθησης των νευρωνικών δικτύων αποτελεί κύριο θέμα της ερευνητικής δραστηριότητας στο πεδίο της μηχανικής μάθησης και πολλές τεχνικές επιβλεπόμενης μάθησης εφαρμόζονται στην επεξεργασία περιεχομένου πολυμεσικών συστημάτων. Χαρακτηριστικό γνώρισμα της μάθησης με εκπαιδευτή είναι η διαθεσιμότητα «labeled» δεδομένων μάθησης, δηλαδή δεδομένων για τα οποία υπάρχει κι η επιθυμητή απόκριση του συστήματος. Συνήθως, αυτά τα labels είναι labels κλάσεων σε προβλήματα ταξινόμησης. Οι αλγόριθμοι επιβλεπόμενης μάθησης διαμορφώνουν μοντέλα χρησιμοποιώντας αυτά τα δεδομένα εκπαίδευσης, τα οποία μοντέλα χρησιμοποιούνται για την ταξινόμηση άλλων δεδομένων που δεν είναι labeled. Γενικά, η μάθηση με εκπαιδευτή αποτελεί τη σημαντικότερη μεθοδολογία της μηχανικής μάθησης [15, 46].

Η επιβλεπόμενη μάθηση έχει ως στόχο την απόκτηση της σχέσης πληροφορίας μεταξύ δεδομένων εισόδου και εξόδου ενός συστήματος, βάσει παραδειγμάτων εκπαίδευσης στα οποία η κάθε είσοδος είναι αντιστοιχισμένη στην κατάλληλη έξοδο. Έτσι, θα δημιουργηθεί ένα τεχνητό σύστημα που έχει τη δυνατότητα να μάθει την αντιστοίχιση εισόδου – εξόδου, και να προβλέψει την έξοδο του συστήματος όταν του δίνονται νέες είσοδοι δεδομένων χωρίς label. Αν η έξοδος αυτή λαμβάνει διακριτές τιμές από ένα πεπερασμένο σύνολο τιμών, οι οποίες υποδεικνύουν τις κλάσεις του προβλήματος, η διαμορφωμένη αντιστοίχιση οδηγεί σε ταξινόμηση (classification) των δεδομένων εισόδου. Αυτό το είδος μάθησης χρησιμοποιήθηκε για τη διεκπεραίωση της παρούσας εργασίας, όπως θα εξηγηθεί και στα επόμενα κεφάλαια. Από την άλλη μεριά, αν η έξοδος λαμβάνει συνεχείς τιμές, η αντιστοίχιση οδηγεί σε παλινδρόμηση (regression) των δεδομένων εισόδου [16].

Όπως φαίνεται στο παρακάτω σχήμα, θεωρούμε ότι ο εκπαιδευτής (teacher) έχει γνώση του περιβάλλοντος (environment) και αυτή η γνώση αντιπροσωπεύεται από ένα σύνολο παραδειγμάτων εισόδου με labeled εξόδους, όπως αναφέρθηκε προηγουμένως. Ωστόσο, το περιβάλλον είναι άγνωστο στο νευρωνικό δίκτυο. Με την είσοδο ενός διανύσματος εκπαίδευσης, ο εκπαιδευτής παρέχει στο νευρωνικό δίκτυο την επιθυμητή απόκριση (desired response) για το συγκεκριμένο διάνυσμα. Η επιθυμητή απόκριση αντιπροσωπεύει τη βέλτιστη ενέργεια που πρέπει να εκτελεστεί από το νευρωνικό δίκτυο. Οι παράμετροι του δικτύου, δηλαδή τα συναπτικά του βάρη, προσαρμόζονται υπό τη συνδυασμένη επιρροή του διανύσματος εκπαίδευσης και του σήματος σφάλματος (error signal). Το σήμα σφάλματος ορίζεται ως η διαφορά μεταξύ της επιθυμητής απόκρισης και της πραγματικής απόκρισης του δικτύου. Η προσαρμογή που περιεγράφηκε εκτελείται με επαναληπτικό τρόπο, όπως υποδεικνύεται από το βρόγχο στο παρακάτω διάγραμμα. Έτσι, το νευρωνικό δίκτυο θα βρεθεί τελικά σε μια κατάσταση που θα προσομοιώνει τη συμπεριφορά του εκπαιδευτή. Χρησιμοποιώντας

τις κατάλληλες στατιστικές μετρικές για το εκάστοτε πρόβλημα, η προσομοίωση αυτή μπορεί να κριθεί.



Σχήμα 3: Σχηματικό διάγραμμα της μάθηση με εκπαιδευτή, από [17]

Με τη μέθοδο που επισημαίνεται στο παραπάνω διάγραμμα, η γνώση του περιβάλλοντος που είναι διαθέσιμη στον εκπαιδευτή μεταφέρεται στο νευρωνικό δίκτυο μέσω εκπαίδευσης και αποθηκεύεται με τη μορφή σταθερών συναπτικών βαρών, τα οποία αντιπροσωπεύουν μακροπρόθεσμη μνήμη. Όταν επιτευχθεί αυτή η συνθήκη, μπορούμε να απαλλαγούμε από τον εκπαιδευτή και να αφήσουμε το νευρωνικό δίκτυο να αντιμετωπίσει το περιβάλλον εντελώς μόνο του, δηλαδή να παραγάγει προβλέψεις για δεδομένα εισόδου για τα οποία δε γνωρίζει την επιθυμητή απόκριση [17].

Αξίζει να σημειωθεί ότι η ελαχιστοποίηση του σφάλματος κατά την εκπαίδευση δεν εξασφαλίζει απαραίτητα καλή απόδοση κατά το testing, δηλαδή την κατάσταση όπου το σύστημα έχει εκπαιδευτεί και αξιολογείται βάσει των προβλέψεών του σε άγνωστα δεδομένα, ήτοι δεδομένα που δε χρησιμοποιήθηκαν κατά τη διάρκεια της εκπαίδευσης. Ο κύριος λόγος για την παρουσίαση του φαινομένου αυτού είναι πιθανό «overfitting» του συστήματος στα δεδομένα εκπαίδευσης, δηλαδή αχρείαστη πολυπλοκότητα του μαθητευόμενου συστήματος στην εκπαίδευση κατά τον καθορισμό της αντιστοίχισης μεταξύ εισόδου και εξόδου. Το πρόβλημα αναφέρεται ως πρόβλημα γενίκευσης (generalizability). Ένα καλός αλγόριθμος εκμάθησης πρέπει να έχει καλή γενίκευση. Για να τη λάβει υπόψη του κατά το σχεδιασμό του μαθητευόμενου συστήματος, ένας αλγόριθμος εκμάθησης πρέπει να ισορροπήσει μεταξύ ελαχιστοποίησης του σφάλματος εκπαίδευσης και πολυπλοκότητας της δομής του για τη βέλτιστη απόδοση [18].

Υπάρχουν διαφορετικές προσεγγίσεις για το σχεδιασμό ενός συστήματος εκπαίδευσης στην επιβλεπόμενη μάθηση. Μερικές από τις πιο γνωστές προσεγγίσεις

είναι οι εξής: logic-based, Multilayer Perceptron, statistical-learning, instance-based, Support Vector Machines, και Boosting.

3.2.2 Μάθηση χωρίς εκπαιδευτή

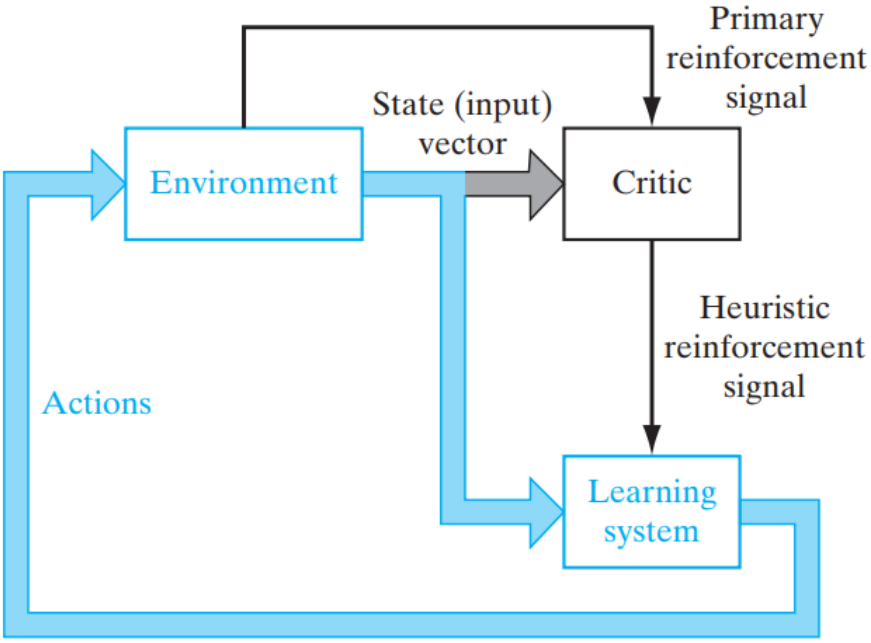
Στην επιβλεπόμενη μάθηση, η διαδικασία μάθησης λαμβάνει χώρα υπό την καθοδήγηση ενός εκπαιδευτή. Ωστόσο, στο παράδειγμα που είναι γνωστό ως μάθηση χωρίς εκπαιδευτή, όπως υποδηλώνει το όνομά του, δεν υπάρχει εκπαιδευτής που να επιβλέπει τη διαδικασία μάθησης. Δηλαδή, δεν υπάρχουν χαρακτηρισμένα παραδείγματα της λειτουργίας που πρέπει να μάθει το νευρωνικό δίκτυο. Στα πλαίσια αυτού του δεύτερου παραδείγματος, μπορούν να οριστούν δύο υποκατηγορίες μάθησης, που εξετάζονται στη συνέχεια: ενισχυτική μάθηση και μη επιβλεπόμενη μάθηση [17].

3.2.2.1 Ενισχυτική μάθηση

Η ενισχυτική μάθηση αντιμετωπίζει την ερώτηση του πώς ένας αυτόνομος πράκτορας (agent) ο οποίος αισθάνεται και ενεργεί εντός ενός περιβάλλοντος μπορεί να μάθει να επιλέγει βέλτιστες ενέργειες για να πετύχει τους στόχους του. Αυτό το πολύ γενικό πρόβλημα περιλαμβάνει εργασίες όπως εκμάθηση ελέγχου ενός κινητού ρομπότ, εκμάθηση βελτιστοποίησης λειτουργιών ενός εργοστασίου, και εκμάθηση του πώς να παίζει βέλτιστα διάφορα επιτραπέζια παιχνίδια. Κάθε φορά που ο πράκτορας εκτελεί μια ενέργεια μέσα στο περιβάλλον, του παρέχεται μια επιβράβευση (reward) ή μια ποινή (penalty) για να του υποδείξει επιθυμητότητα της συνεπαγόμενης κατάστασης. Για παράδειγμα, όταν ένας πράκτορας εκπαιδεύεται για να παίξει κάποιο παιχνίδι, μπορεί να του παρέχεται επιβράβευση όταν κερδίζει το παιχνίδι, ποινή όταν χάνει, και μηδενική επιβράβευση σε όλες τις υπόλοιπες καταστάσεις. Ο στόχος του πράκτορα είναι να μάθει από αυτή την έμμεση, καθυστερημένη επιβράβευση και να επιλέγει αλληλουχίες ενεργειών που παράγουν την καλύτερη αθροιστική επιβράβευση. Οι αλγόριθμοι ενισχυτικής μάθησης σχετίζονται με τους αλγορίθμους που χρησιμοποιούνται στο δυναμικό προγραμματισμό για τη λύση προβλημάτων βελτιστοποίησης [13].

Στην ενισχυτική μάθηση, η εκμάθησης μιας αντιστοίχισης εισόδου – εξόδου εκτελείται μέσω συνεχούς αλληλεπίδρασης με το περιβάλλον, με στόχο την ελαχιστοποίηση ενός βαθμωτού δείκτη απόδοσης. Στο παρακάτω σχήμα παρουσιάζεται το σχηματικό διάγραμμα μιας μορφής ενός συστήματος ενισχυτικής μάθησης που βασίζεται σε ένα μηχανισμό που λειτουργεί ως κριτής (critic), ο οποίος μετατρέπει ένα κύριο σήμα ενίσχυσης (primary reinforcement signal) λαμβανόμενο από το περιβάλλον σε ένα υψηλότερης ποιότητας σήμα ενίσχυσης που αποκαλείται ευρετικό σήμα ενίσχυσης (heuristic reinforcement signal), αμφότερα εκ των οποίων είναι βαθμωτές είσοδοι. Το σύστημα σχεδιάζεται ώστε να μαθαίνει βάσει καθυστερούμενης ενίσχυσης, γεγονός το οποίο σημαίνει ότι το σύστημα παρατηρεί μια χρονική ακολουθία ερεθισμάτων που λαμβάνει από το περιβάλλον, τα οποία καταλήγουν στην παραγωγή του ευρετικού σήματος ενίσχυσης. Ο στόχος της ενισχυτικής μάθησης είναι να

ελαχιστοποιεί μια συνάρτηση τρέχοντος κόστους, η οποία ορίζεται ως πρόβλεψη του αθροιστικού κόστους ενεργειών που εκτελούνται σε μια αλληλουχία βημάτων αντί απλώς του άμεσου κόστους μιας ενέργειας. Αντίστοιχα, θα μπορούσε ο στόχος της εκμάθησης να είναι η μεγιστοποίηση μιας συνάρτησης επιβράβευσης. Ενδεχομένως ορισμένες από τις ενέργειες που έχουν εκτελεστεί σε αυτήν την αλληλουχία χρονικών βημάτων να είναι οι καλύτερες ορίζουσες της συνολικής συμπεριφοράς του συστήματος. Η λειτουργία του συστήματος μάθησης είναι να ανακαλύψει αυτές τις ενέργειες και να τις τροφοδοτήσει πίσω στο περιβάλλον [17].



Σχήμα 4: Σχηματικό διάγραμμα της μάθηση με εκπαιδευτή, από [17]

Κύρια συστατικά ενός συστήματος ενισχυτικής μάθησης αποτελούν τα παρακάτω [19]:

- Η πολιτική (policy) είναι ο τρόπος με τον οποίο ο πράκτορας που αντιλαμβάνεται και ενεργεί εντός ενός περιβάλλοντος θα συμπεριφερθεί υπό κάποιες συγκεκριμένες συνθήκες. Δηλαδή, η πολιτική αντιστοιχίζει καταστάσεις σε ενέργειες. Μπορεί να είναι ένας πίνακας αναζήτησης, μια συνάρτηση, ή να περιλαμβάνει διαδικασία αναζήτησης. Η εύρεση της βέλτιστης πολιτικής είναι ο κύριος στόχος της ενισχυτικής μάθησης.
- Το σήμα επιβράβευσης (reward signal) υποδεικνύει πόσο καλή ή κακή είναι μια ενέργεια και ορίζει το στόχο του προβλήματος όπου ο σκοπός του πράκτορα είναι η μεγιστοποίηση της συνολικής λαμβανόμενης επιβράβευσης. Κατά συνέπεια, η επιβράβευση είναι ο κύριος παράγοντας που ενημερώνει την πολιτική. Η επιβράβευση μπορεί να είναι άμεση ή καθυστερημένη, και για τις καθυστερημένες επιβραβεύσεις ο πράκτορας πρέπει να καθορίσει ποιες ενέργειες είναι πιο σχετικές με αυτές.
- Η συνάρτηση αξίας (value function) είναι μια πρόβλεψη των συνολικών μελλοντικών επιβραβεύσεων, και χρησιμοποιείται για την εκτίμηση καταστάσεων και την επιλογή μεταξύ ενεργειών αντίστοιχα.

Στην ενισχυτική μάθηση, ο πράκτορας επηρεάζει την κατανομή των παραδειγμάτων εκπαίδευσης με τη σειρά ενεργειών που επιλέγει κάθε φορά. Αφού η εξερεύνηση είναι εγγενώς ακριβή όσον αφορά τους πόρους και το χρόνο, μια φυσική και κρίσιμη ερώτηση στην ενισχυτική μάθηση είναι η αντιμετώπιση της διχοτόμησης μεταξύ της εξερεύνησης (exploration) αγνώστων εδαφών και την εκμετάλλευση (exploitation) της ήδη υπάρχουσας γνώσης, δηλαδή ποια στρατηγική πειραματισμού παράγει την πιο αποτελεσματική εκπαίδευση. Πιο συγκεκριμένα, ο πράκτορας πρέπει να ισορροπήσει μεταξύ της άπληστης εκμετάλλευσης αυτών που έχει μάθει έως τώρα για την επιλογή ενεργειών που αποδίδουν υψηλότερες επιβραβεύσεις στο σύντομο μέλλον για τη μεγιστοποίηση της συνολικής επιβράβευσης, και τη συνεχή εξερεύνηση του περιβάλλοντος για την απόκτηση περισσότερων πληροφοριών από άγνωστες μέχρι στιγμής καταστάσεις και ενέργειες για να επιτύχει ενδεχομένως μακροπρόθεσμα οφέλη. Εκτενείς μελέτες έχουν διεξαχθεί για την εύρεση στρατηγικών που δίνουν το βέλτιστο trade-off μεταξύ εκμετάλλευσης και εξερεύνησης [13, 20].

Αξιοσημείωτο φαινόμενο της ενισχυτικής μάθησης είναι ότι συχνά απαιτείται από ένα ρομπότ να μάθει πολλές εργασίες εντός του ίδιου περιβάλλοντος, χρησιμοποιώντας τους ίδιους αισθητήρες. Για παράδειγμα, ένα κινητό ρομπότ μπορεί να χρειαστεί να μάθει πώς να «προσδέσει» στο φορτιστή της μπαταρίας του, πώς να πλοηγηθεί μέσα από στενούς διαδρόμους, και πώς να λάβει σήμα εξόδου από εκτυπωτές laser. Αυτή η σύνθεση δημιουργεί τη δυνατότητα χρήσης εμπειρίας και γνώσης που έχει αποκτήσει στο παρελθόν για τη μείωση της πολυπλοκότητας εκμάθησης των νέων εργασιών εντός του ίδιου περιβάλλοντος.

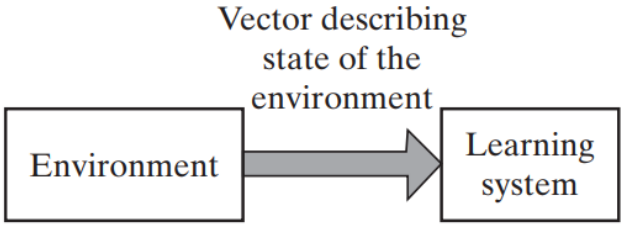
Άξιο συζήτησης είναι, επίσης, το γεγονός ότι υπάρχουν περιπτώσεις όπου ο πράκτορας δε θα έχει πρόσβαση σε όλες τις πληροφορίες μιας κατάστασης. Αν και είναι βολικό να υποθέσουμε ότι οι αισθητήρες του πράκτορα μπορούν να αντιληφθούν τη συνολική κατάσταση του περιβάλλοντος σε κάθε χρονικό βήμα, σε πολλές περιπτώσεις στην πράξη οι αισθητήρες παρέχουν μόνο μερική πληροφορία. Για παράδειγμα, ένα ρομπότ με μια κάμερα που δείχνει μπροστά δε μπορεί να δει τι βρίσκεται πίσω του. Σε αυτές τις περιπτώσεις, μπορεί να είναι υποχρεωτικό για τον πράκτορα να λάβει υπόψη του τις προηγούμενες παρατηρήσεις του για το περιβάλλον μαζί με τις τρέχουσες πληροφορίες από τους αισθητήρες του για την επιλογή της επόμενής του ενέργειας, και η καλύτερη πολιτική μπορεί να είναι η επιλογή ενεργειών που αποσκοπούν στη βελτίωση της ορατότητας των καταστάσεων του περιβάλλοντος [13].

Τέλος, αξίζει να σημειωθεί ότι η ενισχυτική μάθηση με καθυστέρηση είναι δύσκολο να εκτελεστεί. Αρχικά, δεν υπάρχει εκπαιδευτής για να παρέχει μια επιθυμητή απόκριση σε κάθε βήμα της διαδικασίας μάθησης. Επιπλέον, η καθυστέρηση με την οποία παράγεται το κύριο σήμα ενίσχυσης υποδηλώνει ότι η μανθάνουσα μηχανή πρέπει να λύσει ένα χρονικό πρόβλημα ανάθεσης εμπιστοσύνης. Δηλαδή, η μηχανή πρέπει να είναι σε θέση να καθορίζει το βαθμό επιτυχίας ατομικά για κάθε ενέργεια της χρονικής αλληλουχίας βημάτων που οδήγησαν στο τελικό αποτέλεσμα, ενώ ο κύριος μηχανισμός ενίσχυσης μπορεί να αποτιμά μόνο το τελικό αποτέλεσμα. Παρ' όλα αυτά, η ενισχυτική μάθηση με καθυστέρηση είναι μια ελκυστική μέθοδος. Παρέχει στο σύστημα μάθησης μια βάση για να επικοινωνεί με το περιβάλλον του, αναπτύσσοντας έτσι τη δυνατότητα

να μάθει να εκτελεί μια προκαθορισμένη εργασία βασιζόμενο αποκλειστικά στα αποτελέσματα της εμπειρίας του από την αλληλεπίδραση με το περιβάλλον [17].

3.2.2.2 Μη επιβλεπόμενη μάθηση

Στη μη επιβλεπόμενη μάθηση, ή αυτο-οργανούμενη μάθηση, δεν υπάρχει εξωτερικός εκπαιδευτής ή κριτής που να επιβλέπει τη διαδικασία μάθησης. Στη θέση του υπάρχει ένα ανεξάρτητο από την εργασία μέτρο της ποιότητας της αναπαράστασης που καλείται να μάθει το δίκτυο και οι ελεύθερες παράμετροι του δικτύου βελτιστοποιούνται σε σχέση με αυτό το μέτρο. Για ένα συγκεκριμένο ανεξάρτητο από την εργασία μέτρο, αφού το δίκτυο προσαρμοστεί στις στατιστικές κανονικότητες των δεδομένων εισόδου, αναπτύσσει τη δυνατότητα να σχηματίζει εσωτερικές αναπαραστάσεις για την κωδικοποίηση χαρακτηριστικών της εισόδου και, μέσω αυτών να δημιουργεί νέες κλάσεις αυτόματα [17, 46].



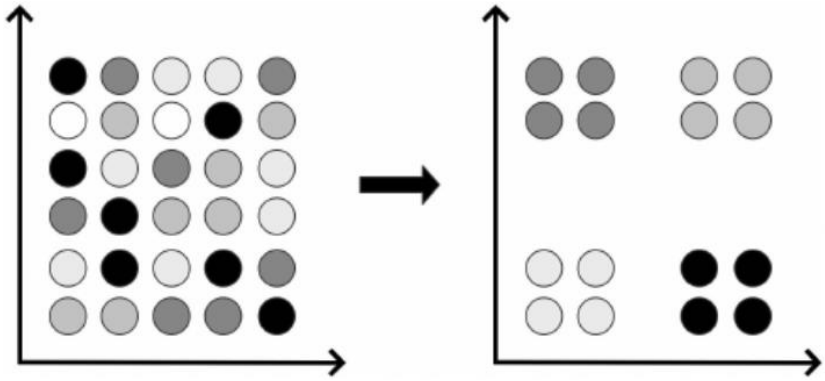
Σχήμα 5: Σχηματικό διάγραμμα μη επιβλεπόμενης μάθησης, από [17]

Η μη επιβλεπόμενη μάθηση αναφέρεται σε αλγορίθμους που χρησιμοποιούνται για την αναγνώριση μοτίβων σε dataset που περιέχουν δεδομένα που δεν είναι κατηγοριοποιημένα και δεν έχουν label. Συνεπώς, οι αλγόριθμοι καλούνται να ταξινομήσουν, να ομαδοποιήσουν και να κάνουν κατάλληλα label τα δεδομένα αυτά χωρίς να έχουν εξωτερική καθοδήγηση από κάποιον εκπαιδευτή κατά τη διάρκεια εκτέλεσης της εργασίας. Με άλλα λόγια, δε δίνονται labels στον αλγόριθμο εκμάθησης, ο οποίος καλείται να ανακαλύψει μια δομή για τα δεδομένα εισόδου. Αυτό το σύστημα τεχνητής νοημοσύνης θα ομαδοποιήσει την αταξινόμητη πληροφορία βασιζόμενο πάνω σε ομοιότητες και διαφορές μεταξύ των δεδομένων, καθώς δεν παρέχονται κλάσεις. Συμπερασματικά, ο κύριος στόχος της μη επιβλεπόμενης μάθησης είναι η ανακάλυψη κρυφών και ενδιαφερόντων μοτίβων σε δεδομένα χωρίς label. Οι τέσσερις τύποι μη επιβλεπόμενων εργασιών παρουσιάζονται παρακάτω και είναι: συσταδοποίηση (clustering), συσχέτιση (association), ανίχνευση ανωμαλιών (anomaly detection), και αυτόματοι κωδικοποιητές (autoencoders) [21].

Clustering

Το clustering είναι μια σημαντική έννοια της μη επιβλεπόμενης μάθησης. Κύριο έργο της είναι η εύρεση μιας δομής ή ενός μοτίβου σε μια συλλογή μη κατηγοριοποιημένων δεδομένων. Παράδειγμα αποτελεί η ομαδοποίηση παρόμοιων ταινιών για να προταθούν στους χρήστες μιας διαδικτυακής πλατφόρμας μετάδοσης

περιεχομένου τηλεοπτικών σειρών και ταινιών. Οι αλγόριθμοι υπεύθυνοι για το clustering επεξεργάζονται τα δεδομένα και βρίσκουν, αν υπάρχουν, clusters (ομάδες) μεταξύ αυτών. Ένα cluster είναι, συνεπώς, μια συλλογή αντικειμένων που είναι όμοια μεταξύ τους και ανόμοια με αντικείμενα που ανήκουν σε άλλα clusters. Υπάρχει και δυνατότητα προσαρμογής του αλγορίθμου για την εύρεση συγκεκριμένου αριθμού από clusters, γεγονός που επιτρέπει την αλλαγή του βαθμού ανάλυσης των ομάδων αυτών.



Σχήμα 6: Παράδειγμα clustering, από [21]

Υπάρχουν πολλοί τύποι clustering ανάλογα με τον τρόπο που ομαδοποιούνται τα δεδομένα, όπως partitioning (τμηματοποίηση), ιεραρχικό (hierarchical), επικαλυπτόμενο (overlapping), και πιθανολογικό (probabilistic). Στο partitioning, τα δεδομένα ομαδοποιούνται έτσι ώστε κάθε δεδομένο να ανήκει μόνο σε ένα cluster. Είναι, επίσης, γνωστό και ως αποκλειστικό (exclusive) clustering. Ένα παράδειγμα αλγορίθμου για αυτόν τον τύπο clustering είναι ο K-means. Έπειτα, στο ιεραρχικό clustering, κάθε δεδομένο αποτελεί ένα cluster. Οι επαναλαμβανόμενες ενώσεις μεταξύ των δύο κοντινότερων clusters μειώνουν τον αριθμό τους. Ύστερα, στο overlapping clustering χρησιμοποιούνται ασαφή σύνολα (fuzzy sets) για την ομαδοποίηση των δεδομένων. Έτσι, κάθε δεδομένο μπορεί να ανήκει σε δύο ή περισσότερα cluster με διαφορετικό βαθμό συμμετοχής. Τέλος, το πιθανολογικό clustering χρησιμοποιεί κατανομές πιθανοτήτων για τη δημιουργία των clusters [21].

Association

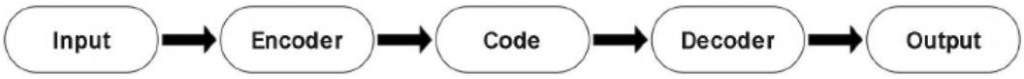
Το association χρησιμοποιείται για τον εντοπισμό ομοιοτήτων μεταξύ διαφορετικών αντικειμένων ενός συνόλου δεδομένων, όπως δεδομένα συναλλαγών ή οποιαδήποτε σχεσιακή βάση δεδομένων. Οι σχέσεις μεταξύ των αντικειμένων συνήθως αναπαρίσταται με μορφή κανόνων ή σετ συχνών αντικειμένων. Χρησιμοποιούνται ευρέως για την ανάλυση του καλαθιού αγοράς (ποια αντικείμενα αγοράζονται μαζί), clustering πελατών καταστημάτων (σε ποια καταστήματα έχουν τάση οι άνθρωποι να επισκέπτονται μαζί), bundling τιμών, διασταυρούμενες πωλήσεις, και άλλα. Μπορεί να θεωρηθεί μια προηγμένη μορφή ενός «what-if» σεναρίου [21].

Anomaly detection

Ανίχνευση ανωμαλιών είναι κάθε διαδικασία η οποία βρίσκει τα έκτοπα δεδομένα (outliers) ενός dataset. Αυτές οι ανωμαλίες μπορεί να αναφέρονται σε ασυνήθιστη δικτυακή κίνηση, ένα χαλασμένο αισθητήρα, ή δεδομένα που πρέπει να αφαιρεθούν πριν την ανάλυση για να μην την επηρεάσουν αρνητικά. Outlier θεωρείται ένα δεδομένα όταν τα μοτίβα του είναι πέρα από τις φυσιολογικές τιμές ή παρεκκλίνουν από τα φυσιολογικά μοτίβα. Για παράδειγμα, ένα outlier σε ένα δίκτυο μπορεί να σημαίνει ότι το χακαρισμένο δίκτυο στέλνει ευαίσθητο περιεχόμενο σε έναν μη εξουσιοδοτημένο εξυπηρετητή. Άλλες εφαρμογές αποτελούν ανίχνευση εισβολής και απάτης, και συστήματα στρατιωτικής παρακολούθησης [21].

Autoencoders

Οι autoencoders είναι μια τεχνική μη επιβλεπόμενης μάθησης που αξιοποιεί τις δυνατότητες των νευρωνικών δικτύων για την εργασία της μάθησης αναπαράστασης (representation learning). Είναι ένας ειδικός τύπος νευρωνικών δικτύων πρόσθιας τροφοδότησης στον οποίο η είσοδος είναι η ίδια με την έξοδο. Ένας αυτόματος κωδικοποιητής αποτελείται από τρία δομικά στοιχεία: τον κωδικοποιητή (encoder), το κωδικοποιημένο μήνυμα (code), και τον αποκωδικοποιητή (decoder). Ο κωδικοποιητής συμπυκνώνει τα δεδομένα εισόδου και, έτσι, παράγεται το κωδικοποιημένο μήνυμα. Ύστερα, ο αποκωδικοποιητής επανακατασκευάζει την είσοδο χρησιμοποιώντας αυτό το code. Για την κατασκευή ενός autoencoder απαιτούνται τρία συστατικά: μια μέθοδο κωδικοποίησης, μια μέθοδο αποκωδικοποίησης, και μια συνάρτηση απώλειας (loss function) για τη σύγκριση μεταξύ της εξόδου του συστήματος και της επιθυμητής απόκρισης. Η κωδικοποίηση επικυρώνεται και βελτιώνεται επιχειρώντας να αναδημιουργηθεί η είσοδος από την κωδικοποίηση. Ο autoencoder μαθαίνει μια αναπαράσταση (κωδικοποίηση) για ένα σύνολο δεδομένων, συνήθως για να πετύχουμε μείωση διαστάσεων, εκπαιδεύοντας το δίκτυο να αγνοεί ασήμαντα δεδομένα, δηλαδή το «θόρυβο» [21].



Σχήμα 7: Τα στάδια ενός autoencoder, από [21]

3.3 Νευρωνικά δίκτυα

Η ανάπτυξη των τεχνητών νευρωνικών δικτύων βασίστηκε στο γεγονός ότι η αλληλουχία ενεργειών και εργασιών που απαιτούνται για την εκτέλεση υπολογισμών στον ανθρώπινο εγκέφαλο διαφέρει δραστικά από την αντίστοιχη διαδικασία σε έναν συμβατικό ψηφιακό υπολογιστή. Ο εγκέφαλος είναι ένα εξαιρετικά πολύπλοκο σύστημα επεξεργασίας πληροφοριών, το οποίο έχει τη δυνατότητα να επεξεργάζεται παράλληλα και μη γραμμικά τα δεδομένα εισόδου. Με την κατάλληλη οργάνωση των δομικών του στοιχείων, δηλαδή των νευρώνων, επιτυγχάνει ιλιγγιώδη ταχύτητα στην εκτέλεση συγκεκριμένων υπολογισμών, πολλαπλάσια μεγαλύτερη από αυτή ενός υπερσύγχρονου

ηλεκτρονικού υπολογιστή. Παραδείγματα τέτοιων υπολογισμών αποτελούν η αναγνώριση προτύπων, η αντίληψη και ο έλεγχος κίνησης, η ανάλυση των ηχητικών κυμάτων, και η κατανόηση του νοήματος μιας πρότασης εντός κειμένου. Πιο συγκεκριμένα, η ανθρώπινη όραση είναι μια διαδικασία κατά την οποία ο εγκέφαλος επεξεργάζεται διαρκώς νέες πληροφορίες που παρέχονται από το οπτικό σύστημα, το οποίο προσφέρει μια συνεχώς μεταβαλλόμενη αναπαράσταση του περιβάλλοντος και, συνεπώς, τροφοδοτώντας τον εγκέφαλο με τα απαραίτητα δεδομένα εισόδου για την εργασία αναγνώρισης. Για να επιτύχουν καλή απόδοση, τα νευρωνικά δίκτυα χρησιμοποιούν τεράστιο αριθμό απλών, διασυνδεδεμένων μεταξύ τους υπολογιστικών κυττάρων, τα οποία αποκαλούνται «νευρώνες» ή «μονάδες επεξεργασίας» [12, 17, 47].

Ένα τεχνητό νευρωνικό δίκτυο ορίζεται ως εξής: ένας τεράστιος παράλληλος επεξεργαστής με κατανεμημένη αρχιτεκτονική, ο οποίος αποτελείται από απλές μονάδες επεξεργασίας και έχει από τη δύση του τη δυνατότητα να αποθηκεύει εμπειρική γνώση και να την καθιστά διαθέσιμη για χρήση. Μοιάζει με τον ανθρώπινο εγκέφαλο σε δύο σημεία [17]:

1. Το δίκτυο προσλαμβάνει τη γνώση από το περιβάλλουν του, μέσω μιας διαδικασίας μάθησης
2. Η ισχύς των συνδέσεων μεταξύ των νευρώνων, που αποκαλείται συναπτικό βάρος, χρησιμοποιείται για την αποθήκευση της γνώσης που αποκτιέται

Η διαδικασία μέσω της οποίας επιτυγχάνεται η μάθηση αποκαλείται αλγόριθμος μάθησης και η λειτουργία του είναι να τροποποιεί τα συναπτικά βάρη του δικτύου με τον κατάλληλο τρόπο για την επίτευξη του επιθυμητού στόχου [48].

3.3.1 Ιδιότητες νευρωνικών δικτύων

Τα νευρωνικά δίκτυα προσφέρουν τις ακόλουθες χρήσιμες ιδιότητες και δυνατότητες [17]:

Μη γραμμικότητα

Οι νευρώνες έχουν τη δυνατότητα να είναι είτε γραμμικοί είτε μη γραμμικοί. Ένα νευρωνικό δίκτυο που απαρτίζεται από διασυνδεδεμένους μη γραμμικούς νευρώνες είναι κι αυτό μη γραμμικό. Η ιδιότητα αυτή είναι σημαντική σε περιπτώσεις στις οποίες ο φυσικός μηχανισμός που παρέχει τα σήματα εισόδου είναι εκ φύσεως μη γραμμικός, για παράδειγμα η ομιλία.

Αντιστοίχιση εισόδου – εξόδου

Για την κατανόηση αυτής της ιδιότητας θα εξετάσουμε το παράδειγμα της μάθησης χωρίς εκπαιδευτή, ή επιβλεπόμενη μάθηση, κατά την οποία τα συναπτικά βάρη

του νευρωνικού δικτύου τροποποιούνται με την εισαγωγή χαρακτηρισμένων παραδειγμάτων εκπαίδευσης. Δηλαδή, τροφοδοτούμε το νευρωνικό δίκτυο με ένα σύνολο δεδομένων εισόδου για τα οποία γνωρίζουμε την αντίστοιχη επιθυμητή απόκριση. Για κάθε νέο σήμα εισόδου, τα συναπτικά βάρη τροποποιούνται για να προσεγγίσουν την επιθυμητή έξοδο, ελαχιστοποιώντας τη διαφορά μεταξύ της επιθυμητής και πραγματικής απόκρισης. Η εκπαίδευση του δικτύου συνεχίζει για τα υπόλοιπα σήματα εισόδου του συνόλου δεδομένων. Τερματίζει όταν το δίκτυο φτάσει σε ευσταθή κατάσταση, δηλαδή όταν με την εισαγωγή νέου σήματος εισόδου δεν εμφανίζεται σημαντική μεταβολή στα συναπτικά βάρη του δικτύου. Τα εφαρμοζόμενα παραδείγματα εκπαίδευσης θα μπορούσαν να εφαρμοστούν εκ νέου κατά τη διάρκεια της εκπαίδευσης, αλλά με διαφορετική σειρά. Έτσι, το δίκτυο μαθαίνει από τα παραδείγματα, κατασκευάζοντας μια αντιστοίχιση εισόδου – εξόδου για το δοθέν πρόβλημα.

Προσαρμοστικότητα

Λόγω της μεθόδου μάθησης που βασίζεται στην τροποποίηση των συναπτικών βαρών για την εκπαίδευση, τα νευρωνικά δίκτυα παρουσιάζουν σημαντική προσαρμοστικότητα καθώς ανάλογα με τις μεταβολές που συμβαίνουν στο περιβάλλουν τους μεταβάλλονται και τα βάρη. Έτσι, λαμβάνοντας νέα σήματα εισόδου, ένα νευρωνικό δίκτυο μπορεί να επανεκπαιδευτεί με σχετική ευκολία για να χειριστεί τις μεταβολές του περιβάλλοντος λειτουργίας του. Σαν γενικό κανόνα, θεωρούμε ότι όσο πιο προσαρμοστικό είναι ένα σύστημα, διασφαλίζοντας ταυτόχρονα ότι παραμένει διαρκώς σταθερό, τόσο πιο εύρωστο θα είναι και τόσο καλύτερα θα αποδίδει όταν κληθεί να λειτουργήσει σε ένα μη σταθερό περιβάλλον. Ωστόσο, αξίζει να σημειωθεί ότι η προσαρμοστικότητα δεν συνεπάγεται πάντα την ευρωστία. Το γεγονός αυτό εξαρτάται από τις σταθερές χρόνου που χρησιμοποιούνται για την (επαν)εκπαίδευση του συστήματος και τις συνθήκες του περιβάλλοντος στο οποίο βρίσκεται. Πιο συγκεκριμένα, σταθερές χρόνου μικρής διαρκείας μπορεί να οδηγήσουν σε γρήγορη αλλαγή κατάστασης, δηλαδή το σύστημα είναι πολύ ευαίσθητο στις διαταραχές. Αν αυτές οι διαταραχές είναι πλασματικές, οι συνεχείς μεταβολές στα συναπτικά βάρη του δικτύου είναι πολύ πιθανό να οδηγήσουν σε δραματική μείωση της απόδοσης του συστήματος. Συνεπώς, για την αξιοποίηση όλων των πλεονεκτημάτων της προσαρμοστικότητας, οι σταθερές χρόνου πρέπει να οριστούν έτσι ώστε το σύστημα να αγνοεί τις πλασματικές διαταραχές, αλλά ταυτόχρονα να αποκρίνεται σε μεταβολές του περιβάλλοντος που έχουν πραγματικά σημασία. Συνολικά, η φυσική αρχιτεκτονική ενός νευρωνικού δικτύου για ταξινόμηση προτύπων, επεξεργασία σήματος και εφαρμογές ελέγχου, σε συνδυασμό με την προσαρμοστική δυνατότητα του δικτύου, το καθιστά χρήσιμο εργαλείο για την προσαρμοστική ταξινόμηση προτύπων, την προσαρμοστική επεξεργασία σήματος, και τον προσαρμοστικό έλεγχο συστημάτων.

Ενδεικτική απόκριση

Στο πλαίσιο της ταξινόμησης προτύπων, υπάρχει η δυνατότητα να σχεδιαστεί νευρωνικό δίκτυο έτσι ώστε να παρέχει πληροφορία και για το ποιο συγκεκριμένο πρότυπο θα επιλεγεί, αλλά και για τον αντίστοιχο βαθμό εμπιστοσύνης για την εκάστοτε επιλογή. Αυτός ο βαθμός εμπιστοσύνης μπορεί να αξιοποιηθεί για την απόρριψη διφορούμενων μοτίβων και, συνεπώς, τη βελτίωση της απόδοσης του συστήματος.

Πληροφορία σχετική με το περιεχόμενο

Η γνώση αντιπροσωπεύεται από την ίδια τη δομή και την κατάσταση ενεργοποίησης ενός νευρωνικού δικτύου. Κάθε νευρώνας στο δίκτυο ενδεχομένως να επηρεάζεται από την συνολική δραστηριότητα όλων των άλλων νευρώνων του δικτύου. Αυτό σημαίνει ότι ένα νευρωνικό δίκτυο χειρίζεται με φυσικό τρόπο τη σχετική με το περιεχόμενο πληροφορία (contextual information).

Ανοχή σε βλάβες

Ένα νευρωνικό δίκτυο, υλοποιημένο σε μορφή hardware, έχει την εγγενή δυνατότητα να είναι ανεκτικό σε βλάβες, ή εύρωστο, υπό την έννοια ότι η απόδοσή του μειώνεται βαθμιαία και ομαλά υπό αντίξοες συνθήκες λειτουργίας. Για να διασφαλιστεί ότι ένα νευρωνικό δίκτυο είναι πράγματι ανθεκτικό σε βλάβες, μπορεί να χρειαστεί να διαμορφωθεί κατάλληλα ο αλγόριθμος που χρησιμοποιείται για την εκπαίδευση του δικτύου ώστε να λάβει διορθωτικά μέτρα σε περίπτωση μερικής αποτυχίας του συστήματος. Γενικότερα, επειδή ένα νευρωνικό δίκτυο αποτελεί ένα κατανεμημένο σύστημα νευρώνων, μια βλάβη σε ένα σημείο του συστήματος δε θα λάβει μεγάλη έκταση και, κατά συνέπεια, η απόδοση του συστήματος μειώνεται ομαλά και δεν παρουσιάζει πλήρη αποτυχία λειτουργίας σε τέτοιες μερικές βλάβες, για παράδειγμα καταστροφή ενός νευρώνα ή των συνδέσεών του.

Δυνατότητα υλοποίησης σε VLSI

Η μαζικά παράλληλη φύση ενός νευρωνικού δικτύου το καθιστά κατάλληλο για το γρήγορο υπολογισμό συγκεκριμένων εργασιών και για την υλοποίησή του με χρήση τεχνολογίας πολύ μεγάλης κλίμακας ολοκλήρωσης (VLSI). Έτσι, τα νευρωνικά δίκτυα μπορούν να αξιοποιήσουν και τα πλεονεκτήματα της τεχνολογίας VLSI, όπως η παροχή ενός μέσου «σύλληψης» ιδιαίτερα πολύπλοκης συμπεριφοράς με εξαιρετικά ιεραρχικό τρόπο.

Ομοιομορφία ανάλυσης και σχεδίασης

Στο σχεδιασμό και την ανάλυση των νευρωνικών συστημάτων χρησιμοποιείται κοινή και ομοιόμορφη σημειογραφία σε όλα τα πεδία εφαρμογής τους. Αυτό πηγάζει από την κοινή δομή του νευρώνα ανεξαρτήτως του νευρωνικού δικτύου στο οποίο

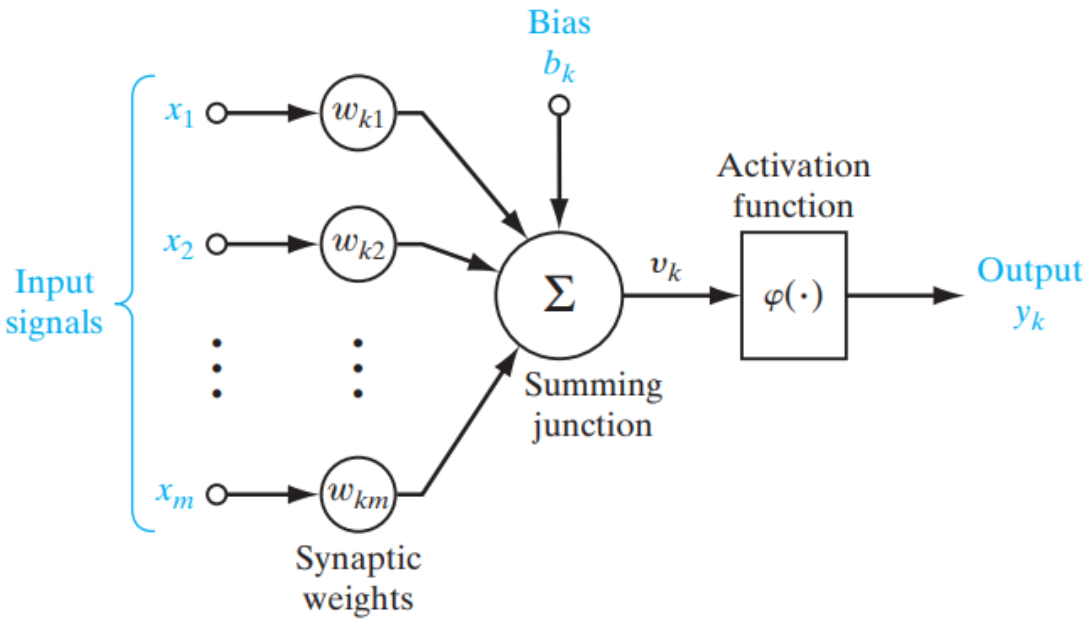
εντάσσεται, αλλάζοντας ουσιαστικά μόνο τα συναπτικά του βάρη κατάλληλα για την εκάστοτε εφαρμογή. Έτσι, είναι εφικτή η καθολική χρήση θεωριών και αλγορίθμων μάθησης σε διαφορετικές εφαρμογές που χρησιμοποιούν νευρωνικά δίκτυα. Αυτό με τη σειρά του καθιστά εφικτή και την κατασκευή δομοστοιχειωτών δικτύων με απρόσκοπτη ενοποίηση επιμέρους λειτουργικών μονάδων.

Αναλογία με τη νευροφυσιολογία του εγκεφάλου

Όπως αναφέρθηκε και στην εισαγωγή του κεφαλαίου, η ιδέα σχεδίασης των νευρωνικών δικτύων είναι βασισμένη στη λειτουργία του ανθρώπινου εγκεφάλου, ο οποίος αποτελεί απόδειξη ότι η εύρωστη, παράλληλη επεξεργασία όχι μόνο είναι φυσικά εφικτή, αλλά ταχύτατη και ισχυρή. Οι νευροβιολόγοι αντιμετωπίζουν τα τεχνητά νευρωνικά δίκτυα ως ένα εργαλείο έρευνας της ερμηνείας νευροβιολογικών φαινομένων. Για παράδειγμα, βασισμένοι στο έργο του Anastasio (1993), συγκρίνονται γραμμικά μοντέλα του αιθουσοοφθαλμικού αντανακλαστικού με μοντέλα νευρωνικών δικτύων που βασίζονται σε ανατροφοδοτούμενα δίκτυα. Παράλληλα, οι μηχανικοί αναζητούν στον τομέα της νευροβιολογίας ιδέες που θα μπορέσουν να επιλύσουν προβλήματα που είναι πιο πολύπλοκα από τις συμβατικές τεχνικές σχεδίασης. Παράδειγμα αποτελεί η προσπάθεια κατασκευής ηλεκτρονικών ολοκληρωμένων κυκλωμάτων τα οποία μιμούνται τη δομή του αμφιβληστροειδή (νευρομορφικά κυκλώματα).

3.3.2 Μοντέλο τεχνητού νευρώνα

Ένας νευρώνας είναι μια μονάδα επεξεργασίας πληροφορίας, η οποία είναι θεμελιώδης για τη λειτουργία ενός νευρωνικού δικτύου. Το μοντέλο ενός τεχνητού νευρώνα παρουσιάζεται στο παρακάτω διάγραμμα:



Σχήμα 8: Μη γραμμικό μοντέλο νευρώνα, από [17]

Αρχικά, στο διάγραμμα φαίνονται τα σήματα εισόδου (input signals) $x_1, x_2, \dots, x_m$, τα οποία σε συνδυασμό με τα συναπτικά βάρη (synaptic weights) αποτελούν ένα σύνολο συνάψεων που καταλήγουν στον κόμβο άθροισης (summing junction). Στον κόμβο αυτό, αθροίζονται τα σήματα εισόδου, σταθμισμένα από τα αντίστοιχα συναπτικά βάρη του νευρώνα. Δηλαδή, μια οποιαδήποτε είσοδος x_j πολλαπλασιάζεται από το αντίστοιχο βάρος w_{kj} της σύναψης j που συνδέεται με τον νευρώνα k. Σημειώνουμε ότι το συναπτικό βάρος ενός τεχνητού νευρώνα μπορεί να λάβει και αρνητικές και θετικές τιμές [17].

Έπειτα, η έξοδος του αθροιστή αποτελεί είσοδο της συνάρτησης ενεργοποίησης (activation function) $\varphi(\cdot)$. Η συνάρτηση αυτή χρησιμοποιείται για τον περιορισμό του πλάτους του σήματος εξόδου ενός νευρώνα. Τυπικά, το κανονικοποιημένο εύρος τιμών πλάτους της εξόδου ενός νευρώνα είναι είτε το διάστημα [0, 1] είτε το [-1, 1]. Οι δύο βασικοί τύποι συναρτήσεων ενεργοποίησης περιγράφονται στην επόμενη παράγραφο.

Τέλος, στο μοντέλο του σχήματος φαίνεται και μια επιπλέον είσοδος στον κόμβο άθροισης, η εξωτερικά εφαρμοζόμενη πόλωση (bias) b_k. Η πόλωση έχει ως αποτέλεσμα την αύξηση ή μείωση της δικτυακής διέγερσης της συνάρτησης ενεργοποίησης, ανάλογα με το εάν είναι θετική ή αρνητική αντίστοιχα.

Από τα παραπάνω προκύπτουν οι εξής μαθηματικές εξισώσεις για τις εξόδους v_k του κόμβου άθροισης (τοπικό πεδίο του νευρώνα) και y_k του νευρώνα [17]:

$$v_k = \sum_{j=1}^{m} w_{kj} x_j + b_k$$

$$\text{και } y_k = \varphi(v_k)$$

3.3.3 Τύποι συνάρτησης ενεργοποίησης

Όπως αναφέρθηκε και στην προηγούμενη παράγραφο, οι συναρτήσεις ενεργοποίησης χρησιμοποιούνται στα νευρωνικά δίκτυα για να ρυθμίζουν τις τιμές εξόδου του δικτύου. Χρησιμοποιούνται ευρέως σε πολλούς διαφορετικούς τομείς, όπως αναγνώριση προτύπων και ταξινόμηση, αναγνώριση φωνής, ανίχνευση δακτυλικού αποτυπώματος, και πρόβλεψη καιρού. Παρακάτω παρουσιάζονται οι κύριοι τύποι των συναρτήσεων ενεργοποίησης που χρησιμοποιούνται στα σύγχρονα νευρωνικά δίκτυα [22].

Συνάρτηση κατωφλίου (threshold function)

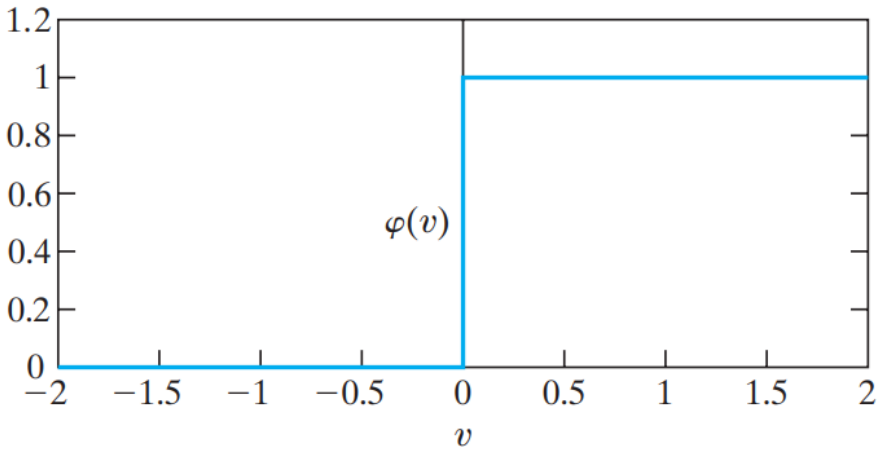
Αυτός ο τύπος συνάρτησης ενεργοποίησης περιγράφεται από την παρακάτω μαθηματική σχέση:

$$\varphi(v) = \begin{cases} 1 \text{ εάν } v \geq 0 \\ 0 \text{ εάν } v < 0 \end{cases}.$$

Στους κλάδους της μηχανικής, αυτή η μορφή συνάρτησης κατωφλίου αναφέρεται ως συνάρτηση Heaviside. Χρησιμοποιώντας την παραπάνω συνάρτηση, η έξοδος ενός νευρώνα k ορίζεται ως:

$$y_k = \begin{cases} 1 \text{ εάν } v_k \geq 0 \\ 0 \text{ εάν } v_k < 0' \end{cases}$$

όπου v_k είναι το τοπικό πεδίο του νευρώνα που αναλύθηκε στην προηγούμενη παράγραφο. Το μοντέλο αυτό αναφέρεται ως μοντέλο McCulloch – Pitts, εις αναγνώριση του πρωτοποριακού έργου των δύο αυτών επιστημόνων στον τομέα των νευρωνικών δικτύων [17].



Σχήμα 9: Γραφική παράσταση συνάρτησης κατωφλίου, από [17]

Σιγμοειδής συνάρτηση (sigmoid function)

Η σιγμοειδής συνάρτηση ενεργοποίησης, η οποία αποκαλείται και λογιστική συνάρτηση σε κάποιες βιβλιογραφίες, είναι μια μη γραμμική, αυστηρά αύξουσα συνάρτηση η οποία χρησιμοποιείται στα επίπεδα εξόδου των νευρωνικών δικτύων για την πρόβλεψη βάσει πιθανοτήτων, για παράδειγμα σε προβλήματα δυαδικής ταξινόμησης. Η μαθηματική σχέση που την εκφράζει είναι η ακόλουθη:

$$\varphi(v) = \frac{1}{1 + e^{-\alpha v}},$$

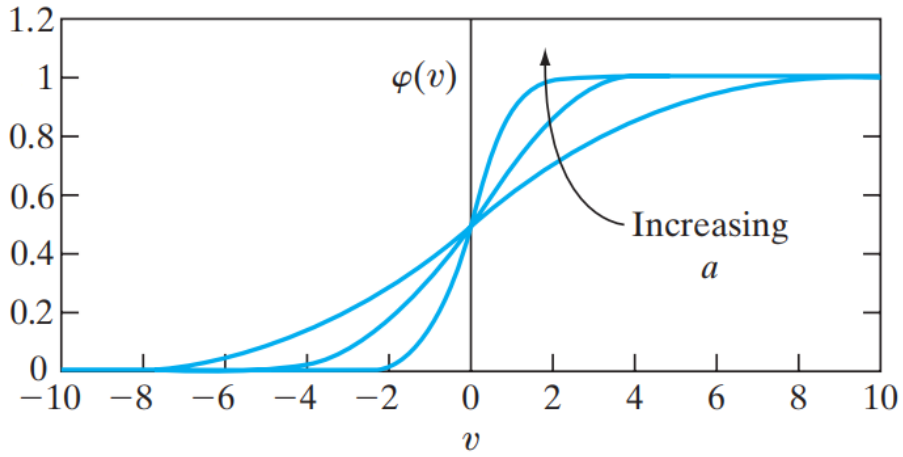
όπου α η παράμετρος κλίσης της συνάρτησης. Μεταβάλλοντας την παράμετρο αυτή λαμβάνουμε σιγμοειδής συναρτήσεις διαφορετικών κλίσεων. Λαμβάνει το όνομά της από τη μορφή της γραφικής της παράσταση, που παρουσιάζεται στο παρακάτω σχήμα και μοιάζει με «S» [17].

Όπως φαίνεται και στο παρακάτω διάγραμμα, η συνάρτηση λαμβάνει τιμές από ένα συνεχές πεδίο τιμών, το διάστημα [0, 1]. Σε ορισμένες περιπτώσεις, ανάλογα με την εκάστοτε εφαρμογή, είναι ωφέλιμο η έξοδος της συνάρτησης ενεργοποίησης να λαμβάνει τιμές από το διάστημα [-1, 1]. Σε αυτήν την περίπτωση, η συνάρτηση κατωφλίου που ορίστηκε παραπάνω παίρνει την εξής μορφή (συνάρτηση προσήμου):

$$\varphi(v) = \begin{cases} 1 \text{ εάν } v > 0 \\ 0 \text{ εάν } v = 0 \\ -1 \text{ εάν } v < 0 \end{cases},$$

ενώ στην αντίστοιχη περίπτωση για τη σιγμοειδή συνάρτηση χρησιμοποιείται η συνάρτηση υπερβολικής εφαπτομένης:

$$\varphi(v) = \tanh(v) = \frac{e^{v} - e^{-v}}{e^{v} + e^{-v}}.$$



Σχήμα 10: Γραφική παράσταση σιγμοειδής συνάρτησης, από [17]

3.3.4 Αρχιτεκτονικές δικτύων

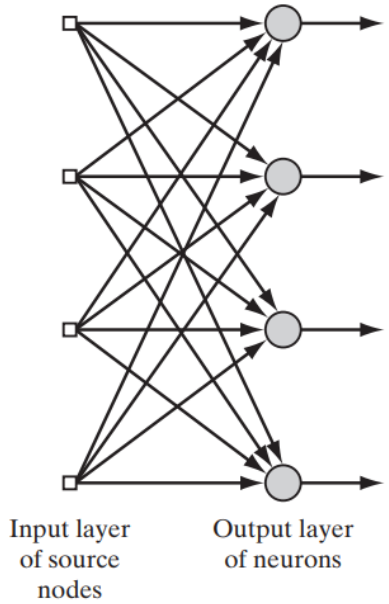
Υπάρχουν δύο κύριες κατηγορίες αρχιτεκτονικών δικτύων ανάλογα με τον τύπο των συνδέσεων μεταξύ των νευρώνων, τα δίκτυα πρόσθιας τροφοδότησης (feedforward networks) και τα αναδρομικά δίκτυα (recurrent networks). Αν δεν υπάρχει «feedback», δηλαδή αν οι έξοδοι των νευρώνων δεν συνδέονται απευθείας με τα σήματα εισόδου, το δίκτυο χαρακτηρίζεται ως δίκτυο πρόσθιας τροφοδότησης. Στην αντίθετη περίπτωση, αν υπάρχει «feedback», δηλαδή μια συναπτική σύνδεση από τις εξόδους προς τις εισόδους (είτε τις δικές τους εισόδους ή τις εισόδους άλλων νευρώνων), το δίκτυο αποκαλείται αναδρομικό [23].

Συνήθως, τα νευρωνικά δίκτυα είναι οργανωμένα σε επίπεδα (layers). Τα δίκτυα πρόσθιας τροφοδότησης χωρίζονται σε δύο κατηγορίες, ανάλογα με τον αριθμό των επιπέδων αυτών. Έτσι, έχουμε τα ενός επιπέδου δίκτυα πρόσθιας τροφοδότησης και τα πολυεπίπεδα δίκτυα πρόσθιας τροφοδότησης. Οι διαφορετικοί τύποι των αρχιτεκτονικών νευρωνικών δικτύων αναλύονται παρακάτω.

Ενός επιπέδου δίκτυα πρόσθιας τροφοδότησης (Single-layer feedforward networks)

Στην απλούστερη δυνατή μορφή ενός δικτύου, ένα νευρωνικό δίκτυο έχει ένα επίπεδο εισόδου αποτελούμενο από πηγαίους κόμβους (input layer of source nodes) το οποίο συνδέεται απευθείας με ένα επίπεδο νευρώνων εξόδου αποτελούμενο από υπολογιστικούς κόμβους (output layer of neurons), και όχι αντίστροφα αφού είναι

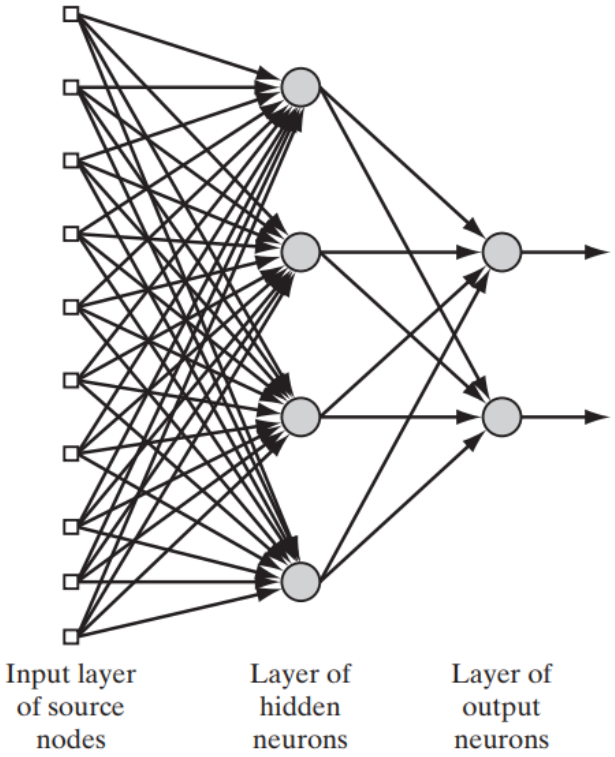
δίκτυο πρόσθιας τροφοδότησης. Αυτό το δίκτυο αποκαλείται ενός επιπέδου, με το επίπεδο αυτό να αναφέρεται στους κόμβους εξόδου, καθώς οι πηγές εισόδου δεν προσμετρούνται γιατί δεν πραγματοποιείται κάποιος υπολογισμός σε εκείνο το επίπεδο [17].



Σχήμα 11: Δίκτυο πρόσθιας τροφοδότησης με ένα μεμονωμένο επίπεδο νευρώνων, από [17]

Πολυεπίπεδα δίκτυα πρόσθιας τροφοδότησης (Multilayer feedforward networks)

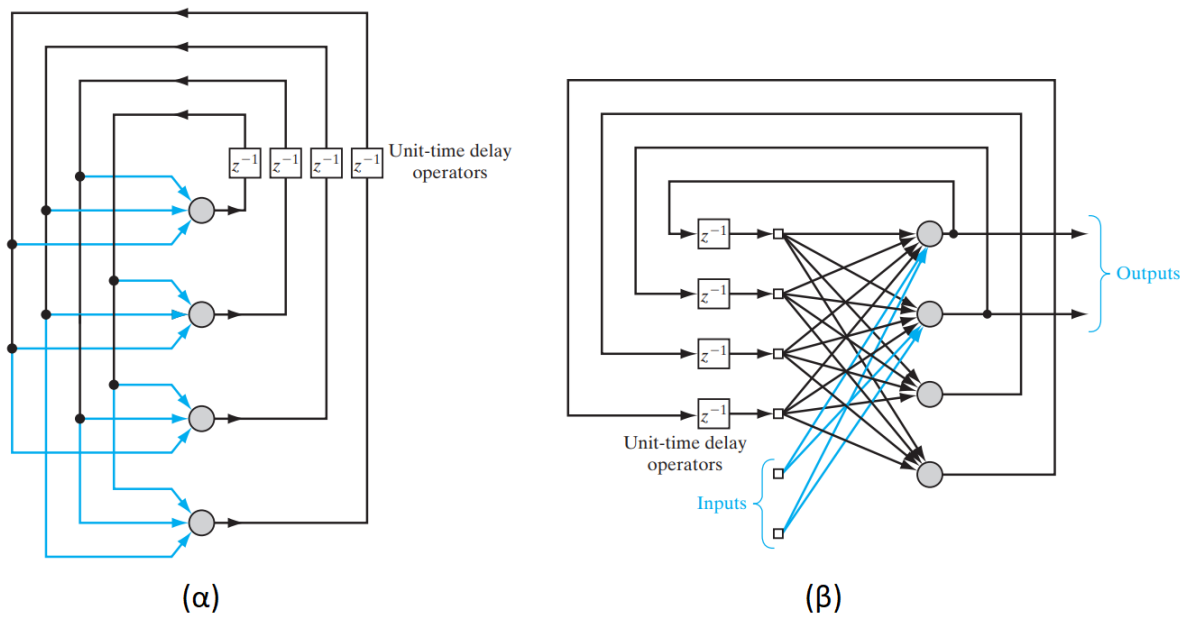
Ως επέκταση των δικτύων ενός επιπέδου πρόσθιας τροφοδότησης δημιουργήθηκαν τα πολυεπίπεδα δίκτυα που χαρακτηρίζονται από την παρουσία ενός ή περισσότερων κρυφών επιπέδων. Οι υπολογιστικοί κόμβοι των επιπέδων αυτών αποκαλούνται κρυφοί νευρώνες ή κρυφές μονάδες, λόγω του γεγονότος ότι τα επίπεδα αυτά δεν είναι ορατά ούτε από την είσοδο ούτε από την έξοδο του δικτύου. Με την προσθήκη ενός ή περισσότερων κρυφών επιπέδων, το δίκτυο αποκτά τη δυνατότητα εξαγωγής στατιστικών υψηλότερης τάξης από την είσοδό του. Το πρώτο κρυφό επίπεδο έχει ως εισόδους τα σήματα εισόδων των πηγών. Τα επόμενα επίπεδα έχουν ως εισόδους τα σήματα εξόδου του προηγούμενου επιπέδου, ενώ το τελικό επίπεδο (επίπεδο εξόδου) αποτελεί τη συνολική απόκριση του δικτύου στο πρότυπο ενεργοποίησης που παρέχεται από τους πηγαίους κόμβους [17].



Σχήμα 12: Πλήρες συνδεδεμένο δίκτυο πρόσθιας τροφοδότησης με ένα κρυφό επίπεδο και ένα επίπεδο εξόδου, από [17]

Αναδρομικά δίκτυα (Recurrent networks)

Τα αναδρομικά δίκτυα, όπως αναφέρθηκε και προηγουμένως, διαφέρουν από τα δίκτυα πρόσθιας τροφοδότησης στο ότι έχουν «feedback», δηλαδή έχουν τουλάχιστον ένα βρόγχο ανάδρασης. Τα δίκτυα αυτά χρησιμοποιούνται κυρίως για την αναγνώριση μοτίβων σε μια σειρά δεδομένων. Τα δεδομένα αυτά μπορεί να είναι για παράδειγμα γραπτό κείμενο, γενετικός κώδικας, ή κάποια χρονοσειρά τιμών [24].



Σχήμα 13: Αναδρομικό δίκτυο (α) χωρίς βρόγχους αυτο-ανάδρασης και κρυφούς νευρώνες (β) με κρυφούς νευρώνες, από [17]

Ένα αναδρομικό δίκτυο μπορεί να περιέχει αυτο-ανάδραση, που ονομάζεται η κατάσταση κατά την οποία η έξοδος ενός νευρώνα επανατροφοδοτείται στην είσοδο του ίδιου νευρώνα. Η παρουσία βρόγχων ανάδρασης έχει βαθιά επίδραση στη δυνατότητα μάθησης του δικτύου και στην απόδοσή του. Σημειώνουμε ότι οι βρόγχοι ανάδρασης προϋποθέτουν τη χρήση συγκεκριμένων κλάδων, αποτελούμενων από στοιχεία μοναδιαίας χρονικής καθυστέρησης (z^{-1}), τα οποία έχουν ως αποτέλεσμα μη γραμμική δυναμική συμπεριφορά, εφόσον το νευρωνικό δίκτυο περιέχει μη γραμμικές μονάδες. Στο παραπάνω διάγραμμα παρουσιάζονται δύο παραδείγματα αναδρομικών νευρωνικών δικτύων [17].

3.4 Logistic Regression Classifier

Υπό γενικές προϋποθέσεις, η εκ των υστέρων πιθανότητα (posterior probability) μιας κλάσης C_1 μπορεί να γραφεί ως εξής:

$$p(C_1|\varphi) = y(\varphi) = \sigma(w^T\varphi),$$

όπου φ η είσοδος, το διάνυσμα των χαρακτηριστικών (feature vector), και $\sigma(\cdot)$ η λογιστική σιγμοειδής συνάρτηση, η οποία αναλύθηκε σε προηγούμενη παράγραφο. Στην ορολογία της στατιστικής, το μοντέλο αυτό είναι γνωστό ως logistic regression, αλλά σημειώνουμε ότι το συγκεκριμένο μοντέλο χρησιμοποιείται για ταξινόμηση και όχι regression. Για ένα M διαστάσεων feature vector φ, το μοντέλο αυτό έχει M ρυθμιζόμενες παραμέτρους [16].

Για την εκμάθηση του συγκεκριμένου μοντέλου χρειάζονται δύο στοιχεία. Αρχικά, μια μετρική του πόσο κοντά είναι η πρόβλεψη του μοντέλου στην επιθυμητή απόκριση. Συνήθως, μετριέται η απόσταση των δύο αυτών εξόδων, και όχι η ομοιότητα μεταξύ τους. Αυτό γίνεται με τη χρήση μιας συνάρτησης σφάλματος. Στο logistic regression συνήθως χρησιμοποιείται η λεγόμενη cross-entropy συνάρτηση, που αναλύεται στη συνέχεια. Το δεύτερο που απαιτείται είναι ένας αλγόριθμος βελτιστοποίησης, για την επαναλαμβανόμενη ενημέρωση των συναπτικών βαρών με στόχο την ελαχιστοποίηση της συνάρτησης σφάλματος [25].

Για τον καθορισμό των παραμέτρων του logistic regression μοντέλου χρησιμοποιούμε μέγιστη πιθανοφάνεια (maximum likelihood). Για να γίνει αυτό, χρησιμοποιούμε την παράγωγο της λογιστικής σιγμοειδής συνάρτησης, η οποία μπορεί να εκφραστεί βάσει του εαυτού της

$$\frac{d\sigma}{d\alpha} = \sigma(1-\sigma).$$

Για ένα dataset $\{\varphi_n, t_n\}$, όπου $t_n \in \{0,1\}$ και $\varphi_n = \varphi(x_n)$ με $n = 1, 2, \dots, N$, η συνάρτηση πιθανοφάνειας μπορεί να γραφεί

$$p(t|w) = \prod_{n=1}^{N} y_n^{t_n} \{1-y_n\}^{1-t_n},$$

όπου $t = (t_1, t_2, \dots, t_N)^T$ και $y_n = p(C_1|\varphi_n)$. Μπορούμε να ορίσουμε μια συνάρτηση σφάλματος παίρνοντας τον αρνητικό λογάριθμο της πιθανοφάνειας, το οποίο μας δίνει την cross-entropy συνάρτηση σφάλματος της μορφής

$$E(w) = -\ln p(t|w) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\},$$

όπου $y_n = \sigma(\alpha_n)$ και $\alpha_n = w^T \varphi_n$. Παίρνοντας την κλίση (ανάδελτα) της συνάρτησης σφάλματος ως προς w, έχουμε

$$\nabla E(w) = \sum_{n=1}^{N} (y_n - t_n)\varphi_n.$$

Παρατηρούμε ότι ο παράγοντας που σχετιζόταν με τη λογιστική σιγμοειδή έχει απαλειφθεί, γεγονός που οδηγεί σε μια απλουστευμένη μορφή για την κλίση του λογαρίθμου της πιθανοφάνειας. Πιο συγκεκριμένο, η συνεισφορά στην κλίση από το δεδομένα n δίνεται από το σφάλμα $y_n - t_n$ μεταξύ της επιθυμητής απόκρισης και της πρόβλεψης του μοντέλου, πολλαπλασιασμένο από το feature vector φ_n [16].

Αξίζει να σημειωθεί ότι η μέγιστη πιθανοφάνεια μπορεί να παρουσιάσει σημαντικό overfitting σε dataset τα οποία είναι γραμμικώς διαχωρίσιμα. Αυτό συμβαίνει, επειδή η λύση της μέγιστης πιθανοφάνειας προκύπτει όταν το υπερεπίπεδο (hyperplane) που αντιστοιχεί στο $\sigma = 0.5$, το οποίο ισοδυναμεί με $w^T \varphi = 0$, διαχωρίζει τις δύο κλάσεις και η τάξη μεγέθους του w τείνει στο άπειρο. Σε αυτήν την περίπτωση, η λογιστική σιγμοειδής συνάρτηση γίνεται απεριόριστα απότομη στο χώρο των χαρακτηριστικών (feature space), έτσι ώστε σε κάθε σημείο εκπαίδευσης από την κλάση k να αντιστοιχίζεται η εκ των υστέρων πιθανότητα $p(C_k|x) = 1$. Επιπλέον, υπάρχει κατά κανόνα ένα συνεχές σετ τέτοιων λύσεων γιατί κάθε διαχωριζόμενο υπερεπίπεδο θα δώσει την ίδια εκ των υστέρων πιθανότητα στα σημεία εκπαίδευσης. Η μέγιστη πιθανοφάνεια δεν προσφέρει κάποιον τρόπο για την επιλογή μιας λύσης αντί κάποιας λύσης, και, συνεπώς, ποια λύση τελικά θα επιλεχθεί εξαρτάται τελικά από τον αλγόριθμο βελτιστοποίησης και την αρχικοποίηση των παραμέτρων. Σημειώνουμε ότι το πρόβλημα θα προκύψει ακόμα και αν ο αριθμός των σημείων δεδομένων είναι μεγάλος σε σχέση με τον αριθμό των παραμέτρων του μοντέλου, αρκεί το dataset εκπαίδευσης να είναι γραμμικά διαχωρίσιμο. Αυτή η ιδιαιτερότητα μπορεί να αποφευχθεί προσθέτοντας έναν παράγοντα κανονικοποίησης στη συνάρτηση σφάλματος. Έτσι, έχουμε το πρόβλημα βελτιστοποίησης στο binary logistic regression με παράγοντα κανονικοποίησης $r(w)$ για την ελαχιστοποίηση της παρακάτω συνάρτησης απώλειας:

$$\min_w C \sum_{n=1}^{N} \{t_n \ln y_n - (1 - t_n) \ln(1 - y_n)\} + r(w)$$

η οποία χρησιμοποιείται από τον ταξινομητή logistic regression της βιβλιοθήκης scikit-learn της γλώσσας Python, και χρησιμοποιήθηκε για τη διεκπεραίωση αυτής της εργασίας [16, 26].

3.5 Gaussian Naive Bayes Classifier

Οι naive Bayes μέθοδοι είναι ένα σύνολο αλγορίθμων επιβλεπόμενης μάθησης που βασίζονται στο θεώρημα Bayes με την «αφελή» υπόθεση της υπό όρους ανεξαρτησίας μεταξύ κάθε ζευγαριού των features, με δεδομένη την τιμή της μεταβλητής της κλάσης. Το θεώρημα του Bayes ορίζει την ακόλουθη σχέση [49]:

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y)P(x_1, x_2, \dots, x_n|y)}{P(x_1, x_2, \dots, x_n)},$$

όπου y η μεταβλητή κλάσης και $x_1, x_2, \dots, x_n$ το εξαρτημένο διάνυσμα των features. Χρησιμοποιώντας την αφελή υπόθεση ανεξαρτησίας:

$$P(x_i|y, x_1, \dots, x_{i-1}, x_{i+1}, x_n) = P(x_i|y)$$

για κάθε i, έχουμε:

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, x_2, \dots, x_n)}.$$

Αφού η πιθανότητα $P(x_1, x_2, \dots, x_n)$ είναι σταθερή για δοσμένη είσοδο, μπορούμε να χρησιμοποιήσουμε τον παρακάτω κανόνα ταξινόμησης:

$$P(y|x_1, x_2, \dots, x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y) \Rightarrow$$

$$\hat{y} = \arg\max_y P(y)\prod_{i=1}^{n} P(x_i|y)$$

Οι διαφορετικοί naive Bayes ταξινομητές διαφέρουν κυρίως στην υπόθεσή τους όσον αφορά την κατανομή της πιθανότητας $P(x_i|y)$. Στον Gaussian Naive Bayes Classifier γίνεται η εξής υπόθεση [27]:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_y^2}\right),$$

όπου μ_y η μέση τιμή και σ_y^2 η διακύμανση και δίνονται από τις σχέσεις [28]:

$$\mu_{Xi|C=c} = \frac{1}{N_c}\sum_{i=1}^{N_c} x_i \ \ και$$

$$\sigma^2_{Xi|C=c} = \frac{1}{N_c}\sum_{i=1}^{N_c} x_i^2 - \mu^2,$$

όπου N_c είναι ο αριθμός των παραδειγμάτων όπου $C = c$ και N είναι ο συνολικός αριθμός των παραδειγμάτων που χρησιμοποιήθηκαν για την εκπαίδευση. Ο υπολογισμός της πιθανότητας $P(C = c)$ για τις κλάσεις είναι απλός, χρησιμοποιώντας σχετικές συχνότητες:

$$P(C = c) = \frac{N_c}{N}.$$

Παρά όλες τις φαινομενικά υπεραπλουστευμένες υποθέσεις, οι naive Bayes ταξινομητές έχουν δουλέψει αρκετά καλά σε πολλές εφαρμογές, με την πιο διαδεδομένη να είναι αυτή της ταξινόμησης αρχείων και φιλτάρισμα της ανεπιθύμητης αλληλογραφίας. Απαιτούν λίγα δεδομένα εκπαίδευσης για την εκτίμηση των απαραίτητων παραμέτρων.

Οι ταξινομητές και μαθητευόμενα συστήματα βασισμένα στον naive Bayes αλγόριθμο μπορούν να παρουσιάσουν σημαντική ταχύτητα σε σύγκριση με πιο προηγμένες μεθόδους. Η απόζευξη των υπό συνθήκη κατανομών των features σημαίνει ότι κάθε κατανομή μπορεί να εκτιμηθεί ανεξάρτητα ως μια κατανομή μίας διάστασης. Έτσι, αποφεύγονται τα προβλήματα της διαστατικότητας.

Συνολικά, τα πλεονεκτήματα των naive Bayes αλγορίθμων περιλαμβάνουν τα παρακάτω [29]:

- Υπολογιστική αποδοτικότητα: ο χρόνος εκπαίδευσης είναι γραμμικός όσον αφορά και τον αριθμό των παραδειγμάτων εκπαίδευσης και τον αριθμό των features, και ο χρόνος ταξινόμησης είναι γραμμικός όσον αφορά τον αριθμό των features και δεν επηρεάζεται από το συνολικό αριθμό των παραδειγμάτων που χρησιμοποιήθηκαν κατά την εκπαίδευση.
- Χαμηλή διακύμανση: επειδή δεν εφαρμόζει αναζήτηση, αλλά αυτό έχει το κόστος του υψηλού bias
- Σταδιακή εκπαίδευση: λειτουργεί χρησιμοποιώντας εκτιμήσεις πιθανοτήτων χαμηλών τάξεων που παράγονται από τα δεδομένα εκπαίδευσης. Αυτές μπορούν να ενημερωθούν εύκολα με την είσοδο νέων δεδομένων εκπαίδευσης.
- Απευθείας πρόβλεψη εκ των υστέρων πιθανοτήτων.
- Ευρωστία στο θόρυβο: ο αλγόριθμος χρησιμοποιεί πάντα όλα τα χαρακτηριστικά για όλες τις προβλέψεις που εκτελεί και, συνεπώς, είναι σχετικά αναίσθητος στο θόρυβο των παραδειγμάτων που ταξινομούνται. Επειδή χρησιμοποιεί πιθανότητες, είναι, επίσης, σχετικά αναίσθητος στο θόρυβο των δεδομένων εκπαίδευσης.
- Ευρωστία στις ελλείπουσες τιμές: επειδή ο naive Bayes χρησιμοποιεί πάντα όλα τα χαρακτηριστικά για όλες τις προβλέψεις, αν η τιμή ενός χαρακτηριστικού λείπει, πληροφορίες από τα υπόλοιπα χαρακτηριστικά χρησιμοποιούνται, με αποτέλεσμα μια «κομψή» πτώση στην επίδοση του συστήματος. Είναι, επίσης, σχετικά αναίσθητο στις ελλείπουσες τιμές λόγω του πιθανολογικού πλαισίου.

Ωστόσο, παρ' όλο που ο naive Bayes είναι ένας σχετικά καλός ταξινομητής, είναι γνωστός ως εκτιμητής (estimator) χαμηλής απόδοσης, οπότε οι πιθανότητες που υπολογίζει ως έξοδο δεν είναι έμπιστες.

3.6 Δέντρα αποφάσεων

Η εκμάθηση δέντρων αποφάσεων για ταξινόμηση είναι μια μέθοδος προσέγγισης μιας συνάρτησης-στόχου διακριτών τιμών, κατά την οποία η συνάρτηση αντιπροσωπεύεται με ένα δέντρο αποφάσεων. Αυτές οι μέθοδοι εκμάθησης είναι μεταξύ των πιο διαδεδομένων αλγορίθμων επαγωγικού συμπεράσματος και έχουν εφαρμοστεί επιτυχημένα σε ένα μεγάλο εύρος εργασιών, από διάγνωση ιατρικών περιπτώσεων έως και εκτίμηση πιστωτικού κινδύνου υποψηφίων για δάνεια [13].

Για την επεξήγηση της λειτουργίας των δέντρων αποφάσεων απαιτείται ο ορισμός των παρακάτω εννοιών:

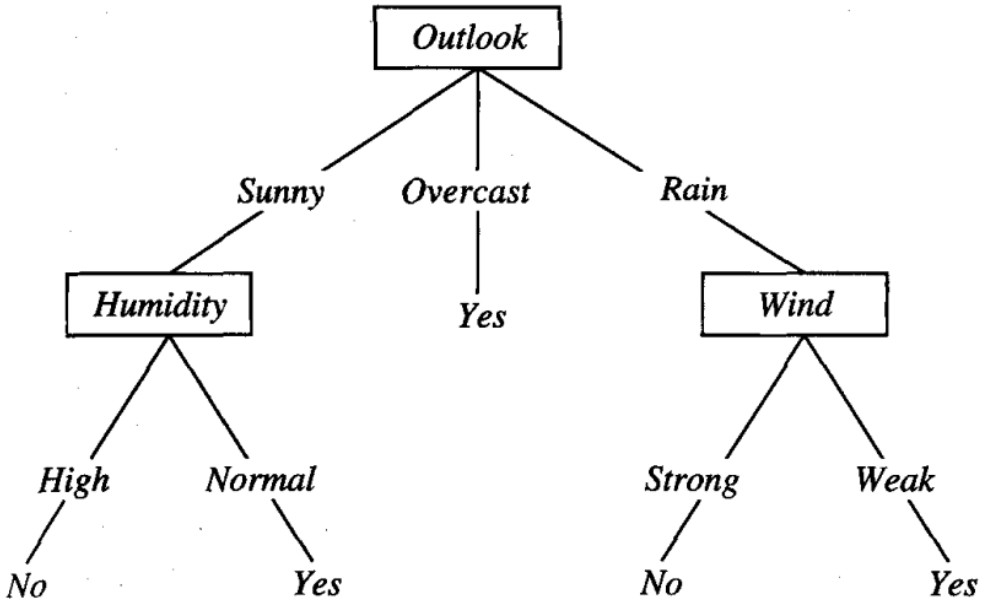
- Οι κόμβοι ενός δέντρου μπορούν να χωριστούν σε επίπεδα, με την έννοια ότι κόμβοι που απέχουν το ίδιο από τη ρίζα βρίσκονται στο ίδιο επίπεδο.
- Ως ρίζα ορίζουμε τον κόμβο του δέντρου ο οποίος βρίσκεται στο ανώτατο επίπεδο του δέντρου και δεν έχει πατέρα, δηλαδή κανένας άλλος κόμβος δεν οδηγεί σε αυτόν.
- Ο πατέρας ενός κόμβου είναι ο αμέσως προηγούμενος κόμβος της διαδρομής από τη ρίζα σε αυτόν. Κάθε κόμβος εκτός από τη ρίζα έχει ένα μοναδικό γονιό.
- Το παιδί ενός κόμβου είναι ένα κόμβος του οποίου αυτός είναι πατέρας.
- Ένας κόμβος χωρίς παιδιά ονομάζεται φύλλο.

Τα δέντρα αποφάσεων ταξινομούν διατρέχοντας το δέντρο από τη ρίζα έως κάποιο φύλλο, το οποίο παρέχει την κατηγοριοποίηση για το συγκεκριμένο παράδειγμα. Κάθε κόμβος του δέντρου προσδιορίζει ένα «τεστ» κάποιου χαρακτηριστικού του παραδείγματος, και κάθε κλαδί που κατεβαίνει από τον κόμβο αυτό αντιστοιχεί σε μια πιθανή τιμή του χαρακτηριστικού. Συνεπώς, για την ταξινόμηση ενός παραδείγματος ξεκινάμε από τη ρίζα του δέντρου, ελέγχοντας το χαρακτηριστικό του παραδείγματος που ορίζεται από αυτόν τον κόμβο, και πηγαίνοντας στον επόμενο κόμβο ανάλογα με την τιμή του χαρακτηριστικού στο συγκεκριμένο παράδειγμα. Η διαδικασία αυτή επαναλαμβάνεται για όλα τα υποδέντρα, μέχρι να καταλήξουμε σε φύλλο, όπου και λαμβάνεται η απάντηση για την ταξινόμηση.

Γενικά, τα δέντρα απόφασης αντιπροσωπεύουν μια διάζευξη συζεύξεων περιορισμών των τιμών των χαρακτηριστικών των παραδειγμάτων. Κάθε διαδρομή από τη ρίζα έως κάποιο φύλλο αντιστοιχεί σε μια σύζευξη ελέγχων των τιμών των χαρακτηριστικών, και το ίδιο το δέντρο είναι μια διάζευξη των συζεύξεων αυτών. Για παράδειγμα, το δέντρο που παρουσιάζεται στο σχήμα 14 αντιστοιχεί στην παρακάτω έκφραση [13]:

$$(Outlook = Sunny \wedge Humidity = Normal) \vee (Outlook = Overcast)$$

$$\vee (Outlook = Rain \wedge Wind = Weak)$$



Σχήμα 14: Παράδειγμα δέντρου απόφασης για τη διεξαγωγή παιχνιδιού τένις, από [13]

Μερικά πλεονεκτήματα των δέντρων απόφασης είναι [30]:

- Κατανοούνται και ερμηνεύονται εύκολα χάρη στην οπτικοποίησή τους.
- Απαιτούν μικρή προεπεξεργασία δεδομένων. Άλλες τεχνικές συχνά απαιτούν κανονικοποίηση των δεδομένων, δημιουργία dummy μεταβλητών, και αφαίρεση ελλειπουσών τιμών.
- Το κόστος της χρήσης ενός δέντρου για την πρόβλεψη δεδομένων είναι λογαριθμικό σε σχέση με τον αριθμό των παραδειγμάτων που χρησιμοποιήθηκαν για την εκπαίδευση του δέντρου.
- Μπορούν να διαχειριστούν και αριθμητικά και κατηγορικά δεδομένα.
- Μπορούν να διαχειριστούν προβλήματα πολλαπλών εξόδων.
- Χρησιμοποιεί το white-box μοντέλο. Αν μια κατάσταση είναι παρατηρήσιμη σε ένα μοντέλο, τότε η εξήγηση για την κατάσταση αυτή δίνεται εύκολα από Boolean λογική. Αντιθέτως, σε ένα black-box μοντέλο, τα αποτελέσματα μπορεί να ερμηνεύονται πιο δύσκολα.
- Είναι εφικτή η εκτίμηση ενός μοντέλου με τη χρήση στατιστικής ανάλυσης. Αυτό επιτρέπει την εκτίμηση και της αξιοπιστίας του μοντέλου.

Τα μειονεκτήματα των δέντρων αποφάσεων περιλαμβάνουν [30]:

- Μερικές φορές μπορεί να δημιουργηθεί ένα πολύ περίπλοκο δέντρο που δε γενικεύει καλά τα δεδομένα. Αυτό αποκαλείται overfitting. Μηχανισμοί όπως το «κλάδεμα» (pruning), ο ορισμός του ελάχιστου αριθμού παραδειγμάτων που απαιτούνται σε ένα φύλλο, ή ο ορισμός του μέγιστου βάθους του δέντρου είναι απαραίτητοι για την αποφυγή του προβλήματος.
- Τα δέντρα αποφάσεων μπορεί να είναι ασταθής εξαιτίας μικρών μεταβολών στα δεδομένα, το οποίο συνεπάγεται τη δημιουργία ενός τελείως διαφορετικού

δέντρου. Αυτό το πρόβλημα περιορίζεται με τη χρήση δέντρων απόφασης εντός ενός ensemble, όπως θα αναλυθεί στη συνέχεια.

- Οι προβλέψεις των δέντρων αποφάσεων δεν είναι ούτε ομαλές ούτε συνεχείς, αλλά τμηματικές σταθερές προσεγγίσεις. Για το λόγο αυτό δεν είναι καλά στην προεκβολή συμπερασμάτων.
- Το πρόβλημα εκμάθησης ενός βέλτιστου δέντρου απόφασης θεωρείται NP-πλήρες ακόμα και για απλές περιπτώσεις. Συνεπώς, οι αλγόριθμοι είναι βασισμένοι σε ευριστικούς αλγορίθμους, όπως είναι ο άπληστος (greedy) όπου λαμβάνονται τοπικά βέλτιστες αποφάσεις σε κάθε κόμβο. Αυτοί οι αλγόριθμοι δε μπορούν να εγγυηθούν το συνολικά βέλτιστο δέντρο απόφασης. Το πρόβλημα αυτό μπορεί να περιοριστεί εκπαιδεύοντας πολλά δέντρα εντός ενός ensemble, όπου τα χαρακτηριστικά και τα παραδείγματα επιλέγονται τυχαία με αντικατάσταση.
- Υπάρχουν έννοιες που είναι δύσκολες να μάθει ένα δέντρο απόφασης γιατί δε μπορούν να εκφραστούν απλά, όπως XOR και προβλήματα πολυπλεκτών.
- Τα δέντρα απόφασης δημιουργούν δέντρα με bias, αν μερικές κλάσεις κυριαρχούν. Προτείνεται, λοιπόν, να προηγηθεί ισορρόπηση του dataset πριν την εκπαίδευση του μοντέλου.

3.6.1 Βασικός αλγόριθμος εκμάθησης

Οι περισσότεροι αλγόριθμοι που έχουν αναπτυχθεί για την εκμάθηση δέντρων αποφάσεων είναι παραλλαγές ενός βασικού αλγορίθμου που επιστρατεύει μια άπληστη αναζήτηση από πάνω έως κάτω δια μέσου του χώρου όλων των πιθανών δέντρων. Σε αυτήν την προσέγγιση έχει βασιστεί ο αλγόριθμος ID3 και ο διάδοχός του C4.5. Στην παράγραφο αυτή θα αναλύσουμε το βασικό αλγόριθμο εκμάθησης δέντρων αποφάσεων, που αντιστοιχεί προσεγγιστικά στον αλγόριθμο ID3 [13].

Ο βασικός αλγόριθμος, ID3, εκπαιδεύει δέντρα αποφάσεων κατασκευάζοντάς τα από πάνω προς τα κάτω, ξεκινώντας με την ερώτηση «ποιο χαρακτηριστικό πρέπει να δοκιμαστεί στη ρίζα του δέντρου;». Για να απαντηθεί η ερώτηση αυτή, κάθε χαρακτηριστικό εκτιμάται χρησιμοποιώντας στατιστική ανάλυση για να καθοριστεί πόσο καλά ταξινομεί μόνο του τα παραδείγματα εκπαίδευσης. Το καλύτερο χαρακτηριστικό επιλέγεται ως ο πρώτος έλεγχος του δέντρου, στη ρίζα του. Ύστερα, ένας απόγονος της ρίζας του δέντρου δημιουργείται για κάθε πιθανή τιμή του κατηγορικού χαρακτηριστικού. Στην περίπτωση των αριθμητικών χαρακτηριστικών, ο έλεγχος που δημιουργείται είναι Boolean λογικής (true ή false) ανάλογα με το αν η τιμή του χαρακτηριστικού είναι πάνω ή κάτω από ένα threshold (που συνήθως ορίζεται ως κάποια μέση τιμή). Τα παραδείγματα εκπαίδευσης ταξινομούνται βάσει του μέχρι τώρα δέντρου και η διαδικασία αυτή επαναλαμβάνεται χρησιμοποιώντας τα παραδείγματα κάθε φορά για την επιλογή του καλύτερου χαρακτηριστικού στο συγκεκριμένο σημείο του δέντρου. Έτσι, έχουμε μια άπληστη αναζήτηση για ένα αποδεκτό δέντρο απόφασης, στην οποία ο αλγόριθμος δεν οπισθοπορεί για να επανεξετάσει προηγούμενες επιλογές.

Η κύρια επιλογή στον αλγόριθμο ID3 είναι η επιλογή ποιου χαρακτηριστικού να ελέγχεται σε κάθε κόμβο του δέντρου. Ιδανικά, επιλέγεται το χαρακτηριστικό που είναι το πιο χρήσιμο για την ταξινόμηση παραδειγμάτων. Για να ληφθεί αυτή η απόφαση, πρέπει να οριστεί ένα κατάλληλο μέτρο για την αξία κάθε χαρακτηριστικού. Ορίζουμε μια στατιστική ιδιότητα, το κέρδος πληροφορίας (information gain), που μετρά πόσο καλά ένα συγκεκριμένο χαρακτηριστικό διαχωρίζει τα παραδείγματα εκπαίδευσης βάσει της επιθυμητής ταξινόμησής τους. Ο αλγόριθμος ID3 χρησιμοποιεί αυτήν την information gain μετρική για να επιλέξει μεταξύ των υποψήφιων χαρακτηριστικών σε κάθε βήμα κατασκευής του δέντρου [13].

Για τον ακριβή ορισμό του information gain, αρχικά ορίζουμε ένα μέτρο που χρησιμοποιείται συχνά στη θεωρία πληροφορίας, την εντροπία, που χαρακτηρίζει τη γνησιότητα (purity) ενός αυθαίρετου συνόλου παραδειγμάτων. Δεδομένου συνόλου S, που περιέχει θετικά και αρνητικά παραδείγματα (1 ή 0 αντίστοιχα, churned και μη στη δική μας περίπτωση), η εντροπία του S αναλογικά με αυτή τη Boolean ταξινόμηση είναι

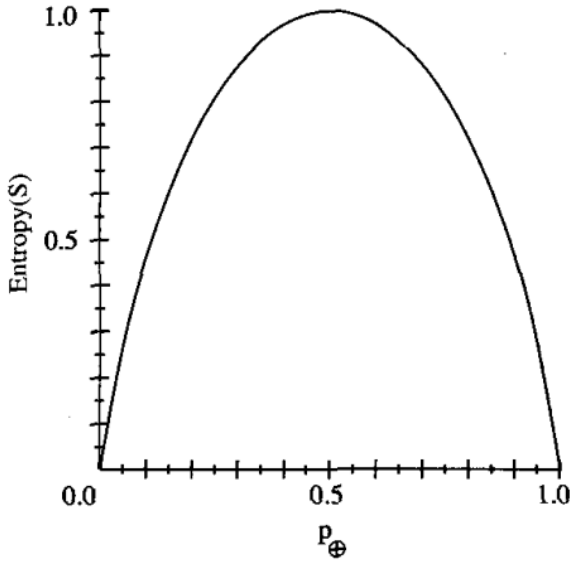
$$Entropy(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-,$$

όπου p_+ και p_- είναι το ποσοστό των θετικών και αρνητικών παραδειγμάτων εντός του S, αντίστοιχα. Σε όλους τους υπολογισμούς που περιλαμβάνουν εντροπία, ορίζουμε

$$0 \log 0 = 0.$$

Έτσι, η εντροπία είναι μηδενική αν όλα τα μέλη του S ανήκουν στην ίδια κλάση, και ίση με 1 αν περιέχει ίσο αριθμό θετικών και αρνητικών παραδειγμάτων. Αν το σύνολο των παραδειγμάτων περιέχει άνισο αριθμό θετικών και αρνητικών δειγμάτων η εντροπία παίρνει τιμές στο διάστημα $(0, 1)$.

Μια ερμηνεία της εντροπίας από τη θεωρία πληροφορίας είναι ότι προσδιορίζει τον ελάχιστο αριθμό από bits (δυαδικά ψηφία) πληροφορίας που απαιτούνται για την κωδικοποίηση της ταξινόμησης ενός αυθαίρετου μέλους του S, δηλαδή ενός μέλους του S που επιλέχθηκε τυχαία με ομοιόμορφη πιθανότητα. Για παράδειγμα αν $p_+ = 1$, ο αποδέκτης γνωρίζει ότι το τυχαίο παράδειγμα είναι θετικό, οπότε δε χρειάζεται να αποσταλεί κάποιο μήνυμα και η εντροπία είναι μηδενική. Από την άλλη, αν $p_+ = 0.5$, απαιτείται ένα bit για να υποδείξει αν το επιλεγμένο παράδειγμα είναι θετικό ή αρνητικό. Αν $p_+ = 0.8$, τότε ένα σύνολο μηνυμάτων μπορούν να κωδικοποιηθούν χρησιμοποιώντας κατά μέσο όρο λιγότερα από ένα bit ανά μήνυμα, αναθέτοντας σύντομα μηνύματα στο σύνολο των θετικών παραδειγμάτων και μεγαλύτερης διαρκείας σε αρνητικά παραδείγματα, που είναι λιγότερο πιθανά [13].



Σχήμα 15: Γραφική παράσταση της συνάρτησης εντροπίας, από [13]

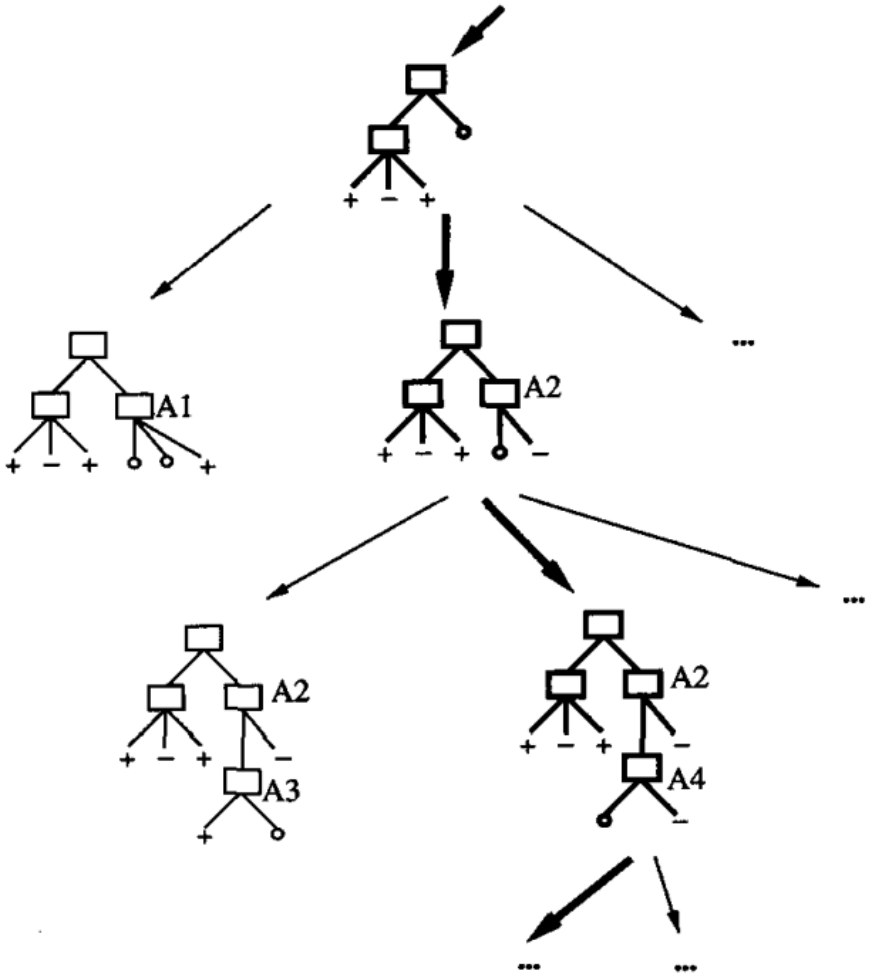
Έχοντας ορίσει την εντροπία ως μέτρο του impurity ενός συνόλου παραδειγμάτων εκπαίδευσης, μπορούμε τώρα να ορίσουμε ένα μέτρο της αποτελεσματικότητας ενός χαρακτηριστικού για την ταξινόμηση των δεδομένων. Το μέτρο που θα χρησιμοποιήσουμε, το information gain, είναι απλά η αναμενόμενη μείωση της εντροπίας που προκαλείται από το διαχωρισμό των παραδειγμάτων βάσει αυτού του χαρακτηριστικού. Πιο συγκεκριμένα, το information gain $G(S, A)$ ενός χαρακτηριστικού A σε σχέση με ένα σύνολο παραδειγμάτων S ορίζεται ως

$$G(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v),$$

όπου $Values(A)$ είναι το σύνολο όλων των πιθανών τιμών του χαρακτηριστικού A, και S_v είναι το υποσύνολο του S για το οποίο το χαρακτηριστικό A έχει τιμή v. Ο πρώτος όρος είναι η εντροπία του αρχικού συνόλου S και ο δεύτερος είναι η αναμενόμενη τιμή της εντροπίας μετά το διαχωρισμό του S χρησιμοποιώντας το χαρακτηριστικό A. Η αναμενόμενη εντροπία είναι απλά το άθροισμα των εντροπιών κάθε υποσυνόλου S_v, σταθμισμένο από το κλάσμα των παραδειγμάτων $\frac{|S_v|}{|S|}$ που ανήκουν στο S_v. Συνεπώς, το $Gain(S, A)$ είναι η αναμενόμενη μείωση στην εντροπία που προκαλείται γνωρίζοντας την τιμή του χαρακτηριστικού A. Βάσει της ερμηνείας για την εντροπία που αναφέρθηκε προηγουμένως, το $Gain(S, A)$ είναι ο αριθμός των bits που εξοικονομούνται με την κωδικοποίηση της τιμής ενός αυθαίρετου μέλους του S, γνωρίζοντας την τιμή του χαρακτηριστικού A [13].

Όπως συμβαίνει με άλλες επαγωγικές μεθόδους εκπαίδευσης, ο αλγόριθμος ID3 μπορεί να χαρακτηριστεί ότι αναζητά ένα χώρο υπόθεσης για κάποιο που ταιριάζει με τα παραδείγματα εκπαίδευσης. Ο χώρος υπόθεσης που εξερευνά ο ID3 είναι το σύνολο όλων των πιθανών δέντρων απόφασης. Εκτελείται ένας αλγόριθμος αναρρίχησης, ξεκινώντας με το άδειο δέντρο, και ύστερα λαμβάνοντας υπόψη προοδευτικά πιο πολύπλοκες υποθέσεις στην αναζήτηση ενός δέντρου απόφασης που να ταξινομεί ορθά

τα δεδομένα εκπαίδευσης. Η ευριστική συνάρτηση που καθοδηγεί τον αλγόριθμο είναι το information gain, όπως αναλύθηκε προηγουμένως.



Σχήμα 16: Σχηματική αναπαράσταση αναζήτησης του αλγορίθμου ID3, από [13]

3.6.2 Ensemble τεχνικές

Η βασική ιδέα της έρευνας στον τομέα των ensemble τεχνικών είναι ότι σε πολλές περιπτώσεις μια επιτροπή ταξινομητών θα παραγάγουν καλύτερα αποτελέσματα από ένα μοναδικό ταξινομητή όσον αφορά τη σταθερότητα και την ακρίβεια. Αυτό ισχύει περισσότερο σε περιπτώσεις κατά τις οποίες τα δομικά στοιχεία των ταξινομητών είναι ασταθή, όπως και είναι για τα νευρωνικά δίκτυα και τα δέντρα αποφάσεων. Αν και η χρήση των ensembles στη μηχανική μάθηση είναι σχετικά νέα, η ιδέα ότι η άθροιση γνωμών μιας επιτροπής ειδικών θα αυξήσει την ακρίβεια δεν είναι [15].

Το θεώρημα της κριτικής επιτροπής του Condorcet δηλώνει ότι αν κάθε ψηφοφόρος έχει πιθανότητα p να είναι σωστός και η πιθανότητα η πλειοψηφία των ψηφοφόρων να είναι σωστοί είναι P, τότε η σχέση $p > 0.5$ συνεπάγεται ότι $P > p$. Όσο ο αριθμός των ψηφοφόρων προσεγγίζει το άπειρο, η πιθανότητα P προσεγγίζει τη μονάδα και ισχύει για κάθε $p > 0.5$. Γνωρίζουμε ότι η πιθανότητα P θα είναι μεγαλύτερη της p μόνο εάν υπάρχει «πολυπολιτισμικότητα» (diversity) στην ομάδα των

ψηφοφόρων και γνωρίζουμε, επίσης, ότι η πιθανότητα να είναι σωστή η επιτροπή αυξάνεται με την αύξηση του ensemble αν αυξηθεί και το diversity του. Συνήθως, το diversity και η ακρίβεια του ensemble σταθεροποιείται για κάποιο μέγεθος μεταξύ 10 και 50 μελών.

Στον τομέα της μηχανικής μάθησης είναι γνωστό ότι οι ensemble τεχνικές θα αυξήσουν την επίδοση των ασταθών μαθητευόμενων συστημάτων, τα οποία είναι συστήματα στα οποία μικρές αλλαγές στα δεδομένα εκπαίδευσης μπορούν να παραγάγουν πολύ διαφορετικά μοντέλα και, συνεπώς, διαφορετικές προβλέψεις. Έτσι, ένας τρόπος για την ύπαρξη diversity είναι η εκπαίδευση των μοντέλων σε διαφορετικά υποσύνολα των δεδομένων εκπαίδευσης. Αυτή η προσέγγιση έχει εφαρμοστεί επιτυχημένα για δύσκολες εργασίες ταξινόμησης, καθώς και regression, σε νευρωνικά δίκτυα και δέντρα αποφάσεων. Οι ensemble τεχνικές βελτιώνουν, επίσης, και πιο σταθερά εκπαιδευόμενα δίκτυα όπως ταξινομητές k nearest neighbors και naive Bayes, ωστόσο τα συγκεκριμένα είναι εύρωστα στις αλλαγές των δεδομένων εκπαίδευσης οπότε πρέπει να επιστρατευθούν άλλες πηγές diversity, για παράδειγμα χρησιμοποιώντας υποσύνολο των χαρακτηριστικών.

Έχει δειχθεί ότι η μείωση στο σφάλμα λόγω ενός ensemble είναι ευθέως ανάλογη με το diversity των προβλέψεων των μελών του ensemble, μετρημένο με διακύμανση. Είναι δύσκολο να δειχθεί ότι αντίστοιχη ευθέα σχέση για τις εργασίες ταξινόμησης, αλλά είναι ξεκάθαρο ότι το πλεονέκτημα λόγω του ensemble εξαρτάται από το diversity των μελών του. Μπορούμε να πούμε ότι αν τα μέλη του ensemble είναι πιο πιθανό κατά μέσο όρο να εκτελέσουν σωστή πρόβλεψη και όταν εκτελούν λανθασμένη, διαφέρουν σε διαφορετικά σημεία μεταξύ τους, τότε οι αποφάσεις της πλειοψηφίας είναι πιο πιθανό να είναι σωστές από αυτές των μεμονωμένων μελών [15].

3.6.2.1 Bagging

Ο απλούστερος τρόπος για τη δημιουργία ενός ensemble ασταθών ταξινομητών όπως δέντρων αποφάσεων είναι η χρήση bootstrap aggregation, που συχνά αποκαλείται και bagging. Έχοντας ένα σύνολο δεδομένων εκπαίδευσης D και ένα δείγμα ερωτημάτων q τα βήματα είναι τα εξής [15]:

1. Για ένα ensemble S μελών, παράγονται S σύνολα εκπαίδευσης T_i, όπου $i = 1, 2, \dots, S$, από το D με bootstrap δειγματοληψία, δηλαδή δειγματοληψία με αντικατάσταση. Συχνά ισχύει $|D_i| = |D|$.
2. Για κάθε D_i, έστω D_{v_i} το σύνολο των παραδειγμάτων εκπαίδευσης που δεν επιλέχθηκαν στο D_i. Το σύνολο αυτό μπορεί να χρησιμοποιηθεί ως validation για τον έλεγχο overfitting του μέλους που εκπαιδεύτηκε με D_i. Το D_{v_i} συχνά αποκαλείται "out-of-bag" (OOB) δεδομένα.
3. Εκπαιδεύονται S ταξινομητές $f_i(, D_i)$ χρησιμοποιώντας D_i σύνολα εκπαίδευσης. Τα validation σετ D_{v_i} μπορούν να χρησιμοποιηθούν για έλεγχο overfitting.
4. Παράγονται S προβλέψεις για q χρησιμοποιώντας τους S ταξινομητές $f_i(, D_i)$.

5. Αθροίζονται αυτές τις S προβλέψεις $f_i(q, D_i)$ για να λάβουμε μια μοναδική πρόβλεψη q χρησιμοποιώντας κάποια συνάρτηση συγκεντρωτικών αποτελεσμάτων.

Ο τύπος για αυτό το ensemble θα είναι:

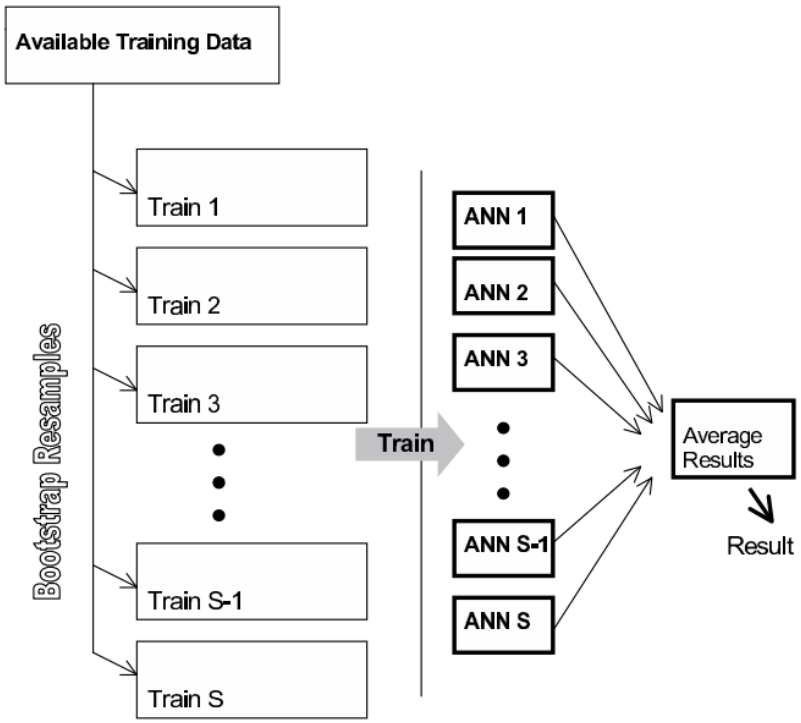
$$f_E(q, D) = F\big(f_1(q, D_1), f_2(q, D_2), \dots, f_S(q, D_S)\big),$$

όπου $F()$ η συνάρτηση συγκεντρωτικών αποτελεσμάτων. Η απλούστερη προσέγγιση για τη συνάρτηση αυτή είναι να πάρουμε τη μέση τιμή ή τη σταθμισμένη μέση τιμή:

$$f_E(q, D) = \sum_{i=1}^{S} w_i \times f_i(q, D_i),$$

όπου $\sum_{i=1}^{S} w_i = 1$. Δεδομένου ότι υπάρχει diversity στο ensemble (και η bootstrap δειγματοληψία πρέπει να το πετύχει αυτό), οι προβλέψεις του ensemble $f_E(q, D)$ θα είναι πιο ακριβείς από τις προβλέψεις των μελών του ensemble $f_i(q, D_i)$ [15].

Είναι σημαντικό να σημειωθεί ότι ο αλγόριθμος bagging, δηλαδή sub-sampling (υποδειγματοληψία) των δεδομένων εκπαίδευσης, είναι η μόνη πηγή diversity αν ο ταξινομητής είναι ασταθής. Όπως αναφέρθηκε και προηγουμένως, για σταθερούς ταξινομητές, όπως ο k nearest neighbors και ο naive Bayes, η τεχνική αυτή δε θα παραγάγει diversity στα μέλη του ensemble. Μια εναλλακτική σε αυτήν την περίπτωση είναι το sub-sample των χαρακτηριστικών, αντί των παραδειγμάτων, το οποίο μπορεί να είναι πολύ αποτελεσματικό όταν τα δεδομένα περιγράφονται από μεγάλο αριθμό χαρακτηριστικών και υπάρχει «πλεονασμός» σε αυτήν την αναπαράσταση.



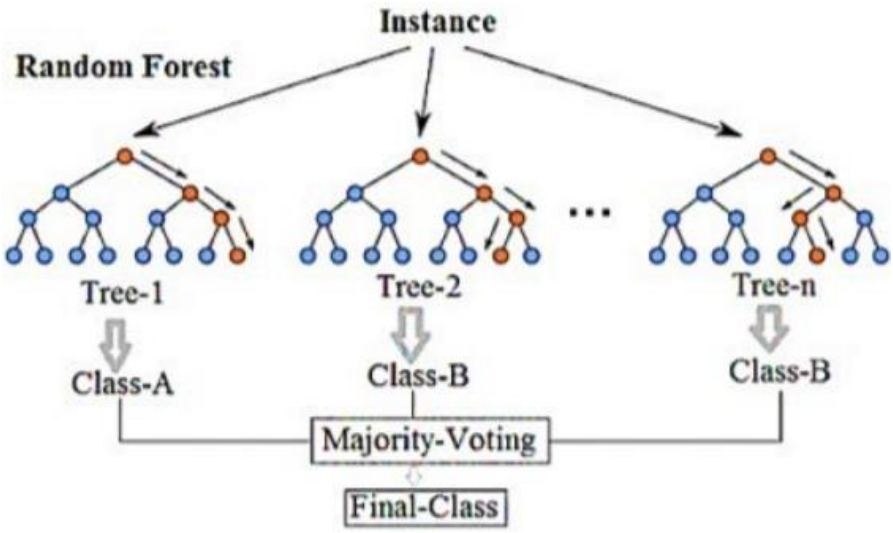
Σχήμα 17: Σχηματική αναπαράσταση του αλγορίθμου bagging, από [15]

Για τα ensembles που χρησιμοποιούνται για regression, υπάρχει σχέση που υποδεικνύει ότι η μείωση του σφάλματος λόγω του ensemble είναι ευθέως ανάλογη με το diversity στις προβλέψεις των μελών. Δυστυχώς, δε φαίνεται να υπάρχει αντίστοιχος τρόπος ποσοτικοποίησης του diversity για ταξινόμηση που να έχει την ίδια ευθέα σχέση με τη μείωση του σφάλματος λόγω του ensemble. Παρ' όλα αυτά, χρησιμοποιούνται διάφορες μετρικές για την εκτίμηση του diversity και της ικανότητάς του να μειώσει το σφάλμα σε ένα ensemble. Μερικές είναι οι: plain disagreement, fail/non-fail disagreement, εντροπία, ασάφεια (ambiguity). Η εντροπία και η ασάφεια ποσοτικοποιούν το diversity του ensemble ως σύνολο. Οι μετρικές των disagreements ποσοτικοποιούν τη διαφωνία μεταξύ ενός ζεύγους ταξινομητών, και η μετρική του diversity προκύπτει ως ο μέσος όρος των διαφωνιών όλων των ζευγών των ταξινομητών του ensemble [15].

3.6.2.2 Random Forests

Στην προηγούμενη παράγραφο αναφέρθηκε ότι οι δύο πιο διαδεδομένες στρατηγικές για την εισαγωγή diversity σε ένα ensemble είναι το sub-sampling των παραδειγμάτων (bagging) και των χαρακτηριστικών (feature selection). Ο Leo Breiman χρησιμοποίησε και τις δύο αυτές ιδέες για τη random forest στρατηγική που ανέπτυξε για τη δημιουργία ensembles [31]. Όπως υποδηλώνει το όνομα, ένα random forest είναι ένα ensemble δέντρων αποφάσεων. Οι δύο πηγές diversity είναι οι ακόλουθες [15]:

1. Όπως στο bagging, για κάθε μέλος του ensemble το σύνολο δεδομένων εκπαίδευσης D δειγματοληπτείται με αντικατάσταση για την παραγωγή ενός συνόλου εκπαίδευσης μεγέθους $|D|$.
2. F είναι το σύνολο των χαρακτηριστικών που περιγράφουν τα δεδομένα, ενώ το $m \ll |F|$ επιλέγεται ως ο αριθμός των χαρακτηριστικών που χρησιμοποιούνται στη διαδικασία του feature selection. Σε κάθε στάδιο, δηλαδή κόμβο, της δημιουργίας του δέντρου m χαρακτηριστικά επιλέγονται τυχαία να είναι υποψήφιοι για το διαχωρισμό αυτού του κόμβου.



Σχήμα 18: Σχηματική αναπαράσταση του αλγορίθμου random forest, από [32]

Για να διασφαλιστεί το diversity μεταξύ των δέντρων που αποτελούν τα συστατικά του ensemble, δεν πραγματοποιείται κάποιο κλάδεμα (pruning), όπως είναι σύνηθες στην κατασκευή δέντρων αποφάσεων. Τα OOB δεδομένα μπορούν να χρησιμοποιηθούν για την εκτίμηση της ακρίβειας γενίκευσης των μελών του ensemble. Στην κατασκευή ενός random forest είναι τυπικό να δημιουργούνται πολύ περισσότερων μελών ensemble σε σχέση με το bagging, αφού υπάρχουν περιπτώσεις που δημιουργούνται 100 ή ακόμα και 1000 δέντρα. Ο «κόπος» που διατίθενται για τη δημιουργία αυτών των δέντρων έχει το πρόσθετο πλεονέκτημα να παρέχει μια ανάλυση των δεδομένων. Τα πλεονεκτήματα αυτά περιλαμβάνουν [15]:

- Εκτίμηση του σφάλματος γενίκευσης. Τα OOB δεδομένα προσφέρουν ευθέως μια εκτίμηση της γενίκευσης του σφάλματος των δέντρων που αποτελούν συστατικά στοιχεία του ensemble. Ωστόσο, μπορούν, επίσης, να χρησιμοποιηθούν για τη λήψη μιας εκτίμησης του σφάλματος γενίκευσης όλου του ensemble συνολικά χωρίς bias. Αφού κάθε παράδειγμα είναι OOB σε περίπου $\frac{1}{3}$ των δέντρων, η πλειοψηφία μπορεί να θεωρηθεί ως η κλάση που θα προέβλεπε το ensemble σε εκείνη την περίπτωση. Το σφάλμα σε αυτές τις αθροιστικές OOB προβλέψεις είναι μια εκτίμηση χωρίς bias του σφάλματος του random forest ως σύνολο.
- Εγγύτητα παραδειγμάτων. Όταν πολλά δέντρα κατασκευάζονται, ένα ενδιαφέρον στατιστικό είναι η συχνότητα με την οποία τα παραδείγματα, και εκπαίδευσης και OOB, βρίσκονται στο ίδιο φύλλο. Κάθε φύλλο σε κάθε δέντρο ελέγχεται και διατηρείται ένας $|D| \times |D|$ πίνακας όπου το κελί (i, j) αυξάνεται κάθε φορά που τα παραδείγματα i και j μοιράζονται το ίδιο φύλλο. Αν ο πίνακας διαιρεθεί με τον αριθμό των δέντρων, έχουμε ένα μέτρο εγγύτητας (proximity) που είναι σε αρμονία με τον αλγόριθμο ταξινόμησης, τον random forest.
- Σημασία μεταβλητών. Η βασική ιδέα για την εκτίμηση της σημασίας μιας μεταβλητής που χρησιμοποιεί random forest είναι να ελέγξουμε την επίδραση των τυχαίων μεταθέσεων των τιμών της μεταβλητής σε OOB παραδείγματα και να τα επαναταξινομήσουμε. Αν το σφάλμα αυξηθεί σημαντικά, τότε η μεταβλητή αυτή είναι σημαντική. Αν το σφάλμα δεν αυξηθεί, τότε η μεταβλητή αυτή δεν είναι χρήσιμη για την ταξινόμηση.

Τα τελευταία χρόνια με την αύξηση της υπολογιστικής ισχύος, υπάρχει μια μετάθεση της έμφασης της έρευνας στο πεδίο της μηχανικής μάθησης. Πλέον, ενδιαφερόμαστε να αξιοποιήσουμε τους εξέχοντες υπολογιστικούς πόρους που είναι διαθέσιμοι, και ο αλγόριθμος random forest είναι μια ιδέα προς αυτήν την κατεύθυνση.

3.6.2.3 Boosting

Η τεχνική του boosting είναι μια πιο προσεκτική προσέγγιση για την κατασκευή ensemble. Αντί να δημιουργούνται όλους τους ταξινομητές με τη μία, όπως συμβαίνει στο bagging, οι ταξινομητές δημιουργούνται σειριακά. Η ιδέα του boosting είναι να χρησιμοποιήσει το σφάλμα μεταξύ της επιθυμητής απόκρισης και της πρόβλεψης των

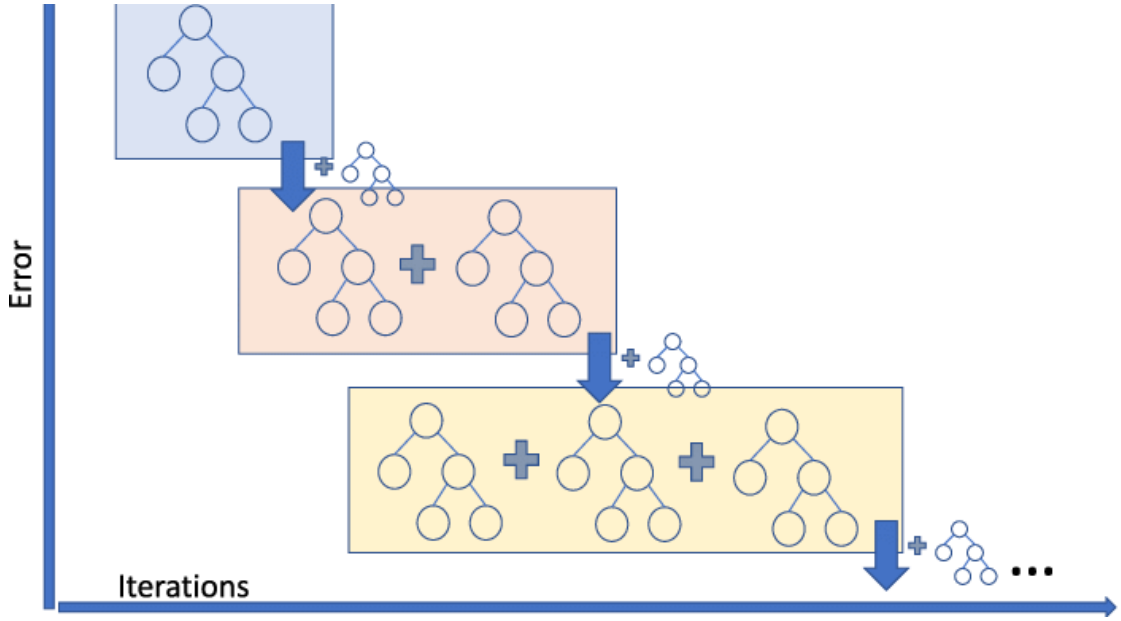
ταξινομητών που έχουν κατασκευαστεί μέχρι εκείνη τη στιγμή για την κατασκευή καλύτερων επόμενων ταξινομητών. Έτσι, μειώνεται και η διακύμανση και το bias. Άρα, στο boosting δημιουργούνται ταξινομητές σειριακά, με τους επακόλουθους ταξινομητές να επικεντρώνονται στα παραδείγματα εκπαίδευσης στα οποία δεν είχαν καλή επίδοση οι προηγούμενοι ταξινομητές. Αυτό επιτυγχάνεται προσαρμόζοντας την κατανομή δειγματοληψίας των δεδομένων εκπαίδευσης. Έχουμε, λοιπόν, τον αλγόριθμο (όπου $(x_1, y_1), (x_2, y_2), \dots, (x_{|D|}, y_{|D|})$ με $x_i \in X, y_i \in Y = \{-1, +1\}$ είναι τα διαθέσιμα δεδομένα εκπαίδευσης) [15]:

1. Αρχικοποίηση της κατανομής δειγματοληψίας $P_1(i) = \frac{1}{|D|}$
2. Για $t = 1, 2, \dots, T$:
   α. Εκπαίδευσε ταξινομητή χρησιμοποιώντας την κατανομή P_t.
   β. Η υπόθεση που αντιπροσωπεύει αυτός ο ταξινομητής είναι $h_t: X \to \{-1, +1\}$.
   γ. Υπολόγισε το σφάλμα του ταξινομητή ως $\varepsilon_t = \sum_{i:h_t(x_i)\neq y_i} P_t(i)$.
   δ. Έστω $\alpha_t = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}$.
   ε. Ενημέρωσε
   $$P_{t+1}(i) = \frac{P_t(i)}{Z_t} \times \begin{cases} e^{-\alpha t} \text{ αν } h_t(x_i) = y_i \\ e^{\alpha t} \text{ αν } h_t(x_i) \neq y_i \end{cases},$$
   όπου Z_t είναι ο παράγοντας κανονικοποίησης που εξασφαλίζει ότι η P_{t+1} είναι κατανομή.
   στ. Συνέχισε την εκπαίδευση νέων ταξινομητών όσο $\varepsilon_t < 0.5$.
3. Αυτό το ensemble ταξινομητών μπορεί να χρησιμοποιηθεί για την παραγωγή ταξινόμησης ως εξής:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$

Όπως είναι φανερό, ο ρόλος που παίζει το σφάλμα ταξινόμησης στον αλγόριθμο αυτό συνεπάγεται ότι αυτή η σύνθεση μπορεί να εφαρμοστεί μόνο σε προβλήματα ταξινόμησης. Ωστόσο, υπάρχουν και επεκτάσεις του boosting για προβλήματα regression.

Γενικά, το σφάλμα γενίκευσης ενός ταξινομητή θα βελτιωθεί με τη δημιουργία ενός ensemble τέτοιων ταξινομητών και συναθροίζοντας τα αποτελέσματά τους. Ακόμα και αν η μείωση είναι περιορισμένη (της τάξης μερικών per cent) και το υπολογιστικό κόστος αυξάνεται κατά μια τάξη μεγέθους, πιθανότατα θα αξίζει αυτή η δημιουργία του ensemble καθώς οι υπολογιστικοί πόροι είναι διαθέσιμοι [15].



Σχήμα 19: Σχηματική αναπαράσταση του αλγορίθμου boosting, από [33]

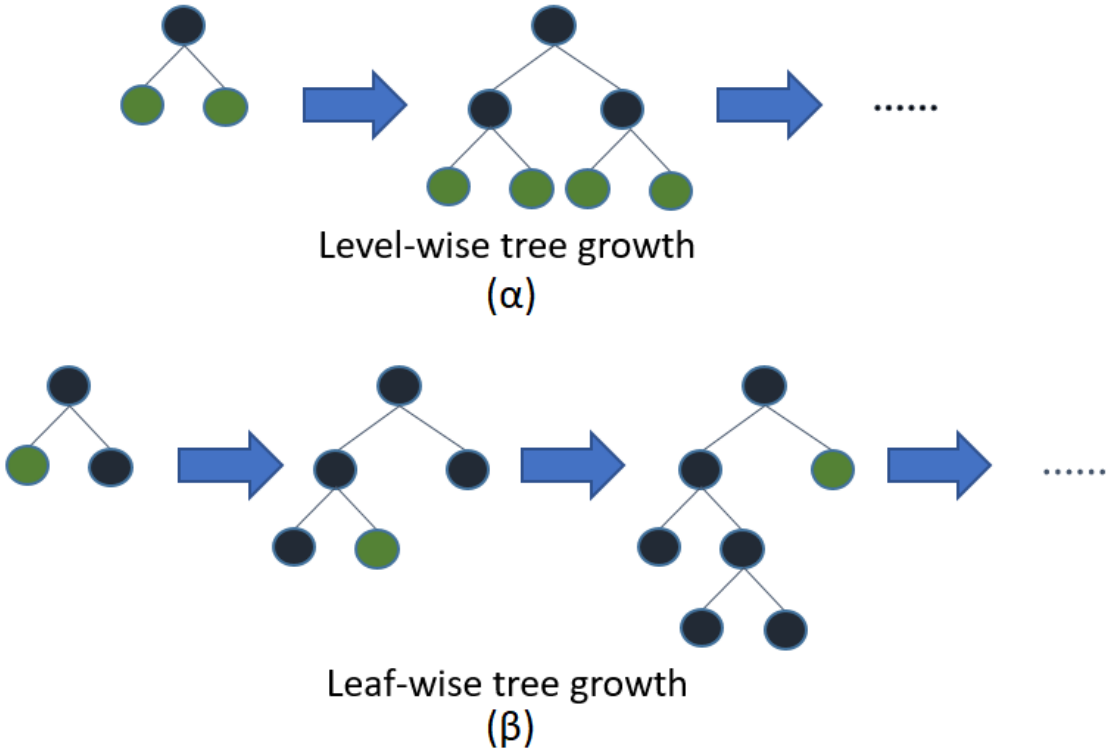
3.6.2.4 Light Gradient Boosting Machine

Ο αλγόριθμος Light Gradient Boosting Machine (LightGBM) [50] αποτελεί επέκταση του αλγορίθμου Gradient Boosting, που αναλύθηκε στην προηγούμενη παράγραφο. Εφαρμόζοντας συγκεκριμένες παραλλαγές στον αλγόριθμο, επιτυγχάνει σημαντική αύξηση στην ταχύτητα και την ακρίβεια, καθώς και μείωση στη χρήση μνήμης.

Οι βελτιστοποιήσεις στην ταχύτητα και τη χρήση μνήμης στηρίζονται στο γεγονός ότι ο LightGBM χρησιμοποιεί αλγορίθμους βασισμένους σε ιστογράμματα, που ομαδοποιούν τα χαρακτηριστικά με συνεχείς τιμές σε διακριτά «καλάθια» (bins). Πλεονεκτήματα των αλγορίθμων που βασίζονται σε ιστογράμματα περιλαμβάνουν [34]:

- Μειωμένο κόστος υπολογισμού του κέρδους κάθε διαχωρισμού
- Χρήση αφαίρεσης ιστογράμματος (histogram subtraction) για επιπλέον επιτάχυνση
- Μειωμένη χρήση μνήμης
- Μειωμένο κόστος επικοινωνίας για εκπαίδευση σε κατανεμημένα συστήματα

Η βελτιστοποίηση στην ακρίβεια προέρχεται από τη μέθοδο που ακολουθεί για τη δημιουργία των δέντρων αποφάσεων. Οι περισσότεροι αλγόριθμοι εκπαίδευσης δέντρων αποφάσεων αναπτύσσουν το δέντρο ανά επίπεδο (level-wise), ενώ ο LightGBM αναπτύσσει τα δέντρα ανά φύλλο (leaf-wise), επιλέγοντας το φύλλο με τη μεγαλύτερη ελαχιστοποίηση στο σφάλμα κάθε φορά. Αυτή η μεθοδολογία μπορεί να οδηγήσει σε overfitting αν τα δεδομένα είναι λίγα σε αριθμό, οπότε ο κατάλληλος ορισμός της παραμέτρου “max_depth” είναι σημαντικός για τον περιορισμό της ανάπτυξης του δέντρου και, συνεπώς, περιορισμό του overfitting.



Σχήμα 20: Σχηματική αναπαράσταση tree growth (α) level-wise (β) leaf-wise, από [34]

Αν αναπτυχθεί όλο το δέντρο, οι μέθοδοι leaf-wise και level-wise θα οδηγήσουν στο ίδιο δέντρο. Η διαφορά τους είναι στη σειρά με την οποία επεκτείνεται το δέντρο. Με δεδομένο ότι συνήθως τα δέντρα δεν αναπτύσσονται μέχρι το τελικό βάθος, η σειρά αυτή έχει σημασία, καθώς η εφαρμογή κριτηρίων διακοπής (όπως η αλλαγή της παραμέτρου "max_depth") και μέθοδοι κλαδέματος (pruning) μπορούν να οδηγήσουν σε τελείως διαφορετικά δέντρα. Επειδή το leaf-wise επιλέγει διαχωρισμούς βασισμένο στη συνολική συμμετοχή τους στο συνολικό σφάλμα του ταξινομητή και όχι απλά στο σφάλμα του κλαδιού εκείνου, συνήθως βρίσκει δέντρα μικρότερου σφάλματος πιο γρήγορα από το level-wise. Με την προσθήκη περισσότερων κόμβων, χωρίς διακοπή ή pruning, θα συγκλίνουν στην επίδοση, αφού τελικά δημιουργούν το ίδιο δέντρο [35].

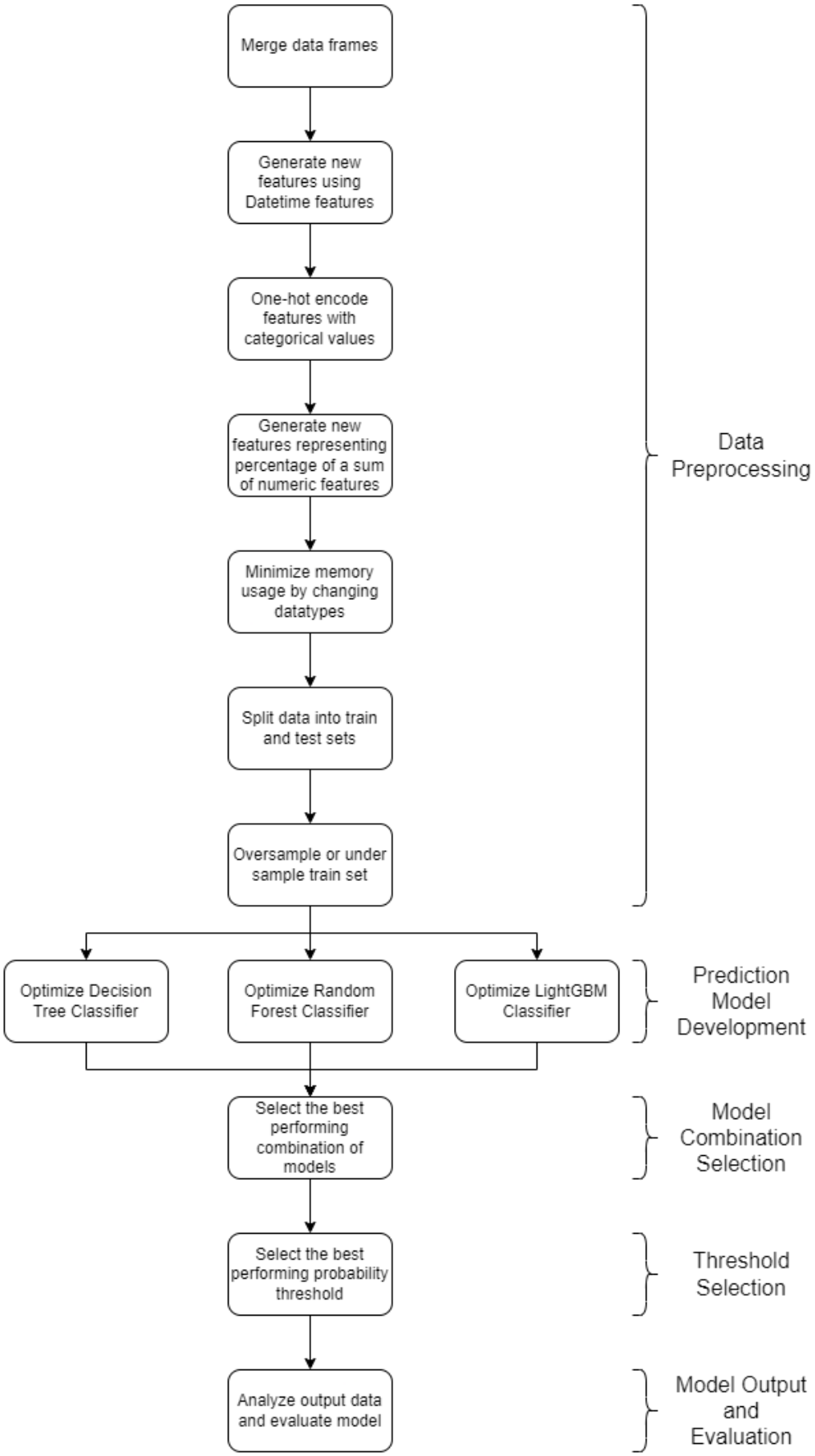
Κεφάλαιο 4. Προτεινόμενη Μεθοδολογία

4.1 Γενική περιγραφή προτεινόμενου πλαισίου – Plug & Play

Η βασική ιδέα του προτεινόμενου πλαισίου είναι η αυτοματοποίηση της διαδικασίας παραγωγής προβλέψεων από σύνολα δεδομένων που αφορούν το churn. Το πλαίσιο αυτό, λειτουργεί ως ένα μαύρο κουτί που δέχεται ως είσοδο ένα πλαίσιο δεδομένων και τα χαρακτηριστικά του, καθώς και ένα σύνολο δεδομένων που επιδέχονται πρόβλεψη, και παράγει ως έξοδο τις προβλέψεις των δεδομένων αυτών, καθώς και μετρικές για την αξιολόγησή τους. Συνήθως, αφού το πρόβλημα αφορά το churn των πελατών ή των υπαλλήλων μιας εταιρείας, ο ορίζοντας πρόβλεψης είναι ο επόμενος μήνας, και τα δεδομένα εισόδου είναι είτε του προηγούμενου μήνα είτε όλα τα δεδομένα που έχουν συγκεντρωθεί. Οι προβλέψεις αυτές αποτελούν ένα καλό και σημαντικό πρώτο βήμα για την ανάλυση της απώλειας πελατών μιας εταιρείας, αλλά, λόγω τη δυσκολίας γενίκευσης και αυτοματοποίησης προβλημάτων που αφορούν ειδικές περιπτώσεις για κάθε σύνολο δεδομένων (για παράδειγμα ελλείπουσες ή παράλογες τιμές, δημιουργία νέων χαρακτηριστικών), προτείνεται περαιτέρω βελτιστοποίηση από το χρήστη. Ωστόσο, με τη βοήθεια των γραφικών παραστάσεων και των μετρικών που βασίζονται στις προβλέψεις του μοντέλου που επιλέχθηκε, παρέχονται στο χρήστη σημαντικές πληροφορίες για το churn του συνόλου δεδομένων που έχει χρησιμοποιήσει.

Η μεθοδολογία που ακολουθήθηκε είναι βασισμένη στην πειραματική διαδικασία που περιγράφεται στο επόμενο κεφάλαιο και αναπτύχθηκε σε περιβάλλον Jupyter Notebook και, συνεπώς, σε γλώσσα προγραμματισμού Python. Ο κώδικας μπορεί να βρεθεί στο GitHub https://github.com/MariosMitropoulos/Plug-and-Play/blob/main/plug-and-play.ipynb. Περιλαμβάνει τα εξής διακριτά βήματα, τα οποία θα αναλυθούν στη συνέχεια:

- Προεπεξεργασία δεδομένων. Το στάδιο αυτό περιλαμβάνει την επεξεργασία των χαρακτηριστικών των δεδομένων εισόδου για την καλύτερη εκπαίδευση των μοντέλων πρόβλεψης και, άρα, την παραγωγή καλύτερων αποτελεσμάτων.
- Ανάπτυξη μοντέλων πρόβλεψης. Εδώ δημιουργούνται και εκπαιδεύονται τα μοντέλα πρόβλεψης, με τη βελτιστοποίηση των υπερπαραμέτρων τους, για την παραγωγή προβλέψεων.
- Επιλογή συνδυασμού μοντέλων. Σε αυτό το βήμα δοκιμάζονται οι συνδυασμοί των μοντέλων παίρνοντας το μέσο όρο τους και επιλέγεται ως βέλτιστο το μοντέλο με τα καλύτερα αποτελέσματα.
- Επιλογή κατωφλίου πρόβλεψης. Στο καλύτερο μοντέλο του προηγούμενου βήματος δοκιμάζονται διάφορα κατώφλια για τις πιθανότητες που παράγει το μοντέλο και επιλέγεται ως βέλτιστο αυτό με τα καλύτερα αποτελέσματα.
- Εξαγωγή αποτελεσμάτων και αξιολόγηση μοντέλου. Στο τελευταίο στάδιο αξιολογείται το συνολικό μοντέλο που προέκυψε με χρήση γραφικών παραστάσεων και μετρικών.



Σχήμα 21: Σχηματική αναπαράσταση της προτεινόμενης μεθοδολογίας

4.2 Προεπεξεργασία δεδομένων

Πρώτο βήμα αποτελεί η προεπεξεργασία του συνόλου δεδομένων. Το βήμα αυτό είναι σημαντικό, καθώς περιλαμβάνει την κατάλληλη επεξεργασία των δεδομένων για να μπορούν να εκπαιδευτούν ορθά τα μοντέλα πρόβλεψης. Όπως αναφέρθηκε και στην εισαγωγή, η απόλυτη αυτοματοποίηση και γενίκευση αυτής της διαδικασίας είναι σχεδόν αδύνατη, καθώς κάθε σύνολο δεδομένων είναι διαφορετικό και περιέχει διαφορετικά πολύπλοκα σημεία. Για παράδειγμα, η αντιμετώπιση των ελλειπουσών τιμών για ένα συγκεκριμένο χαρακτηριστικό εξαρτάται εξ ολοκλήρου από τον τύπο του χαρακτηριστικού. Μερικές περιπτώσεις μπορεί να συμπληρωθούν με μέσο όρο ή με πρόβλεψη βάσει των υπόλοιπων χαρακτηριστικών του συγκεκριμένου παραδείγματος. Σε άλλες είναι καλύτερο να συμπληρωθούν ως ελλείπουσες, ίσως με την προσθήκη νέου χαρακτηριστικού και χρήση one-hot encoding. Επίσης, πολλές φορές είναι χρήσιμη η δημιουργία νέων χαρακτηριστικών από τα υπάρχουσα, τα οποία δίνουν πιο σημαντικές και ουσιαστικές πληροφορίες. Ένα παράδειγμα τέτοιου χαρακτηριστικού είναι η διάρκεια συνδρομής ενός μέλους (αν δεν υπάρχει ήδη το χαρακτηριστικό), χρησιμοποιώντας τα χαρακτηριστικά που περιγράφουν την ημερομηνία λήξης της συνδρομής και την ημερομηνία της συγκεκριμένης συναλλαγής. Παρά τη δυσκολία γενίκευσης αυτής της διαδικασίας, τα βήματα που ακολουθήθηκαν στο προτεινόμενο πλαίσιο είναι απαιτούμενα και αποφέρουν καλή επίδοση για τους σκοπούς εφαρμογής της μεθοδολογίας.

4.2.1 Συγχώνευση δεδομένων

Το αρχικό βήμα της προεπεξεργασίας είναι η συγχώνευση, αν χρειάζεται, των διαφορετικών πλαισίων δεδομένων (data frames) από τα οποία ενδεχομένως να αποτελείται το σύνολο δεδομένων. Για παράδειγμα, διαφορετικοί τύποι data frames μπορεί να είναι τα δημογραφικά δεδομένα των πελατών, τα δεδομένα συναλλαγών μεταξύ εκείνων και της εταιρείας, και τα δεδομένα χρήσης της υπηρεσίας. Μέσα από την πειραματική διαδικασία αναδείχθηκε ότι τα καλύτερα αποτελέσματα παράγονται με τη χρήση όλων των χαρακτηριστικών. Συνεπώς, είναι σημαντικό να συγχωνευτούν τα δεδομένα αυτά σε ένα data frame, το οποίο θα χρησιμοποιηθεί για την εκπαίδευση των μοντέλων πρόβλεψης.

Η συγχώνευση αυτή γίνεται βάσει του μοναδικού αναγνωριστικού που αντιστοιχεί στον κάθε πελάτη ή υπάλληλο, και οι πολλαπλές γραμμές για το ίδιο άτομο επεξεργάζονται κατάλληλα ανάλογα με το χαρακτηριστικό (για παράδειγμα άθροισμα ή μέσος όρος). Ο χρήστης καλείται να συμπληρώσει μια λίστα, η οποία για κάθε data frame περιέχει το ίδιο το data frame, ένα λεξικό με τα χαρακτηριστικά που θα παραμείνουν μετά τη συγχώνευση και ο τρόπος επεξεργασίας πολλαπλών γραμμών τους, και τα χαρακτηριστικά στα οποία πιθανώς απαιτείται μετονομασία. Αυτή η λίστα λαμβάνεται από το πλαίσιο και συγχωνεύει τα data frames εντός ενός βρόγχου.

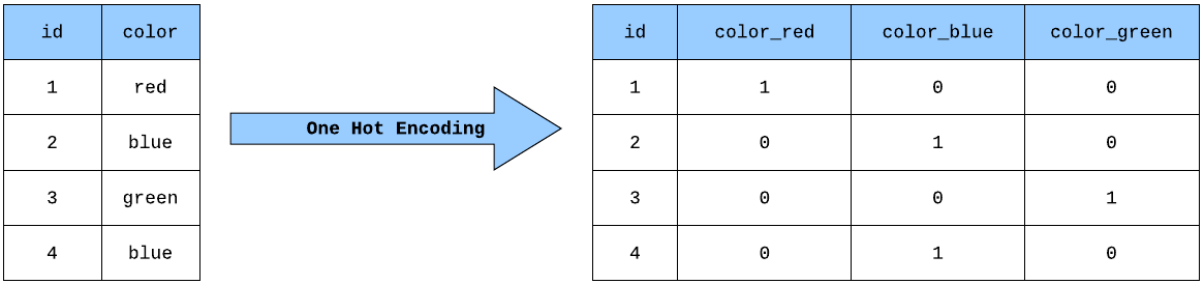
4.2.2 Μετατροπή χαρακτηριστικών ημερομηνιών

Το επόμενο στάδιο της προεπεξεργασίας είναι η μετατροπή των χαρακτηριστικών με τιμές που αντιπροσωπεύουν ημερομηνίες σε μορφή που καθιστά εύκολη την επεξεργασία τους. Η μορφή αυτή είναι το Datetime της βιβλιοθήκης pandas για την Python. Με τον τρόπο αυτό μπορούμε εύκολα να δημιουργήσουμε νέα χαρακτηριστικά βάσει της ημέρας (σπάνια), του μήνα, ή του έτους των αρχικών χαρακτηριστικών. Για παράδειγμα, ένα χαρακτηριστικό που ταιριάζει σε αυτό το βήμα είναι η ημερομηνία εγγραφής του πελάτη στην υπηρεσία. Από αυτή, προκύπτει ο μήνας και το έτος εγγραφής που μπορούν να χρησιμοποιηθούν για διαφορετικές αλλά εξίσου σημαντικές αναλύσεις, όπως στατιστικά εγγραφών ανά χρόνο και ανά μήνα, που το πρώτο αντιστοιχεί σε πληροφορίες σχετικά με την ανάπτυξη και εξέλιξη της εταιρείας, ενώ το δεύτερο στην ανάλυση της εποχικότητας.

Ο χρήστης καλείται να συμπληρώσει μια λίστα με τα ονόματα των χαρακτηριστικών για τα οποία, αρχικά, θα γίνει η μετατροπή σε Datetime και, ύστερα, θα δημιουργηθούν τα νέα χαρακτηριστικά. Αυτή η λίστα λαμβάνεται από το πλαίσιο και εκτελεί τη διαδικασία εντός ενός βρόγχου. Πρέπει να σημειωθεί ότι τα χαρακτηριστικά σε μορφή Datetime αφαιρούνται πριν την εισαγωγή τους στα μοντέλα πρόβλεψης, καθώς δε μπορούν να τα διαχειριστούν.

4.2.3 One-hot encoding κατηγορηματικών χαρακτηριστικών

Επόμενο στη σειρά είναι το one-hot encoding των κατηγορικών χαρακτηριστικών. Το one-hot encoding είναι η τεχνική κατά την οποία ένα κατηγορικό χαρακτηριστικό μετατρέπεται σε πολλαπλά χαρακτηριστικά, ίσα με τον αριθμό των διαφορετικών διακριτών τιμών που λαμβάνει το αρχικό χαρακτηριστικό. Τα νέα χαρακτηριστικά είναι τύπου Boolean, δηλαδή λαμβάνουν τις τιμές 1 ή 0, με τον άσσο να μπαίνει μόνο σε ένα από τα νέα χαρακτηριστικά και τα υπόλοιπα να είναι μηδενικά. Η μετατροπή αυτή απαιτείται για την εκπαίδευση των μοντέλων πρόβλεψης, καθώς αντιμετωπίζουν πιο εύκολα τέτοιου είδους δεδομένα σε σχέση με κατηγορικά. Στο παρακάτω σχήμα παρουσιάζεται ένα παράδειγμα εφαρμογής της τεχνικής one-hot encoding.



Σχήμα 22: Παράδειγμα one-hot encoding, από [36]

Ο χρήστης καλείται να συμπληρώσει μια λίστα με τα ονόματα των χαρακτηριστικών στα οποία θα πραγματοποιηθεί το one-hot encoding. Αυτή η λίστα λαμβάνεται από το πλαίσιο και εκτελεί τη διαδικασία εντός ενός βρόγχου, στον οποίο δημιουργεί τα νέα χαρακτηριστικά, τα συγχωνεύει με το εισαγόμενο data frame, και αφαιρεί τα αρχικά χαρακτηριστικά.

4.2.4 Δημιουργία ποσοστιαίων χαρακτηριστικών

Επόμενο βήμα αποτελεί η δημιουργία χαρακτηριστικών που αντιπροσωπεύουν ποσοστά αντί για απόλυτες τιμές, τα οποία παρέχουν πολύ πιο ουσιαστική πληροφορία που μπορεί να χρησιμοποιηθεί από τα μοντέλα για την παραγωγή καλύτερων προβλέψεων. Δηλαδή, χαρακτηριστικά που όλα μαζί αποτελούν ένα σύνολο και λαμβάνοντας όλα υπόψη έχουμε το 100% μιας πληροφορίας, μπορούν να χρησιμοποιηθούν για τη δημιουργία νέων χαρακτηριστικών βασισμένων στα αρχικά που θα έχουν το ποσοστό που αντιστοιχεί σε κάθε χαρακτηριστικό. Αυτό επιτυγχάνεται με τη διαίρεση της τιμής του κάθε χαρακτηριστικού με το άθροισμα όλων των τιμών των εν λόγω χαρακτηριστικών. Παράδειγμα αποτελεί η δραστηριότητα του χρήστη σε μια υπηρεσία διαδικτυακού streaming μουσικής, όπου για κάθε χρήστη μπορεί να υπάρχουν τα χαρακτηριστικά που περιγράφουν πόσα τραγούδια έχει ακούσει ανάλογα με το ποσοστό της διάρκειάς τους. Δηλαδή, να υπάρχει ένα χαρακτηριστικό για τον αριθμό των τραγουδιών που έχει ακούσει έως το 25% της διάρκειάς του, μεταξύ 25 – 50%, 50 – 75%, και 75 – 100%. Διαιρώντας τα ποσά αυτά με το συνολικό τους άθροισμα, λαμβάνουμε το ποσοστό των τραγουδιών που έχει ακούσει έως το 25% της διάρκειάς του, και ούτω καθεξής. Η πληροφορία αυτή είναι πολύ πιο ουσιαστική από τον απόλυτο αριθμό των τραγουδιών, καθώς πλέον είναι συγκρίσιμα και μεταξύ τους και μεταξύ διαφορετικών χρηστών με διαφορετικές συνήθειες ακρόασης μουσικής. Οι ποσοστιαίες αυτές ενδείξεις μαρτυρούν τη συμπεριφορά του χρήστη για το streaming μουσικής και μπορεί να χρησιμοποιηθεί για την πρόταση κατάλληλων τραγουδιών από την πλατφόρμα προς το χρήστη. Πιο συγκεκριμένα για την προτεινόμενη μεθοδολογία, τα χαρακτηριστικά αυτά χρησιμοποιούνται για την καλύτερη επίδοση των μοντέλων στην πρόβλεψη του churn.

Ο χρήστης καλείται να συμπληρώσει μια λίστα με λίστες ονομάτων χαρακτηριστικών που κάθε εισαγόμενη λίστα αποτελείται από ένα ξεχωριστό σύνολο χαρακτηριστικών που αθροίζονται στο 100%. Αυτή η λίστα λαμβάνεται από το πλαίσιο και εκτελεί τη διαδικασία εντός ενός βρόγχου, στον οποίο δημιουργούνται κατάλληλα τα νέα χαρακτηριστικά. Σημειώνουμε ότι δεν αφαιρούνται τα χαρακτηριστικά στα οποία βασίστηκαν τα νέα.

4.2.5 Ελαχιστοποίηση χρήση μνήμης συστήματος

Μετά από αυτά τα στάδια προεπεξεργασίας του συνόλου δεδομένων, χρησιμοποιείται μια συνάρτηση για την αλλαγή των τύπων των δεδομένων στο

αντίστοιχο μικρότερο δυνατό για τη βέλτιστη αξιοποίηση της μνήμης του συστήματος. Η ελαχιστοποίηση της χρήσης της μνήμης του συστήματος είναι σημαντική, καθώς βελτιστοποιούμε τη χρήση των υπολογιστικών πόρων που έχουμε. Είναι καλή πρακτική, αφού θα υπάρχουν περιπτώσεις που οι υπολογιστικοί πόροι θα είναι περιορισμένοι και η βελτιστοποίηση της χρήσης τους θα επιτρέψει τη μέγιστη αξιοποίησή τους και θα αποτρέψει τη σπατάλη ενέργειας.

Η συνάρτηση που χρησιμοποιείται στο πλαίσιο ελέγχει τη μέγιστη και ελάχιστη τιμή κάθε χαρακτηριστικού και εφαρμόζει το βέλτιστο τύπο σε αυτό.

4.2.6 Διαχωρισμός συνόλου δεδομένων σε train και test set

Πριν την τροφοδοσία των δεδομένων στα μοντέλα πρόβλεψης, πρέπει να προηγηθεί ο διαχωρισμός του συνόλου δεδομένων σε train και test set. Το train set είναι το σύνολο δεδομένων που θα χρησιμοποιηθεί για την εκπαίδευση των μοντέλων, ενώ το test set θα χρησιμοποιηθεί μόνο στο τέλος για την αξιολόγηση των προβλέψεων των μοντέλων πάνω σε άγνωστα δεδομένα. Ο στόχος της επιβλεπόμενης μάθησης είναι η δημιουργία ενός μοντέλου που έχει καλή επίδοση σε προβλέψεις νέων δεδομένων. Έτσι, για την προσομοίωση των νέων αυτών δεδομένων χρησιμοποιείται το test set.

Ο διαχωρισμός που ακολουθήθηκε στο πλαίσιο είναι 80 – 20%, αντίστοιχα, μέσω τυχαίας επιλογής παραδειγμάτων από το αρχικό data frame.

4.2.7 Ισορρόπηση train set

Επόμενη προεργασία που απαιτείται πριν την εκπαίδευση των μοντέλων πρόβλεψης, είναι η ισορρόπηση του train set. Τα δεδομένα που προέρχονται από τον πραγματικό κόσμο συχνά δεν είναι ισορροπημένα, δηλαδή υπάρχει συντριπτική πλειοψηφία δεδομένων μιας κλάσης σε σχέση με την άλλη. Στην περίπτωσή μας, αυτό θα συνέβαινε αν το σύνολο δεδομένων είχε πολλούς περισσότερους churners από non churners, και το αντίστροφο. Είναι σημαντικό να ισορροπηθεί το σύνολο δεδομένων εκπαίδευσης πριν την εισαγωγή τους στα μοντέλα, καθώς η ανισορροπία μπορεί να οδηγήσει τον ταξινομητή να έχει bias προς την κλάση με την πλειοψηφία. Συνεπώς, για να έχουμε την καλύτερη επίδοση δυνατή, πρέπει να ισορροπήσουμε τα δεδομένα εκπαίδευσης με τη χρήση κάποιας τεχνικής. Οι πιο συνηθισμένες τεχνικές είναι το oversampling και το under sampling. Το oversampling χρησιμοποιείται για τη δημιουργία νέων πλασματικών δεδομένων της κλάσης μειοψηφίας με τα λιγότερα δεδομένα, βασιζόμενο στα ήδη υπάρχοντα δεδομένα του συνόλου. Αντιθέτως, το under sampling αφαιρεί δεδομένα της κλάσης πλειοψηφίας, έτσι ώστε να καταλήξει σε ισορροπία. Υπάρχουν πολλές μέθοδοι που εφαρμόζονται για την εφαρμογή των δύο παραπάνω τεχνικών. Σημειώνουμε ότι γενικά το oversampling των δεδομένων θεωρείται καλύτερη πρακτική σε σχέση με το under sampling, καθώς δε χάνουμε δεδομένα και άρα περιπτώσεις που μπορούν να βοηθήσουν τον ταξινομητή να

εκπαιδευτεί καλύτερα. Ωστόσο, επειδή το oversampling μπορεί να αυξήσει σημαντικά το πλήθος των δεδομένων, το under sampling εγγυάται μικρότερους χρόνους εκπαίδευσης. Έτσι, είναι σημαντική η απόφαση μεταξύ τους, έχοντας υπόψη το παραπάνω trade-off. Προφανώς, αν το training set είναι ήδη ισορροπημένο, το βήμα αυτό δεν εκτελείται.

Μέσα από την πειραματική διαδικασία, καταλήξαμε ότι την καλύτερη επίδοση έδειξαν η μέθοδος Synthetic Minority Oversampling Technique (SMOTE) και η random under sampling. Η απόφαση μεταξύ oversampling και under sampling λαμβάνεται βάσει του μεγέθους του συνόλου δεδομένων, ως μια απόπειρα απάντησης στο δίλημμα που περιγράφει το παραπάνω trade-off. Πιο συγκεκριμένα, αν ο αριθμός των γραμμών (παραδειγμάτων) πολλαπλασιασμένος με τον αριθμό των στηλών (χαρακτηριστικών) είναι μεγαλύτερος του 100,000, τότε επιλέγεται oversample, αλλιώς under sample. Ο αριθμός αυτός είναι αυθαίρετος και μπορεί να ρυθμιστεί βάσει της υπολογιστικής ισχύς που είναι διαθέσιμη (η αύξησή του συνεπάγεται περισσότερη χρήση oversampling που είναι υπολογιστικά βαρύτερο).

4.3 Ανάπτυξη μοντέλων πρόβλεψης

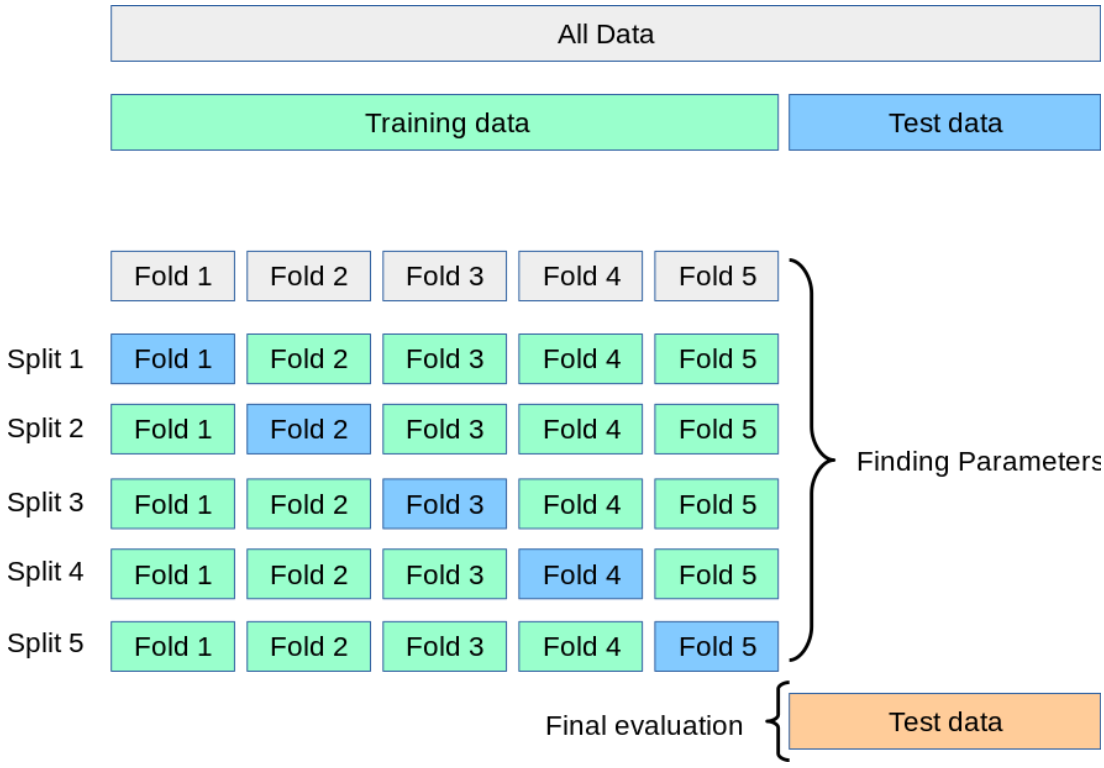
Μέσα από την πειραματική διαδικασία, αναδείχθηκε ότι οι ταξινομητές με την καλύτερη επίδοση σε προβλήματα που αφορούν το churn είναι οι: Decision Tree, Random Forest, και LightGBM. Αποτελούν μοντέλα κλιμακούμενης πολυπλοκότητας, από το Decision Tree προς το LightGBM, βασισμένα στη θεωρία των δέντρων απόφασης, το οποίο τα κάνει κατάλληλα για τη συγκεκριμένη εργασία ταξινόμησης. Χρησιμοποιήθηκαν οι υλοποιήσεις της βιβλιοθήκης scikit learn για τους ταξινομητές Decision Tree και Random Forest, και της Microsoft για το LightGBM.

Η βελτιστοποίηση των υπερπαραμέτρων των ταξινομητών αυτών είναι σημαντική διαδικασία του πλαισίου, καθώς μέσα από αυτή βελτιώνεται η ακρίβεια των προβλέψεων των μοντέλων σε άγνωστα δεδομένα, δηλαδή δεδομένα που δε χρησιμοποιήθηκαν κατά την εκπαίδευση. Η βελτιστοποίηση πραγματοποιείται με την αναζήτηση στο χώρο των υπερπαραμέτρων που έχουμε ορίσει για την εύρεση του συνδυασμού αυτών που ελαχιστοποιούν τη συνάρτηση που έχουμε θέσει ως στόχο. Στην περίπτωσή μας, χρησιμοποιούμε τη μετρική F2-score, η οποία αποτελεί συνδυασμό precision και recall και θα αναλυθεί μαζί με την επεξήγηση της επιλογής της στο επόμενο κεφάλαιο.

Η αναζήτηση αυτή μπορεί να γίνει με διάφορους αλγορίθμους, και δύο από τους πιο διαδεδομένους είναι το grid search και το random search. Στο grid search γίνεται εξαντλητική αναζήτηση όλων των περιπτώσεων και συνδυασμών τιμών των υπερπαραμέτρων που έχουμε ορίσει. Με τον τρόπο αυτό εγγυάται η εύρεση του καλύτερου συνδυασμού, ωστόσο υπολογιστικά είναι πολύ ακριβή τεχνική. Στο random search επιλέγονται τυχαία κάποιοι συνδυασμοί που θα δοκιμαστούν και, στη συνέχεια, επιλέγεται ο καλύτερος από αυτούς. Προφανώς, δεν αποτελεί μέθοδο για τη σίγουρη εύρεση των καλύτερων τιμών των υπερπαραμέτρων, αλλά είναι γρήγορη γεγονός που την καθιστά κατάλληλη σε μερικές περιπτώσεις. Στο προτεινόμενο πλαίσιο

χρησιμοποιήθηκε ένας αλγόριθμος που συνδυάζει στοιχεία και από τους δύο προαναφερθέντες. Χρησιμοποιήθηκε η βιβλιοθήκη HyperOpt, η οποία εκτελεί την αναζήτηση επιλέγοντας υποσύνολο των συνδυασμών των υπερπαραμέτρων, όπως το random search, αλλά αντί για τυχαία επιλογή, επιλέγει βάσει των τιμών που έχουν ήδη δοκιμαστεί για την κάθε υπερπαράμετρο προσπαθώντας να προβλέψει ποια θα είναι η βέλτιστη. Έτσι, έχουμε ένα πέρα από ικανοποιητικό αποτέλεσμα (χωρίς να είναι απαραίτητα το βέλτιστο κάθε φορά) χωρίς να εξετάσουμε κάθε περίπτωση, κερδίζοντας έτσι χρόνο και υπολογιστικούς πόρους.

Για τη μείωση της τυχαιότητας των αποτελεσμάτων και την πιο σίγουρη επιλογή των βέλτιστων τιμών των υπερπαραμέτρων των μοντέλων, χρησιμοποιούμε την τεχνική cross-validation (CV). Το CV αποτελεί τεχνική αξιολόγησης κατά τη διάρκεια της εκπαίδευσης ενός μοντέλου, χωρίς τη χρήση του test set. Πιο συγκεκριμένα, το λεγόμενο k-fold CV, που είναι η πιο διαδεδομένη μορφή του CV, διαχωρίζει το σύνολο δεδομένων εκπαίδευσης σε k μικρότερα σύνολα. Για καθένα από τα k "folds", δηλαδή επαναλήψεις εντός του αλγορίθμου, το μοντέλο χρησιμοποιεί $k-1$ από τα μικρότερα σύνολα στα οποία διαχωρίστηκε προηγουμένως το training set. Το k-οστό από αυτά τα σύνολα χρησιμοποιείται για τη μέτρηση της επίδοσης του μοντέλου. Αυτό επαναλαμβάνεται k φορές, ούτως ώστε να έχει χρησιμοποιηθεί κάθε τμήμα του training set ως σύνολο εκτίμησης της επίδοσης του κάθε fold, δηλαδή validation set. Η συνολική επίδοση του κάθε συνδυασμού ορίζεται ως ο μέσος όρος της επίδοσής του σε καθένα από τα k folds. Στο συγκεκριμένο πλαίσιο χρησιμοποιήθηκε 5-fold CV. Η διαδικασία αυτή περιγράφεται από το παρακάτω σχήμα.



Σχήμα 23: Σχηματική αναπαράσταση k-fold cross-validation, από [37]

Παρακάτω αναγράφονται οι χώροι αναζήτησης για το κάθε μοντέλο που αναπτύχθηκε:

Decision Tree Classifier	
Υπερπαράμετρος	Χώρος αναζήτησης
criterion	{'gini', 'entropy'}
max_depth	{None, 1, 2, …, 32}
max_features	{'sqrt', 'log2', 0.6, 0.7, 0.8, 0.9, 1.0}
min_samples_leaf	{1, 2, …, 20}
min_samples_split	{2, 3, …, 40}

Πίνακας 2: Χώρος αναζήτησης υπερπαραμέτρων Decision Tree Classifier

Random Forest Classifier	
Υπερπαράμετρος	Χώρος αναζήτησης
criterion	{'gini', 'entropy'}
max_features	{0.6, 0.7, 0.8, 0.9, 1.0}
n_estimators	{100, 250, 500, 1000}

Πίνακας 3: Χώρος αναζήτησης υπερπαραμέτρων Random Forest Classifier

LightGBM Classifier	
Υπερπαράμετρος	Χώρος αναζήτησης
colsample_bytree	{0.8, 0.85, 0.9, 0.95, 1.0}
learning_rate	{0.005, 0.01, 0.05, 0.1, 0.3}
max_depth	{-1, 1, 2, …, 32}
n_estimators	{100, 250, 500, 1000}
num_leaves	{31, 50, 100, 200}
subsample	{0.6, 0.7, 0.8, 0.9, 1.0}

Πίνακας 4: Χώρος αναζήτησης υπερπαραμέτρων LightGBM Classifier

4.4 Επιλογή συνδυασμού μοντέλων

Ο συνδυασμός μοντέλων πρόβλεψης αποτελεί τεχνική με στόχο την καλύτερη γενίκευση του συστήματος παραγωγής προβλέψεων. Με το συνδυασμό των μοντέλων, δηλαδή παίρνοντας το μέσο όρο των πιθανοτήτων που προκύπτουν από τα μοντέλα, αυξάνουμε την ευρωστία του συστήματος σε μεμονωμένα σφάλματα ενός από τους ταξινομητές. Για παράδειγμα, αν ένα μοντέλο παραγάγει υπερβολικά υψηλές τιμές και ένα άλλο υπερβολικά χαμηλές, τα σφάλματα αυτά αλληλοακυρώνονται. Επίσης, εξασφαλίζεται η επιλογή του καλύτερου μοντέλου σε περιπτώσεις overfitting, στις οποίες η εκτίμηση των τιμών των υπερπαραμέτρων δεν είναι η βέλτιστη και, ενώ στο validation set το μοντέλο φαίνεται ισχυρό, σε άγνωστα δεδομένα να μην αντεπεξέρχεται κατάλληλα.

Σε αυτό το στάδιο το πλαίσιο έχει ως είσοδο τα μοντέλα πρόβλεψης που αναπτύχθηκαν στο προηγούμενο βήμα μαζί με τις τιμές των υπερπαραμέτρων που τα βελτιστοποιούν. Στόχος είναι η επιλογή του καλύτερου συνδυασμού των παραπάνω ταξινομητών για τη μεγιστοποίηση του δείκτη F2-score. Αφού είχαμε τρία μοντέλα, οι

πιθανοί συνδυασμοί είναι επτά, λαμβάνοντας υπόψη και τις τρεις περιπτώσεις όπου τα καλύτερα μοντέλα είναι ένας από τους ταξινομητές. Στις περιπτώσεις όπου υπάρχει συνδυασμός δύο ταξινομητών ή και των τριών, για την παραγωγή των προβλέψεων λαμβάνεται ο μέσος όρος των πιθανοτήτων που προκύπτουν από κάθε ταξινομητή του συνδυασμού. Ύστερα οι πιθανότητες που προκύπτουν στρογγυλοποιούνται, δηλαδή θέτουμε κατώφλι το 0.5, και προκύπτουν οι προβλέψεις για 0 ή 1.

Για τη μείωση της τυχαιότητας των αποτελεσμάτων και την πιο σίγουρη ανάδειξη του καλύτερου μοντέλου, χρησιμοποιούμε και πάλι την τεχνική CV. Η συνολική επίδοση του κάθε συνδυασμού ορίζεται ως ο μέσος όρος της επίδοσής του σε καθένα από τα k folds. Στο συγκεκριμένο πλαίσιο χρησιμοποιήθηκε 5-fold CV. Αξίζει να σημειωθεί ότι αυτή η τεχνική επιφέρει καλύτερη γενίκευση του προβλήματος και, συνεπώς, καλύτερες προβλέψεις των μοντέλων, ωστόσο είναι υπολογιστικά απαιτητική ανάλογα, βέβαια, και από τους χρόνους εκπαίδευσης των μοντέλων που χρησιμοποιούνται.

4.5 Επιλογή κατωφλίου πρόβλεψης

Συνήθως, η βελτιστοποίηση του κατωφλίου πρόβλεψης είναι τεχνική που χρησιμοποιείται για την αντιμετώπιση της ανισορροπίας ενός συνόλου δεδομένων. Ωστόσο, είναι χρήσιμη τεχνική και για τη συνολική βελτιστοποίηση του συστήματος για την καλύτερη γενίκευση και, συνεπώς, τις καλύτερες δυνατές προβλέψεις πάνω σε άγνωστα δεδομένα. Το πιο συνηθισμένο κατώφλι πρόβλεψης είναι το 0.5. Με τη χρήση αυτού, οι πιθανότητες ουσιαστικά στρογγυλοποιούνται για να μετατραπούν σε προβλέψεις 0 ή 1, με όποια πιθανότητα μεγαλύτερη ή ίση του 0.5 να ανήκει στην κλάση 1, ενώ τις υπόλοιπες στην κλάση 0. Εντούτοις, πολλές φορές η επιλογή διαφορετικού κατωφλίου μπορεί να αποφέρει πολύ καλύτερα και έμπιστα αποτελέσματα. Διαισθητικά, περιμένουμε αύξηση της ακρίβειας με την αύξηση του κατωφλίου, καθώς τα παραδείγματα με μεγαλύτερες πιθανότητες είναι πιο πιθανά να ανήκουν στην κλάση 1 και αναμένουμε στατιστικά ότι με την αύξηση του κατωφλίου όλο και περισσότερα παραδείγματα θα έχουν προβλεφθεί ορθά και θα ανήκουν στην κλάση 1. Ωστόσο, με την αύξηση του κατωφλίου αναμένουμε να μειωθεί η μετρική του recall, δηλαδή το ποσοστό των συνολικών παραδειγμάτων που ανήκουν στην κλάση 1 και προβλέφθηκαν ορθά, αφού μερικές περιπτώσεις που ήταν κοντά στο 50% πλέον ταξινομούνται ως αντικείμενα της άλλης κατηγορίας. Το αντίστοιχο ισχύει και για την κλάση 0, με μικρότερες πιθανότητες και τη μείωση του κατωφλίου.

Έτσι, για την επίτευξη της βέλτιστης δυνατής επίδοσης του συστήματος, στο στάδιο αυτό εφαρμόζεται και πάλι 5-fold CV για την ανάδειξη του καλύτερου κατωφλίου από 0.1 έως και 0.9 με βήμα 0.1. Για την επιλογή μεταξύ των διαφορετικών επιλογών χρησιμοποιείται και πάλι η μετρική F2-score.

4.6 Εξαγωγή αποτελεσμάτων και αξιολόγηση μοντέλου

Το τελευταίο στάδιο της προτεινόμενης μεθοδολογίας είναι η εξαγωγή των αποτελεσμάτων και η αξιολόγηση του συνολικού μοντέλου. Για την αξιολόγηση του μοντέλου χρησιμοποιείται, αρχικά, το test set από το διαχωρισμό του συνόλου δεδομένων που πραγματοποιήθηκε σε προηγούμενο βήμα. Με τον τρόπο αυτό, προκύπτουν οι μετρικές τις οποίες χρησιμοποιούμε για την αξιολόγηση του συστήματος. Οι μετρικές αυτές είναι οι ακόλουθες: precision, recall, specificity, accuracy, F1-score, και F2-score, με τη μεγαλύτερη βαρύτητα να δίνεται στο F2-score καθώς είναι και αυτό που χρησιμοποιήθηκε ως συνάρτηση μεγιστοποίησης για τη βελτιστοποίηση κάθε σταδίου της μεθοδολογίας.

Πέρα από τις μετρικές, χρησιμοποιώντας τις τελικές πιθανότητες που παράγει το μοντέλο πρόβλεψης δημιουργούνται τρεις γραφικές παραστάσεις. Η πρώτη αποτελεί ένα ιστόγραμμα που αναπαριστά γραφικά το πλήθος των πιθανοτήτων σε κάθε διάστημα από το 0 έως το 1 μήκους 0.1. Έτσι, έχουμε μια εικόνα για τα αποτελέσματα του συστήματος και μαζί με την επιλογή του τελικού κατωφλίου εξάγεται πληροφορία σχετικά με αυτά.

Έπειτα, η δεύτερη παράγεται από τη συνάρτηση “plot_importance” του LightGBM και μας παρέχει τη σημαντικότητα κάθε χαρακτηριστικού που χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου LightGBM. Ακόμα και στην περίπτωση που εν τέλει δεν επιλεχθεί το μοντέλο LightGBM ως βέλτιστο, μας παρέχει ουσιαστική πληροφορία για το ποια χαρακτηριστικά παίζουν το σημαντικότερο ρόλο για την παραγωγή των προβλέψεων των μοντέλων.

Τέλος, η τρίτη γραφική παράσταση αφορά το ποσοστό της απώλειας των πελατών σε υποσύνολα των πελατών, τα οποία δημιουργούνται οργανώνοντας τις πιθανότητες σε φθίνουσα σειρά και παίρνοντας κάθε φορά μεγαλύτερο ποσοστό (από 10% έως 100%). Με την αύξηση του ποσοστού των πελατών αναμένουμε μείωση στο churn rate, καθώς οι πιθανότητες είναι ταξινομημένες σε φθίνουσα σειρά. Έτσι, με τη γραφική αυτή μπορεί να αποφασιστεί το κατάλληλο ποσοστό των πελατών στο οποίο αξίζει να επενδυθούν χρήματα για τη διατήρηση των πελατών στην υπηρεσία και την πρόληψη του churn. Αυτό μπορεί να επιτευχθεί με κάποια ειδική προσφορά ή έκπτωση, όπως έχει αναλυθεί και στο δεύτερο κεφάλαιο της εργασίας.

Κεφάλαιο 5. Πειραματική Διαδικασία και Αποτελέσματα

5.1 Πειραματικό σύνολο δεδομένων

Για τη μελέτη του churn, αναζητήσαμε σύνολα δεδομένων που αφορούν συνδρομητικές εταιρείες, καθώς είναι εκείνες τις οποίες ενδιαφέρει η ανάλυση και η πρόβλεψη της απώλειας πελατών της υπηρεσίας που παρέχουν. Σε αυτό το πλαίσιο, το σύνολο δεδομένων που επιλέχθηκε ήταν τα δεδομένα της KKBox Inc., μιας εταιρείας που παρέχει υπηρεσία διαδικτυακού streaming μουσικής. Ακολουθεί μοντέλο freemium, δηλαδή παρέχει μια δωρεάν έκδοση της υπηρεσίας, ενώ ο χρήστης μπορεί να αναβαθμίσει σε συνδρομή για επιπλέον προνόμια. Επικεντρώνεται σε περιοχές όπως η Ταϊβάν, το Χονγκ Κονγκ, η Μαλαισία, η Ιαπωνία, και η Σιγκαπούρη. Τα δεδομένα αυτά δημοσιεύθηκαν από την εταιρεία με τη μορφή διαγωνισμού στην πλατφόρμα Kaggle της Google, στον οποίο οι συμμετέχοντες καλούνταν να προβλέψουν αν οι χρήστες για τους οποίους έληγε η συνδρομή τον επόμενο μήνα από εκείνον που ανέβηκε ο διαγωνισμός θα οδηγούνταν σε churn ή όχι. Ο διαγωνισμός μπορεί να βρεθεί στον ακόλουθο σύνδεσμο https://www.kaggle.com/competitions/kkbox-churn-prediction-challenge.

Το σύνολο δεδομένων αποτελείται από τα δεδομένα που θα χρησιμοποιηθούν για την εκπαίδευση και την αξιολόγηση του μοντέλου που θα αναπτυχθεί και έναν πίνακα με μοναδικά αναγνωριστικά χρηστών για τους οποίους απαιτείται η πρόβλεψη του αν θα συνεχίσουν να χρησιμοποιούν την υπηρεσία. Ο πίνακας αυτός δε θα χρησιμοποιηθεί κατά τη διάρκεια της πειραματικής διαδικασίας, καθώς στόχος της διαδικασίας είναι η ανάλυση και η εκτίμηση των διάφορων μοντέλων και τεχνικών, ενώ ο πίνακας αυτός χρησιμοποιείται μόνο για την υποβολή των αποτελεσμάτων στο διαγωνισμό. Το σύνολο δεδομένων που θα χρησιμοποιηθεί για την εκπαίδευση περιέχει τέσσερις πίνακες δεδομένων με κοινό χαρακτηριστικό ένα μοναδικό αναγνωριστικό για κάθε χρήστη. Το δεύτερο χαρακτηριστικό του πρώτου πίνακα είναι το label του χρήστη, δηλαδή αν θα οδηγηθεί σε churn στο τέλος του μήνα ή όχι. Οι υπόλοιποι πίνακες περιέχουν διαφορετικούς τύπους δεδομένων για τον κάθε χρήστη: δημογραφικά δεδομένα (πόλη, ηλικία, φύλο, και άλλα), δεδομένα των συναλλαγών που έχουν πραγματοποιηθεί μέσω της υπηρεσίας (τρόπος πληρωμής, ημερομηνία λήξης συνδρομής, πόσο πληρωμής, και άλλα), και δεδομένα χρήσης της υπηρεσίας streaming μουσικής (αριθμός μοναδικών τραγουδιών που έχει ακούσει, συνολικά δευτερόλεπτα ακρόασης).

Ο λόγος επιλογής αυτού του συνόλου δεδομένων είναι διττός. Αρχικά, το μέγεθος των δεδομένων είναι σημαντικό (725,722 γραμμές και 72 στήλες μετά την προεπεξεργασία), γεγονός που είναι απαραίτητο για τη βέλτιστη εκπαίδευση των μοντέλων και την εξαγωγή συμπερασμάτων σχετικά με την ανάλυση του churn. Έπειτα, το συγκεκριμένο σύνολο δεν έχει σχεδόν καθόλου ελλείπουσες τιμές, με τις μόνες να εμφανίζονται στο χαρακτηριστικό του φύλου του χρήστη, το οποίο δε φάνηκε να έχει ουσιαστική σχέση με το churn του χρήστη, όπως αναδείχθηκε από την εφαρμογή κατάλληλης τεχνικής στατιστικής ανάλυσης.

5.1.1 Περιγραφή χαρακτηριστικών συνόλου δεδομένων

Παρουσιάζουμε παρακάτω τα χαρακτηριστικά του συνόλου δεδομένων που χρησιμοποιήθηκαν στην πειραματική διαδικασία, δηλαδή τα τέσσερα data frames που προαναφέρθηκαν, καθώς και μερικά συμπεράσματα που προέκυψαν από τη χρήση στατιστικής ανάλυσης για τη σχέση μεταξύ των χαρακτηριστικών και του churn.

Train data frame	
Χαρακτηριστικό	Περιγραφή
msno	μοναδικό αναγνωριστικό χρήστη (mission number)
is_churn	label για το churn (1 ή 0)

Πίνακας 5: Train data frame

Εδώ αξίζει να σημειωθεί ότι έχουμε σημαντική πλειοψηφία της κλάσης 0 (όχι churn), με περίπου 91% να ανήκει σε αυτή, και το υπόλοιπο 9% στην κλάση 1. Πρέπει, δηλαδή, να ισορροπήσουμε κατάλληλα τα δεδομένα πριν την εκπαίδευση των μοντέλων πρόβλεψης. Ακολουθεί το data frame με τα δημογραφικά δεδομένα των χρηστών.

Members data frame	
Χαρακτηριστικό	Περιγραφή
msno	μοναδικό αναγνωριστικό χρήστη (mission number)
city	πόλη του χρήστη
bd	ηλικία του χρήστη
gender	φύλο του χρήστη
registered_via	μέθοδος εγγραφής
registration_init_time	ημερομηνία εγγραφής

Πίνακας 6: Members data frame

Αξιοσημείωτα συμπεράσματα από τη στατιστική ανάλυση αυτών των χαρακτηριστικών αποτελούν το γεγονός ότι η ηλικιακή ομάδα 10 – 17, και σε μικρότερο βαθμό η 18 – 24, παρουσίασε σημαντικά μεγαλύτερο churn rate σε σχέση με τις υπόλοιπες, οι οποίες ήταν κοντά στο μέσο churn rate. Επίσης, φάνηκε ότι οι χρήστες που δεν είχαν συμπληρώσει το φύλο τους είχαν μικρότερο μέσο churn rate συγκριτικά με τους καταγεγραμμένους άνδρες και γυναίκες, που παρουσίασαν σχεδόν ίσο μέσο churn rate. Τέλος, όσον αφορά τις μεθόδους εγγραφής, οι μέθοδοι που φαίνονται συνδεδεμένες με την αύξηση του churn είναι η 3 και η 4, ενώ η 7 φάνηκε να έχει το αντίθετο αποτέλεσμα. Ωστόσο, δε μας παρέχονται τα στοιχεία για το ποιες είναι οι μέθοδοι αυτοί. Ακολουθεί το data frame με τα δεδομένα συναλλαγών.

Transactions data frame	
Χαρακτηριστικό	Περιγραφή
msno	μοναδικό αναγνωριστικό χρήστη (mission number)
payment_method_id	αναγνωριστικό της μεθόδου πληρωμής
payment_plan_days	διάρκεια πλάνου συνδρομής σε μέρες

plan_list_price	αρχική τιμή του πλάνου σε νέα δολάρια Ταϊβάν (NTD)
actual_amount_paid	τελικό ποσό πληρωμής σε νέα δολάρια Ταϊβάν (NTD)
is_auto_renew	αν ο χρήστης έχει ενεργοποιήσει αυτόματη ανανέωση συνρδομής
transaction_date	ημερομηνία συναλλαγής
membership_expire_date	ημερομηνία λήξης συνδρομής
is_cancel	αν ο χρήστης διέκοψε τη συνδρομή του σε αυτή τη συναλλαγή

Πίνακας 7: Transactions data frame

Όπως αναμέναμε, τα χαρακτηριστικά "is_auto_renew" και "is_cancel" είναι άμεσα συνδεδεμένα με την τελική ταξινόμηση του χρήστη ως churner ή όχι. Με την εφαρμογή χ^2 τεστ, μπορούμε να αποφανθούμε ότι το churn είναι ανάλογο με τη διακοπή της συνδρομής του χρήστη σε αυτή τη συναλλαγή ("is_cancel"), δηλαδή αν είναι αληθής η τιμή του χαρακτηριστικού αυτού τότε η πιθανότητα του churn αυξάνεται σημαντικά, ενώ είναι αντιστρόφως ανάλογο με την αυτόματη ανανέωση συνδρομής, δηλαδή η ενεργοποίησή της μειώνει την πιθανότητα του churn. Ακολουθεί το data frame με τα δεδομένα χρήσης της υπηρεσίας για τον κάθε χρήστη.

User logs data frame	
Χαρακτηριστικό	Περιγραφή
msno	μοναδικό αναγνωριστικό χρήστη (mission number)
date	ημερομηνία του συγκεκριμένου log
num_25	αριθμός τραγουδιών για τα οποία παίχτηκε λιγότερο από 25% της διάρκειάς τους
num_50	αριθμός τραγουδιών για τα οποία παίχτηκε μεταξύ 25% και 50% της διάρκειάς τους
num_75	αριθμός τραγουδιών για τα οποία παίχτηκε μεταξύ 50% και 75% της διάρκειάς τους
num_985	αριθμός τραγουδιών για τα οποία παίχτηκε μεταξύ 75% και 98.5% της διάρκειάς τους
num_100	αριθμός τραγουδιών για τα οποία παίχτηκε περισσότερο από 98.5% της διάρκειάς τους
num_unq	αριθμός μοναδικών τραγουδιών που παίχτηκαν
total_secs	συνολικά δευτερόλεπτα ακρόασης

Πίνακας 8: User logs data frame

Αξιοσημείωτο εδώ είναι το γεγονός ότι το churn δε φάνηκε να επηρεάζεται σημαντικά βάσει της χρήσης της υπηρεσίας από τους πελάτες στα τεστ στατιστικής ανάλυσης που εκτελέστηκαν.

5.2 Προεπεξεργασία συνόλου δεδομένων εκπαίδευσης

Η προεπεξεργασία του συνόλου δεδομένων είναι απαραίτητη για τη βελτιστοποίηση των αποτελεσμάτων των μοντέλων πρόβλεψης και, συνεπώς, είναι σημαντική για την ορθή ανάλυση της πειραματικής διαδικασίας. Το σύνολο δεδομένων που επιλέχθηκε δεν έχει πολλές ελλείπουσες τιμές, ωστόσο επιδέχεται κατάλληλη τροποποίηση για την εξαγωγή καλύτερων αποτελεσμάτων.

5.2.1 Members data frame

Η πρώτη τροποποίηση που πραγματοποιήθηκε είναι η μετατροπή των χαρακτηριστικών με κατηγορικές τιμές με one-hot encoding, τεχνική που αναφέρθηκε στο προηγούμενο κεφάλαιο της εργασίας. Η τεχνική αυτή εφαρμόστηκε για τα χαρακτηριστικά “city”, “gender”, “registered_via", και “registration_init_time”. Ειδικότερα, στο χαρακτηριστικό για το φύλο του χρήστη υπήρχαν ελλείπουσες τιμές, όπως αναφέρθηκε και στην ανάλυση του συνόλου δεδομένων. Έτσι, πέρα από τα χαρακτηριστικά για άνδρα και γυναίκα, δημιουργήθηκε και ένα χαρακτηριστικό που λαμβάνει την τιμή 1 (και τα άλλα δύο την τιμή 0) στις περιπτώσεις που ο χρήστης δεν είχε συμπληρώσει το φύλο του. Επιπλέον, το “registration_init_time” μετατράπηκε, αρχικά, από την εξαψήφια αριθμητική μορφή YYMMDD σε μορφή Datetime για να μπορεί να διαχειριστεί με μεγαλύτερη ευκολία, καθώς η αρχική μορφή μπορεί να είναι εύκολα αντιληπτή από τον άνθρωπο αλλά όχι από τον υπολογιστή. Ύστερα, δημιουργήθηκε νέο χαρακτηριστικό βάσει του χρόνου εγγραφής, στο οποίο τελικά εφαρμόστηκε one-hot encoding, και διαγράφτηκε το αρχικό χαρακτηριστικό.

Επόμενο εμπόδιο του data frame αυτού αποτελούν μερικές μη λογικές τιμές του χαρακτηριστικού “bd”, το οποίο λαμβάνει τιμές από -7168 έως και 2016. Αρχικά, αντικαθιστούμε τις αρνητικές τιμές με 0, το οποίο χρησιμοποιούμε ως ένδειξη έλλειψης της συγκεκριμένης πληροφορίας. Σε αυτό το σημείο απαιτείται η λήψη μιας αυθαίρετης απόφασης για τις ηλικίες που επιτρέπονται, καθώς καλούμαστε να μετατρέψουμε τα έτη γεννήσεως σε ηλικίες και να απορρίψουμε αν χρειάζεται ηλικίες που δεν αντιστοιχούν σε πιθανό χρήστη της υπηρεσίας, δηλαδή να βρίσκεται σε υπερβολικά νεαρή ή μεγάλη ηλικία. Έτσι, επιλέγουμε ένα εύρος από 10 έως και 80 ετών. Μετατρέπουμε σε αριθμό ηλικίας όποια έτη γεννήσεως ανήκουν σε αυτό το διάστημα, και τις υπόλοιπες τιμές τις μετατρέπουμε σε 0, ως ένδειξη απουσίας πληροφορίας μαζί με τις αρνητικές τιμές.

5.2.2 Transactions data frame

Και πάλι, το πρώτο βήμα που εκτελέστηκε είναι η εφαρμογή one-hot encoding στα κατηγορηματικά χαρακτηριστικά. Αυτό περιλαμβάνει το “payment_method_id”, δηλαδή τους μεθόδους πληρωμής. Σημειώνουμε ότι επειδή δε βρέθηκε άμεση αντιστοίχιση του χαρακτηριστικού αυτού με το churn rate και λόγω του ότι οι συνολικές μοναδικές τιμές που λαμβάνει το χαρακτηριστικό αυτό είναι 37, αριθμός που είναι μη ρεαλιστικός για πλήθος μεθόδων πληρωμής, τα χαρακτηριστικά αυτά τελικά αφαιρέθηκαν από τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση, όπως θα αναφερθεί και στη συγχώνευση των data frames.

Τα χαρακτηριστικά “transaction_date” και “membership_expire_date” ακολουθούν την ίδια μορφή με το “registration_init_time” που συναντήσαμε προηγουμένως. Και πάλι, για τη διαχείρισή τους τα μετατρέπουμε σε Datetime αλλά αυτή τη φορά τα χρησιμοποιούμε για τη δημιουργία ενός νέου χαρακτηριστικού, του “membership_duration”. Αφαιρώντας το “transaction_date” από το “membership_expire_date” λαμβάνουμε τη διάρκεια της συνδρομής που προκύπτει μέσα από τη συγκεκριμένη συναλλαγή. Το χαρακτηριστικό αυτό παρέχει ουσιαστική πληροφορία για το χρήστη, αξιοποιώντας βέλτιστα τα χαρακτηριστικά που δε θα χρησιμοποιηθούν κατά την εκπαίδευση, αφού είναι τύπου ημερομηνίας.

Τα υπόλοιπα χαρακτηριστικά δε χρειάστηκαν κάποια αλλαγή, αλλά με τη χρήση κάποιων από αυτά δημιουργήσαμε δύο νέα χαρακτηριστικά που ίσως παράσχουν σημαντική πληροφορία και βοηθήσουν τα μοντέλα στην εκπαίδευση και τις προβλέψεις τους. Το πρώτο χαρακτηριστικό είναι το “price_difference” που προκύπτει από την αφαίρεση του “actual_amount_paid” από το “plan_list_price” και αντιπροσωπεύει την έκπτωση που έλαβε ο χρήστης στη συναλλαγή αυτή. Δεύτερο χαρακτηριστικό είναι το “amount_per_day” που προκύπτει από τη διαίρεση του “payment_plan_days” από το “actual_amount_paid” και αντιπροσωπεύει το μέσο όρο χρημάτων που ξοδεύει ο χρήστης για τη χρήση της υπηρεσίας καθημερινά βάσει του πακέτου συνδρομής της συγκεκριμένης συναλλαγής.

5.2.3 User logs data frame

Σε αυτό το data frame χρησιμοποιούμε τα χαρακτηριστικά που αντιπροσωπεύουν τον αριθμό των τραγουδιών που άκουσε ο χρήστης μέχρι κάποιο ποσοστό της διάρκειάς τους για τη δημιουργία παρόμοιων χαρακτηριστικών, που όμως έχουν πιο ουσιαστική πληροφορία. Αυτό το πετυχαίνουμε δημιουργώντας χαρακτηριστικά που αντιπροσωπεύουν ποσοστό αντί για τον απόλυτο αριθμό των τραγουδιών που άκουσε ο χρήστης μέχρι κάποιο ποσοστό της διάρκειάς τους. Για τη δημιουργία των χαρακτηριστικών αυτών, προσθέτουμε τις στήλες έτσι ώστε να προκύψει ένας αριθμός για το κάθε user log, και διαιρούμε το κάθε αρχικό χαρακτηριστικό (“num_25”, “num_50”, “num_75”, “num_985”, και “num_100”) με το αντίστοιχο άθροισμα. Με τον τρόπο αυτό, έχουμε πλέον μια πιο ουσιαστική εικόνα για

τις συνήθειες του χρήστη εντός της υπηρεσίας. Η πληροφορία αυτή είναι πολύ πιο χρήσιμη από τον απόλυτο αριθμό των τραγουδιών που είχαμε αρχικά, καθώς πλέον οι τιμές των χαρακτηριστικών αυτών είναι εύκολα συγκρίσιμες και με άλλους χρήστες της εφαρμογής.

5.2.4 Συγχώνευση data frames

Η συγχώνευση όλων των data frames στην κατάσταση που βρίσκονται μέχρι στιγμής θα παρήγαγε απαγορευτικά μεγάλο σύνολο δεδομένων, καθώς τα transactions και user logs περιέχουν πολλαπλές γραμμές για τον κάθε χρήστη. Με τη συγχώνευσή τους στο κοινό τους χαρακτηριστικό, το μοναδικό αναγνωριστικό του χρήστη, θα είχαμε όλους τους διαφορετικούς συνδυασμούς που προκύπτουν. Για την αποφυγή του φαινόμενου αυτού και την καλύτερη οργάνωση της πληροφορίας, συγχωνεύουμε πρώτα τις πολλαπλές γραμμές των transactions και user logs data frames σε μια γραμμή ανά χρήστη.

Όσον αφορά το transactions data frame, οι αλλαγές που πραγματοποιήθηκαν για τη συγχώνευση των χαρακτηριστικών παρουσιάζονται στον παρακάτω πίνακα, ενώ τα χαρακτηριστικά “payment_method_id”, “payment_plan_days”, και “plan_list_price” δε χρησιμοποιήθηκαν, καθώς οι μέθοδοι πληρωμής δεν είχαν ρεαλιστικές τιμές και τα τελευταία δύο χρησιμοποιήθηκαν για την παραγωγή νέων, χρησιμότερων χαρακτηριστικών.

Transactions data frame	
Χαρακτηριστικό	Τρόπος συνάθροισης
membership_duration	άθροισμα
actual_amount_paid	άθροισμα
price_difference	μέσος όρος
amount_per_day	μέσος όρος
is_auto_renew	μέσος όρος (μετατροπή σε “uses_auto_renew”)
is_cancel	μέσος όρος (μετατροπή σε “has_cancelled”)

Πίνακας 9: Χαρακτηριστικά συνάθροισης του transactions data frame

Αυτό που πρέπει να σχολιαστεί από τον παραπάνω πίνακα, είναι οι δύο μετατροπές των χαρακτηριστικών που αφορούν την αυτόματη ανανέωση της συνδρομής και τη διακοπή της. Κατά τη συνάθροιση των χαρακτηριστικών αυτών παίρνουμε το μέσο όρο τους, και ουσιαστικά έχουμε το ποσοστό των συναλλαγών στο οποίο τα χαρακτηριστικά αυτά παίρνουν την τιμή 1. Έτσι, για τη δημιουργία του χαρακτηριστικού “has_cancelled” μετατρέπουμε όποιο ποσοστό δεν είναι 0% σε 1, και αφήνουμε ως 0 τα υπόλοιπα. Για το “uses_auto_renew” θέτουμε αυθαίρετα ένα κατώφλι για το οποίο θεωρείται ότι ο χρήστης χρησιμοποιεί γενικά την αυτόματη ανανέωση συνδρομής. Στο πλαίσιο αυτής της πειραματικής διαδικασίας το όριο αυτό τέθηκε ως 0.75. Δημιουργήσαμε, επίσης, άλλο ένα νέο χαρακτηριστικό που αντιπροσωπεύει το

συνολικό αριθμό των συναλλαγών του χρήστη. Ακολουθεί ο πίνακας για τη συνάθροιση των δεδομένων του user logs data frame.

User logs data frame	
Χαρακτηριστικό	Τρόπος συνάθροισης
num_unq	άθροισμα
total_secs	άθροισμα
num_25_percentage	μέσος όρος
num_50_percentage	μέσος όρος
num_75_percentage	μέσος όρος
num_985_percentage	μέσος όρος
num_100_percentage	μέσος όρος

Πίνακας 10: Χαρακτηριστικά συνάθροισης του user logs data frame

Σημειώνουμε ότι παίρνοντας το μέσο όρο ποσοστών που αθροίζουν στο 100% ανά γραμμή, η σχέση αυτή παραμένει ισχύουσα και μετά τη συνάθροιση. Επίσης, δημιουργήσαμε και σε αυτήν την περίπτωση ένα νέο χαρακτηριστικό που αντιπροσωπεύει το συνολικό αριθμό των user logs του χρήστη.

Μετά από αυτήν την προεπεξεργασία, συγχωνεύουμε τα data frames και προκύπτει ένα ενιαίο με 725,722 γραμμές, καθεμία από τις οποίες αντιπροσωπεύει έναν χρήστη, και 72 στήλες, καθεμία από τις οποίες αντιπροσωπεύει ένα χαρακτηριστικό (σε αυτά δεν προσμετρούνται το μοναδικό αναγνωριστικό και η στήλη του label).

5.3 Δείκτες αξιολόγησης

Οι δείκτες αξιολόγησης που χρησιμοποιήθηκαν στην πειραματική διαδικασία είναι οι εξής: accuracy, precision, recall, specificity, F1-score, και F2-score. Η περιγραφή και ο ορισμός τους έχει προηγηθεί στην εργασία αυτή, στο δεύτερο κεφάλαιο, ωστόσο αξίζει να σημειωθεί ότι το F2-score έχει επιλεχθεί ως η σημαντικότερη μετρική και, έτσι, χρησιμοποιείται για τη βελτιστοποίηση του μοντέλου που θα αναπτυχθεί. Η βελτιστοποίηση αυτή περιλαμβάνει την επιλογή των βέλτιστων υπερπαραμέτρων των μοντέλων πρόβλεψης κατά την πειραματική διαδικασία, αλλά και την επιλογή του βέλτιστου συνδυασμού μοντέλων και του βέλτιστου κατωφλίου ταξινόμησης για τις πιθανότητες εντός της προτεινόμενης μεθοδολογίας, Plug & Play. Για την εξήγηση της επιλογής αυτής αξίζει να επανεξετάσουμε το μαθηματικό τύπου του F_β-score:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}.$$

Για $\beta = 1$ έχουμε το F1-score που αποτελεί τον αρμονικό μέσο μεταξύ precision και recall. Για μικρότερες τιμές του β δίνεται μεγαλύτερο βάρος στο precision, ενώ για μεγαλύτερες τιμές στο recall. Για $\beta = 2$ έχουμε:

$$F_2 = \frac{5 \times Precision \times Recall}{4 \times Precision + Recall}.$$

Γνωρίζουμε ότι το precision είναι μια μετρική που υπολογίζει το ποσοστό των σωστών προβλέψεων της θετικής κλάσης, ενώ το recall υπολογίζει το ποσοστό των πραγματικών στιγμιότυπων της θετικής κλάσης που προέβλεψε το μοντέλο. Έτσι, καταλαβαίνουμε ότι μεγιστοποιώντας το precision ελαχιστοποιούμε τα false positives, ενώ μεγιστοποιώντας το recall ελαχιστοποιούμε τα false negatives. Συνεπώς, η ερώτηση που πρέπει να απαντηθεί είναι ποιο από τα δύο είναι σημαντικότερο στο συγκεκριμένο πρόβλημα, δηλαδή αν είναι σημαντικότερη η σωστή ταξινόμηση όσο το δυνατόν περισσότερων churners ή η μεγιστοποίηση του αριθμού των σωστών προβλέψεων γενικά.

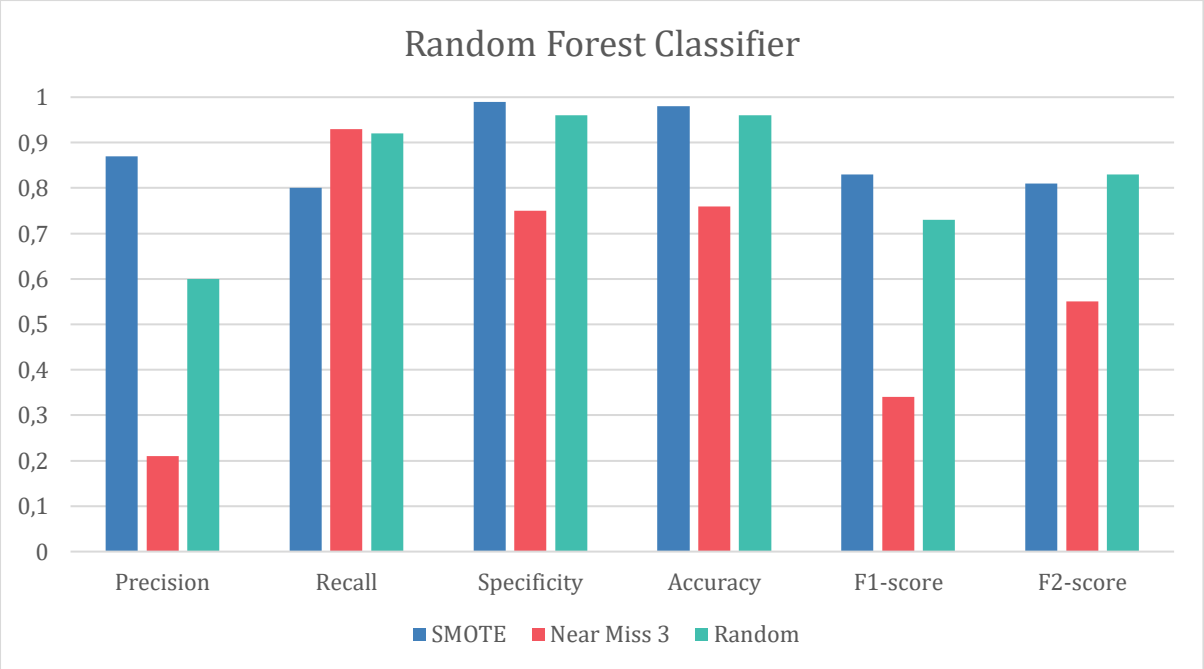
Θεωρούμε ότι η εταιρεία θα προσφέρει στους χρήστες που θα ταξινομηθούν ως churners κάποιους είδους έκπτωση και στους non churners όχι. Με αυτήν την υπόθεση, τα false positives υπονοούν ότι θα δοθεί προσφορά σε κάποιον χρήστη ο οποίος δε θα σταματούσε να χρησιμοποιεί την υπηρεσία (και άρα η εταιρεία δε μεγιστοποιεί το κέρδος της), ενώ τα false negatives υπονοούν ότι ο χρήστης δε θα λάβει κάποια προσφορά και, συνεπώς, δε θα έχει λόγο να αλλάξει γνώμη για την αποχώρησή του από την υπηρεσία, η οποία χάνει τον πελάτη αυτό. Έτσι, η ερώτηση που προκύπτει είναι σε ποια περίπτωση η εταιρεία έχει μεγαλύτερο κέρδος. Υποθέτουμε ότι, γενικά, η εταιρεία έχει μεγαλύτερο κέρδος με την απώλεια του ελάχιστου δυνατού αριθμού πελατών. Άρα, θέτουμε μεγαλύτερο βάρος στο recall, χωρίς να πάψει να είναι σημαντικό το precision, και επιλέγουμε ως κύρια μετρική και συνάρτηση στόχο για τα cross-validation το F2-score.

5.4 Πειραματική διαδικασία και αποτελέσματα

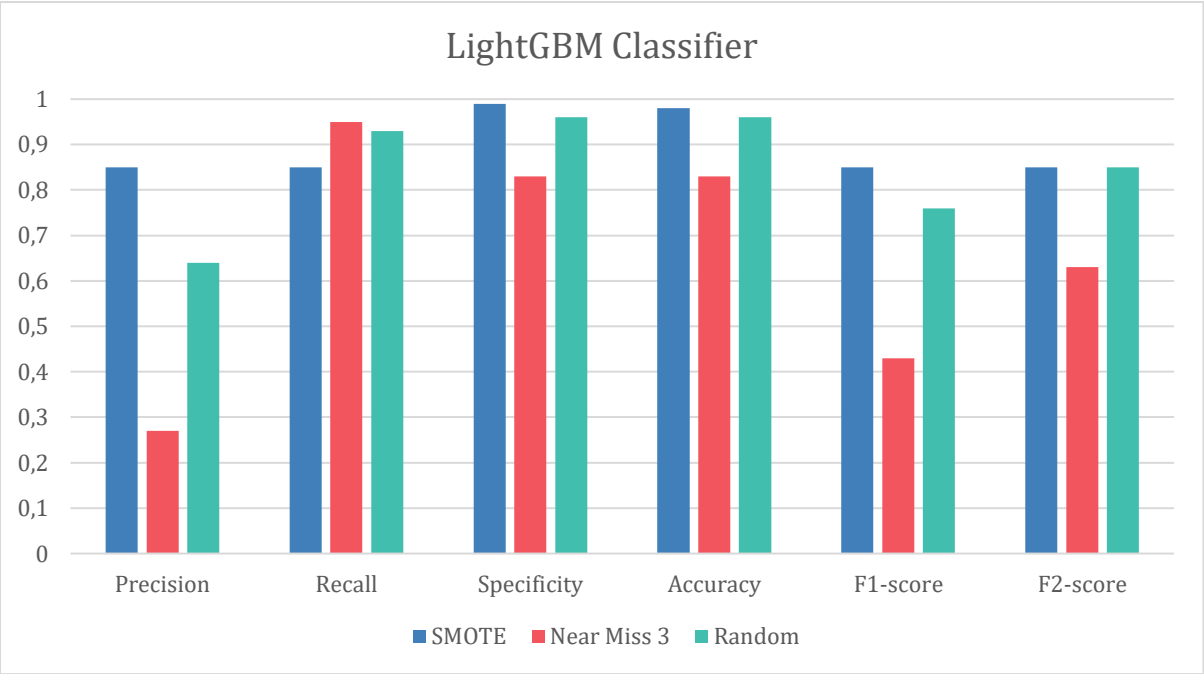
Για να καταλήξουμε σε συμπεράσματα για τη διάταξη που προσφέρει τα καλύτερα αποτελέσματα, δοκιμάζονται πολλές διαφορετικές επιλογές για τις παραμέτρους του συστήματος. Αρχικά, όπως είδαμε στην ανάλυση των δεδομένων, το σύνολο δεδομένων δεν είναι ισορροπημένο και, συνεπώς, θα χρειαστεί η εφαρμογή κάποια τεχνικής oversampling ή under sampling. Η αξία της ισορρόπησης είναι μεγάλη, όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο. Έπειτα, πρέπει να καθοριστεί η χρονική περίοδος για την οποία έχουμε βέλτιστα αποτελέσματα, για παράδειγμα δεδομένα από τον τελευταίο χρόνο ή δεδομένα μόνο του τελευταίου μήνα. Ύστερα, εξετάζουμε αν η χρήση ενός υποσυνόλου των δεδομένων μπορεί να δώσει καλύτερα αποτελέσματα. Σε όλα αυτά τα πειράματα σημειώνουμε και ελέγχουμε την επίδοση των μοντέλων που έχουμε επιλέξει να χρησιμοποιήσουμε, και βάσει όλων αυτών των πειραμάτων καταλήγουμε στη μεθοδολογία που θα ακολουθήσουμε. Σημειώνουμε ότι τα αποτελέσματα που παρουσιάζονται παρακάτω αφορούν ταξινομητές στους οποίους έχει πραγματοποιηθεί αναζήτηση για τη βέλτιστη τιμή κάποιων υπερπαραμέτρων τους.

5.4.1 Ισορρόπηση συνόλου δεδομένων

Αρχικά, εξετάζουμε τη μέθοδο ισορρόπησης του συνόλου δεδομένων που αποφέρει τα καλύτερα αποτελέσματα. Στη συγκεκριμένη περίπτωση δοκιμάστηκαν η τεχνική SMOTE για oversampling των δεδομένων και οι τεχνικές Near Miss 3 και Random για το under sampling. Τα αποτελέσματα που παρουσιάζονται παρακάτω χρησιμοποιούν όλα τα δεδομένα, αν και τα πειράματα εκτελέστηκαν και για τα δεδομένα του προηγούμενου μήνα. Το πείραμα με το οποίο καταλήξαμε ότι όλα τα δεδομένα δίνουν καλύτερα αποτελέσματα περιγράφεται στην επόμενο παράγραφο. Παρακάτω παρουσιάζονται τα αποτελέσματα για επιλεγμένους ταξινομητές, που είχαν την καλύτερη επίδοση.



Σχήμα 24: Αποτελέσματα πειράματος ισορρόπησης δεδομένων για Random Forest Classifier



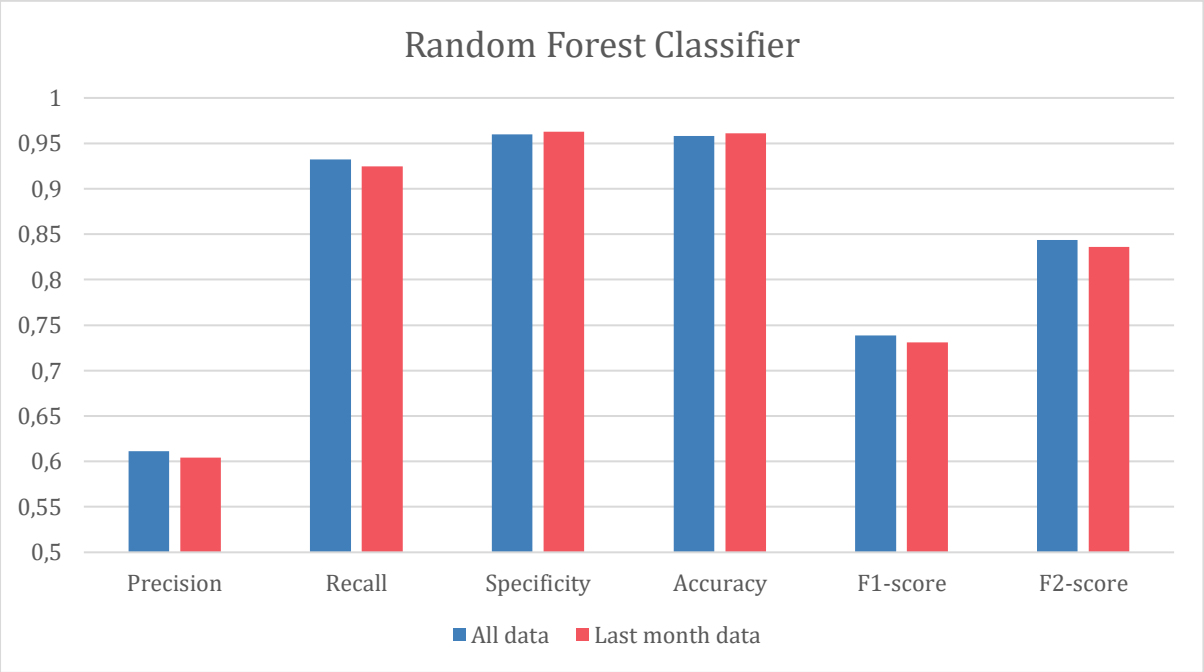
Σχήμα 25: Αποτελέσματα πειράματος ισορρόπησης δεδομένων για LightGBM Classifier

Από τα παραπάνω σχήματα, παρατηρούμε ότι η τεχνική Near Miss 3 έχει σημαντικά χειρότερα αποτελέσματα από τις υπόλοιπες, με εξαίρεση τη μετρική recall που παρουσιάζει αισθητά καλύτερα αποτελέσματα σε σχέση με την τεχνική SMOTE. Ωστόσο, η υπεροχή στη μετρική αυτή δεν είναι αρκετή για να υπερκαλύψει την επίδοσή της στις υπόλοιπες. Όσον αφορά τις υπόλοιπες δύο, δηλαδή τη SMOTE και το Random under sampling, βλέπουμε ότι είναι παρόμοιες, αλλά η διαφορά τους βρίσκεται στο γεγονός ότι η SMOTE έχει υψηλότερο precision, ενώ η Random σημειώνει υψηλότερο recall. Έτσι, προκύπτει μεγαλύτερο F2-score για το Random under sampling, ωστόσο η διαφορά στο precision είναι αξιοσημείωτη. Λαμβάνοντας υπόψη ότι το oversampling καθιστά τους χρόνους των υπολογισμών και εκπαίδευσης των μοντέλων απαγορευτικά μεγάλους σε μερικές περιπτώσεις, επιλέγουμε ως καλύτερη μέθοδο το Random under sampling για τη συνέχεια της πειραματικής διαδικασίας, καθώς το σύνολο δεδομένων είναι αρκετά μεγάλο. Ωστόσο, στην προτεινόμενη μεθοδολογία, αν το εισαγόμενο στο σύστημα σύνολο δεδομένων είναι επαρκώς μικρό βάσει των γραμμών και των στηλών του, χρησιμοποιείται η τεχνική SMOTE, καθώς τα αποτελέσματά της ήταν σχετικά καλύτερα και, ταυτόχρονα, δε χάνεται πληροφορία, όπως συμβαίνει με την αφαίρεση παραδειγμάτων από το σύνολο δεδομένων κατά την εκπαίδευση.

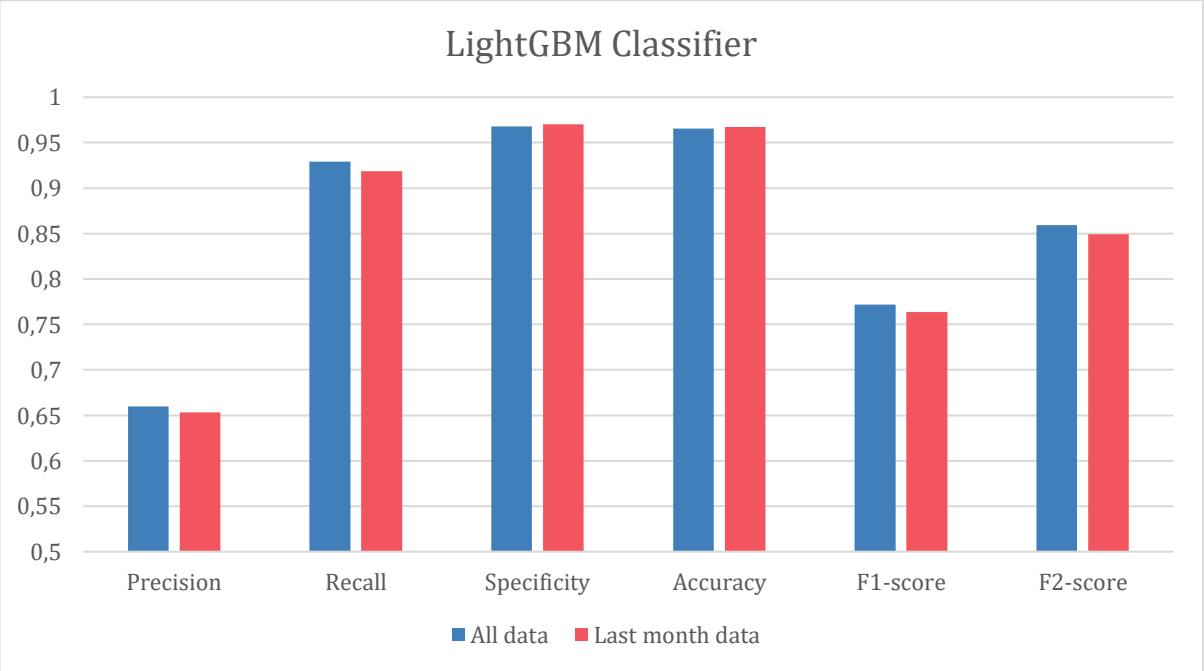
5.4.2 Χρονικό διάστημα δεδομένων

Στο πείραμα αυτό, ελέγχουμε αν είναι καλύτερο να χρησιμοποιήσουμε όλα τα δεδομένα που έχουμε στη διάθεσή μας, ή να τα περιορίσουμε θέτοντας κάποιο χρονικό όριο. Μεγάλο ποσοστό του συνόλου δεδομένων είναι ήδη από τον τρέχοντα χρόνο του διαγωνισμού, οπότε ο διαχωρισμός μεταξύ όλων των δεδομένων και δεδομένων του τελευταίου χρόνου δεν έχει σχεδόν καμία διαφορά στα αποτελέσματα. Έτσι, για να

υπάρχει κάποια διαφορά στον όγκο των παραδειγμάτων και, συνεπώς, στα αποτελέσματα της εκπαίδευσης, επιλέχθηκαν ως εναλλακτική τα δεδομένα του τελευταίου μήνα. Παρακάτω παρουσιάζονται τα αποτελέσματα για επιλεγμένους ταξινομητές, που είχαν την καλύτερη επίδοση. Όπως αναφέρθηκε και παραπάνω, το πείραμα αυτό εκτελέστηκε με Random under sampling.



Σχήμα 26: Αποτελέσματα πειράματος χρονικού διαστήματος δεδομένων για Random Forest Classifier



Σχήμα 27: Αποτελέσματα πειράματος χρονικού διαστήματος δεδομένων για LightGBM Classifier

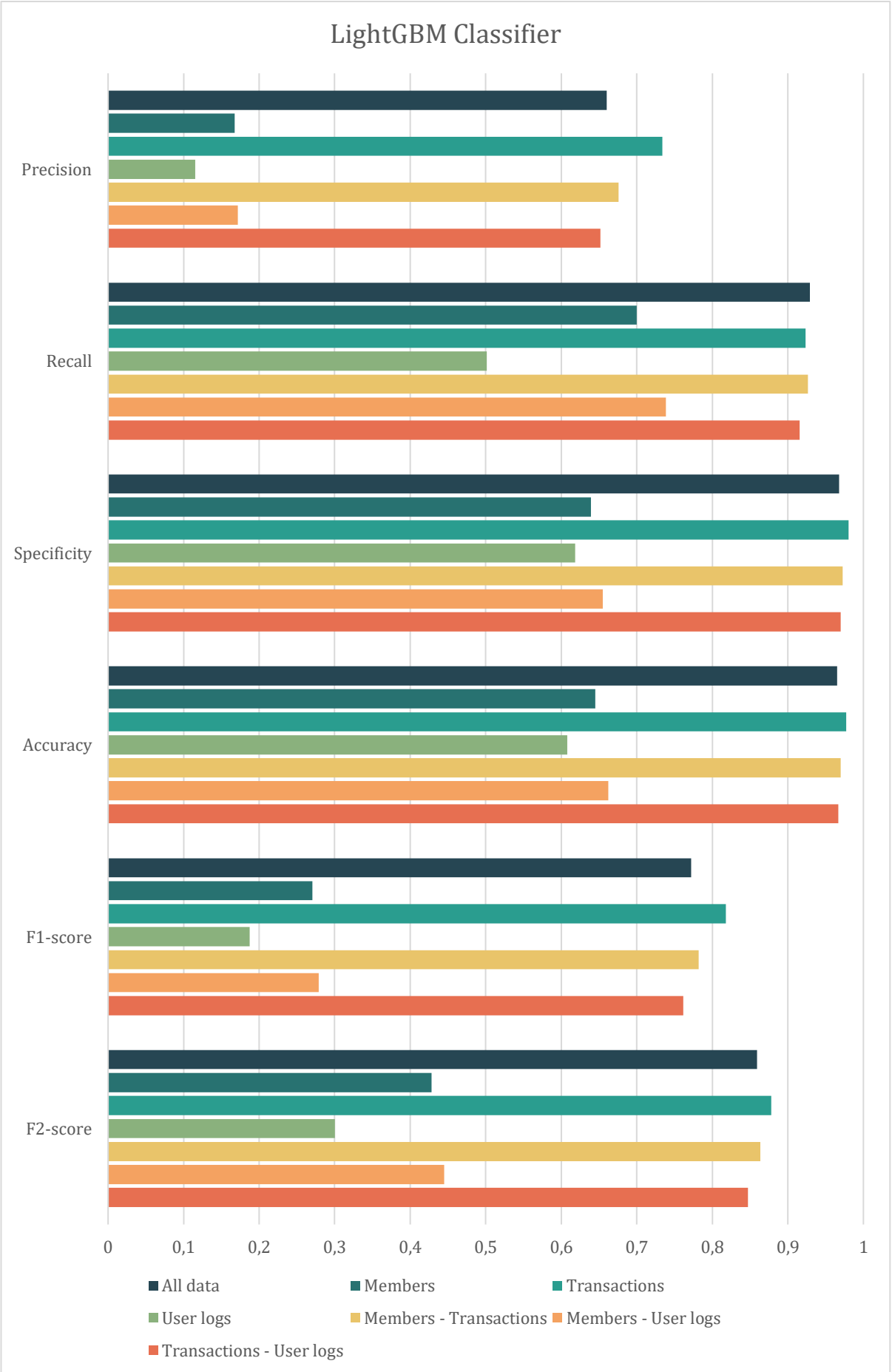
Από τα παραπάνω αποτελέσματα, βλέπουμε ότι σε όλες τις μετρικές όλα τα διαθέσιμα δεδομένα παρουσιάζουν ελαφρώς καλύτερες τιμές σε σχέση με τα δεδομένα

του τελευταίου μήνα. Αυτή η αύξηση στις μετρικές δεν είναι απαραίτητα μεγάλη αλλά, αφού οι υπολογιστικοί χρόνοι δεν επηρεάζονται σημαντικά και είναι καλή πρακτική να μην αφαιρούμε παραδείγματα για την καλύτερη εκπαίδευση των μοντέλων, είναι προτιμότερο να επιλέξουμε να μην περιορίσουμε τα δεδομένα σε κάποιο χρονικό διάστημα.

5.4.3 Τύπος δεδομένων – υποσύνολο συνόλου δεδομένων

Το επόμενο πείραμα που εκτελέσαμε αφορά την επιλογή των τύπων των δεδομένων που είναι καλύτεροι για την εκπαίδευση των μοντέλων πρόβλεψης. Στο συγκεκριμένο σύνολο δεδομένων που διαθέτουμε, έχουμε τους τρεις πίνακες που περιέχουν δημογραφικά δεδομένα, δεδομένα των συναλλαγών που έχουν πραγματοποιηθεί μέσω της υπηρεσίας, και δεδομένα χρήσης της υπηρεσίας streaming μουσικής, αντίστοιχα. Συμπεριλαμβανομένης της συγχώνευσης όλων των πινάκων, οι πιθανοί συνδυασμοί που προκύπτουν είναι επτά. Δοκιμάζουμε τους διαφορετικούς αυτούς συνδυασμούς για να καταλήξουμε σε εκείνον που δίνει τα καλύτερα αποτελέσματα και, ταυτόχρονα, να βγάλουμε κάποια συμπεράσματα για το ποια δεδομένα φαίνονται πιο σημαντικά για την εκπαίδευση. Τα παρακάτω πειράματα εκτελέστηκαν με Random under sampling και τη χρήση όλων των δεδομένων, χωρίς χρονικό όριο. Παρουσιάζονται τα αποτελέσματα για τον ταξινομητή LightGBM, που φάνηκε να έχει καλή απόδοση και μικρούς υπολογιστικούς χρόνους.

Από το παρακάτω διάγραμμα, είναι φανερό ότι τα σημαντικότερα δεδομένα είναι αυτά που αφορούν τις συναλλαγές του χρήστη. Πιο συγκεκριμένα, παρατηρούμε ότι τη χειρότερη επίδοση έχουν τα user logs και, ύστερα, τα δημογραφικά δεδομένα των πελατών. Ωστόσο, παρατηρούμε ότι σε μερικές μετρικές, όπως στο recall, ο συνδυασμός των υποσυνόλων δεδομένων αποφέρει καλύτερα αποτελέσματα από μεμονωμένους πίνακες. Για το λόγο αυτό, και για να μην πετάξουμε δεδομένα εκπαίδευσης που μπορεί να χρησιμεύσουν στη γενίκευση και, συνεπώς, στις καλύτερες προβλέψεις των μοντέλων, επιλέγουμε να συνεχίσουμε να χρησιμοποιούμε όλα τα δεδομένα για το υπόλοιπο της πειραματικής διαδικασίας, αφού και οι υπολογιστικοί χρόνοι δεν επηρεάζονται σημαντικά.

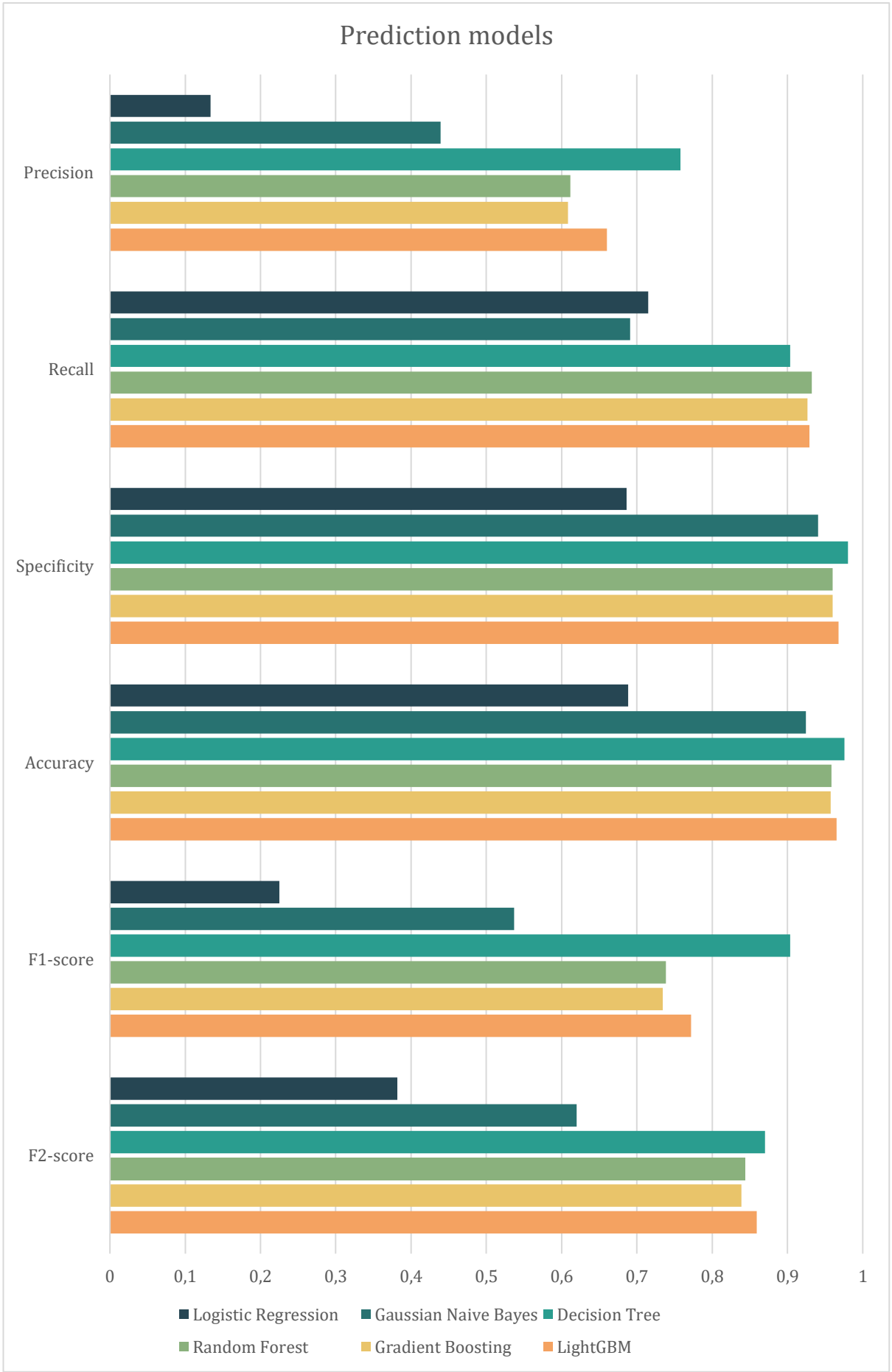


Σχήμα 28: Αποτελέσματα πειράματος συνδυασμών υποσυνόλων του data set

5.4.4 Μοντέλα πρόβλεψης

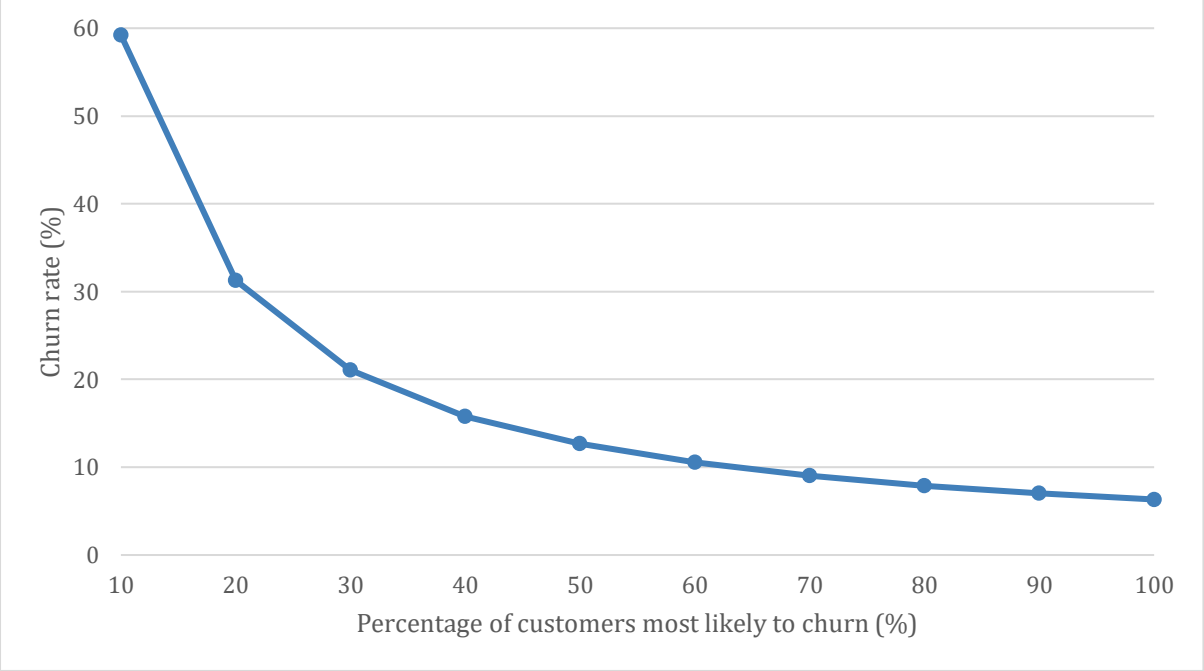
Το επόμενο πείραμα περιλαμβάνει τον έλεγχο διαφόρων ταξινομητών για την εύρεση αυτών που αρμόζουν καλύτερα στο συγκεκριμένο πρόβλημα, δηλαδή την πρόβλεψη της απώλειας πελατών. Εξετάστηκαν οι ταξινομητές Logistic Regression, Gaussian Naive Bayes, Decision Tree, Random Forest, Gradient Boosting, και LightGBM. Το πείραμα αυτό διεξάχθηκε με τα συμπεράσματα των προηγούμενων πειραμάτων, δηλαδή χρησιμοποιήθηκε Random under sampling σε όλο το σύνολο δεδομένων, χωρίς χρονικούς περιορισμούς. Τα αποτελέσματα παρουσιάζονται παρακάτω, στο σχήμα 29.

Από την παρακάτω γραφική παράσταση φαίνεται ότι οι ταξινομητές Logistic Regression και Gaussian Naive Bayes δεν έχουν καλή επίδοση. Πιο συγκεκριμένα, υπάρχει σημαντική διαφορά μεταξύ αυτών και των υπόλοιπων ταξινομητών στη μετρική του precision και, κατά συνέπεια, στο F1 και F2-score. Παράλληλα, βλέπουμε ότι οι ταξινομητές που βασίζονται σε δέντρα αποφάσεων έχουν πολύ καλή απόδοση στο συγκεκριμένο πρόβλημα. Ειδικότερα, την καλύτερη επίδοση φαίνεται να παρουσιάζει ο Decision Tree, αλλά αυτό θα μπορούσε να είναι λόγω του συγκεκριμένου διαχωρισμού train και test set που έτυχε στο πείραμα. Αφού και οι υπόλοιποι ταξινομητές δέντρων αποφάσεων φαίνεται να έχουν καλή απόδοση, θα χρησιμοποιήσουμε συνδυασμό μοντέλων για την ανάπτυξη της προτεινόμενης μεθοδολογίας, όπως εξηγήθηκε στο προηγούμενο κεφάλαιο της εργασίας. Σημειώνουμε ότι η τελική επιλογή είναι οι Decision Tree, Random Forest, και LightGBM που είναι ταξινομητές καθένας από τους οποίους χρησιμοποιεί τα δέντρα απόφασης με διαφορετικό τρόπο, όπως εξηγήθηκε στο τρίτο κεφάλαιο, αφαιρώντας τον Gradient Boosting καθώς ο LightGBM αποτελεί ταχύτερη και καλύτερη έκδοση αυτού.



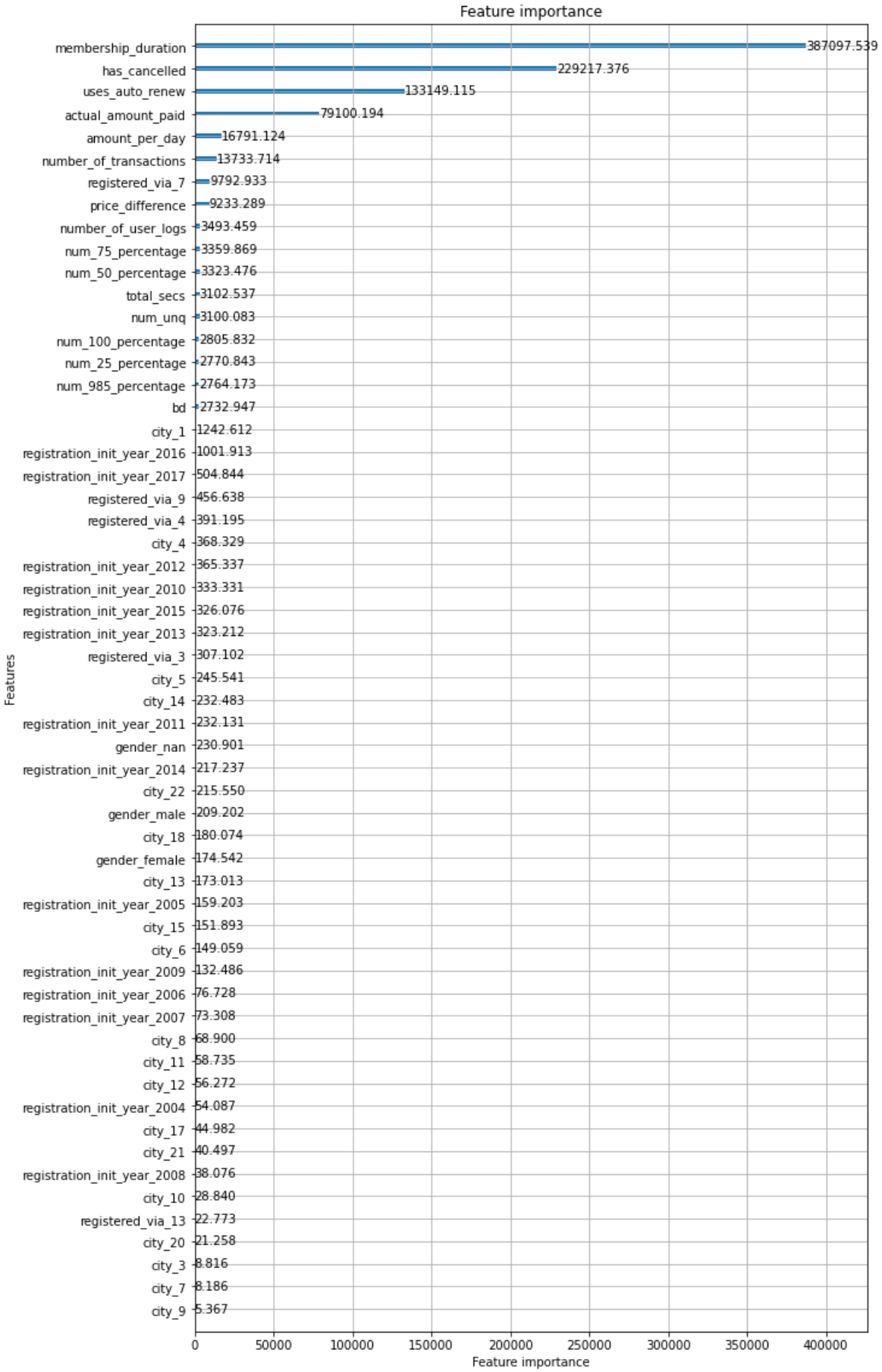
Σχήμα 29: Αποτελέσματα πειράματος διαφορετικών ταξινομητών

Πέρα από το παραπάνω πείραμα, από τα μοντέλα αυτά μπορούμε να παραγάγουμε κάποιες γραφικές παραστάσεις για την περαιτέρω κατανόηση του συνόλου δεδομένων και των αποτελεσμάτων. Η πρώτη γραφική παράγεται μέσω της ταξινόμησης των πιθανοτήτων σε φθίνουσα σειρά και, ύστερα, λαμβάνοντας κάθε φορά μεγαλύτερο ποσοστό πελατών και ελέγχοντας το churn rate αυτών. Μια καλή ένδειξη ότι οι προβλέψεις του μοντέλου έχουν καλή γενίκευση είναι πως ανάμεσα στους πελάτες τους οποίους το μοντέλο αναγνωρίζει ως πιο πιθανούς να κάνουν churn, το πραγματικό churn rate είναι πολύ υψηλό. Άρα, αν επιλεγεί ένα περιορισμένο ποσοστό πελατών με τις υψηλότερες πιθανότητες σύμφωνα με το μοντέλο, το churn rate του υποσυνόλου αναμένεται να είναι πολύ υψηλότερο από το churn rate στο σύνολο του δείγματος. Αντίστοιχα, καθώς το ποσοστό αυτό αυξάνεται, και στο υπό εξέταση υποσύνολο εντάσσονται πελάτες με μικρότερες πιθανότητες (σύμφωνα με το μοντέλο), το churn rate του υποσυνόλου αρχίζει να προσεγγίζει το churn rate όλου του πληθυσμού. Παρακάτω παρουσιάζεται αυτή η γραφική παράσταση, χρησιμοποιώντας το συνδυασμό μοντέλων Decision Tree και LightGBM, ο οποίος αναδείχθηκε ο καλύτερος στο πείραμα που παρουσιάζεται στην επόμενη παράγραφο.



Σχήμα 30: Γραφική παράσταση churn rate – ποσοστό πελατών

Με τη γραφική αυτή, μπορεί να αποφασιστεί το κατάλληλο ποσοστό των πελατών στο οποίο αξίζει να επενδυθούν χρήματα για τη διατήρηση των πελατών στην υπηρεσία και την πρόληψη του churn. Άλλη αξιοσημείωτη γραφική παράσταση είναι αυτή που αναδεικνύει τη σημασία των χαρακτηριστικών που χρησιμοποιήθηκαν για την εκπαίδευση, και συγκεκριμένα του μοντέλου LightGBM που παρέχει τη δυνατότητα δημιουργίας αυτής της γραφικής παράστασης. Για τον υπολογισμό της αξίας αυτής, υπολογίζει το συνολικό information gain των διαχωρισμών που χρησιμοποιούν το εκάστοτε χαρακτηριστικό. Η γραφική αυτή παρουσιάζεται παρακάτω, στο σχήμα 30.



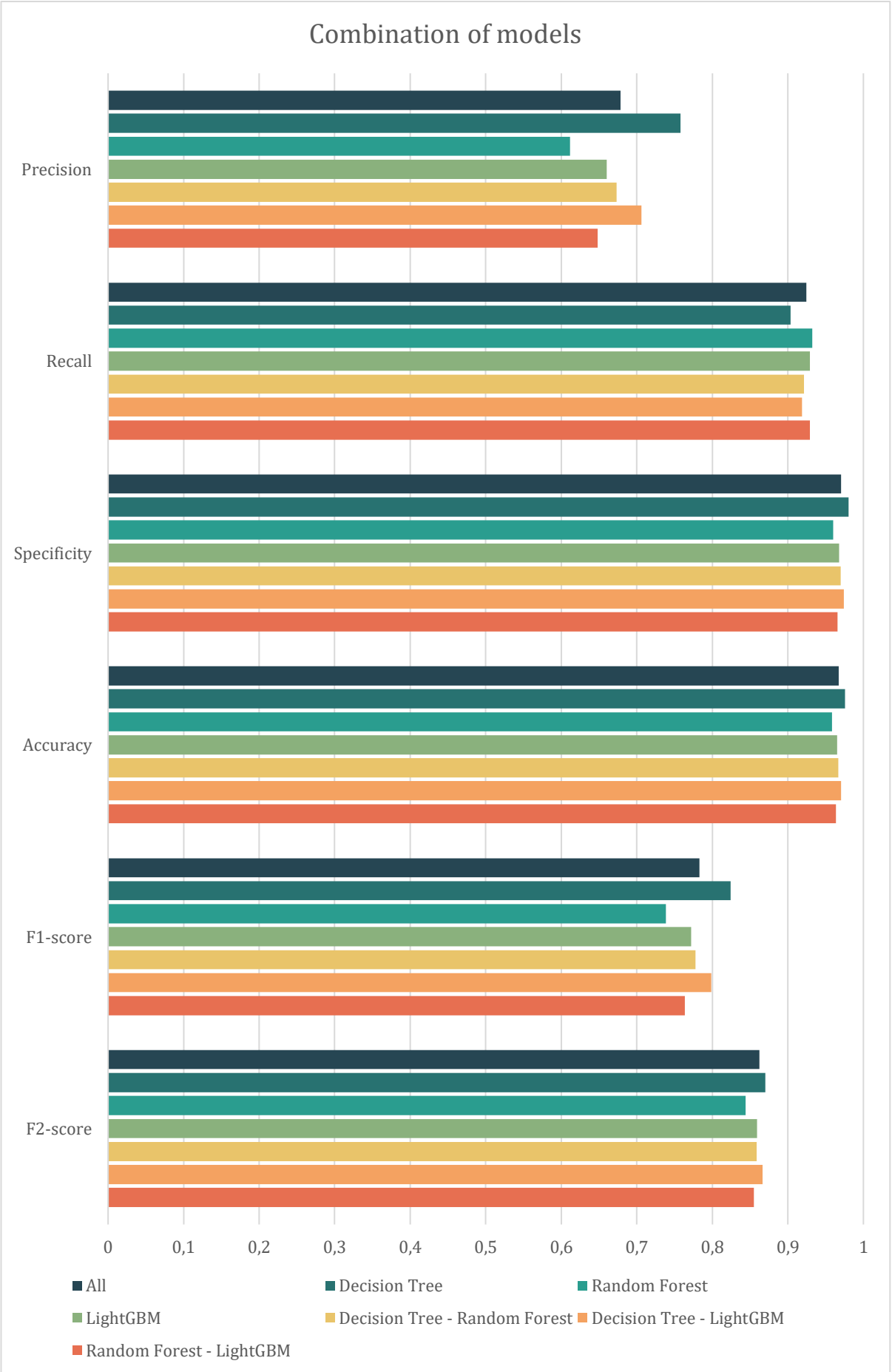
Σχήμα 31: Αξία των χαρακτηριστικών για την εκπαίδευση του LightGBM στην πειραματική διαδικασία

Από αυτή, βλέπουμε ότι πλέον έχουμε περαιτέρω ενδείξεις ότι τα χαρακτηριστικά του πίνακα συναλλαγών είναι τα σημαντικότερα, όπως είχαμε αποφανθεί προηγουμένως στο πείραμα με τους διαφορετικούς τύπους δεδομένων, καθώς τα έξι σημαντικότερα χαρακτηριστικά ανήκουν σε αυτόν. Τα τέσσερα σημαντικότερα χαρακτηριστικά με διαφορά, όπως αναμέναμε, είναι η συνολική διάρκεια συνδρομής του χρήστη, αν ο χρήστης έχει ακυρώσει τη συνδρομή του ποτέ, αν ο χρήστης γενικά έχει ενεργοποιημένη τη λειτουργία αυτόματης ανανέωσης της συνδρομής, και τα συνολικά χρήματα που έχει ξοδέψει ο χρήστης στην υπηρεσία.

5.4.5 Συνδυασμοί μοντέλων

Το επόμενο πείραμα περιλαμβάνει τον έλεγχο των συνδυασμών των ταξινομητών που επιλέχθηκαν από το προηγούμενο πείραμα, δηλαδή Decision Tree, Random Forest, και LightGBM, με σκοπό την καλύτερη γενίκευση του συστήματος παραγωγής προβλέψεων. Ο συνδυασμός προκύπτει από το μέσο όρο των πιθανοτήτων των συνδυαζόμενων μοντέλων και εφαρμόζοντας σε αυτόν κατώφλι 0.5. Εξετάστηκαν όλοι οι συνδυασμοί των παραπάνω ταξινομητών, οι οποίοι είναι επτά, συμπεριλαμβάνοντας τις περιπτώσεις που κάθε μοντέλο είναι μόνο του. Το πείραμα αυτό διεξάχθηκε με τα συμπεράσματα των προηγούμενων πειραμάτων, δηλαδή χρησιμοποιήθηκε Random under sampling σε όλο το σύνολο δεδομένων, χωρίς χρονικούς περιορισμούς. Τα αποτελέσματα παρουσιάζονται παρακάτω, στο σχήμα 32.

Στο σχήμα φαίνεται όλοι οι συνδυασμοί να έχουν ικανοποιητικά αποτελέσματα. Ωστόσο, ο Random Forest και οι συνδυασμοί που τον περιλαμβάνουν παρουσιάζουν ελαφρώς χειρότερη απόδοση. Ο ταξινομητής που φαίνεται να τα πήγε καλύτερα μεμονωμένα είναι ο Decision Tree, ωστόσο ο συνδυασμός Decision Tree και LightGBM παρουσιάζει γενικά τα καλύτερα αποτελέσματα. Έτσι, για την εξαγωγή του επόμενου πειράματος επιλέγουμε τη χρήση του συνδυασμού αυτών των δύο ταξινομητών, ωστόσο στην προτεινόμενη μεθοδολογία χρησιμοποιούνται και οι τρεις ταξινομητές, όπου και διεξάγεται αναζήτηση για τον καλύτερο συνδυασμό μεταξύ τους κάθε φορά. Αυτό γίνεται για τη μείωση των μεμονωμένων σφαλμάτων και την εξασφάλιση του καλύτερου μοντέλου πρόβλεψης ανάλογα με το σύνολο δεδομένων.

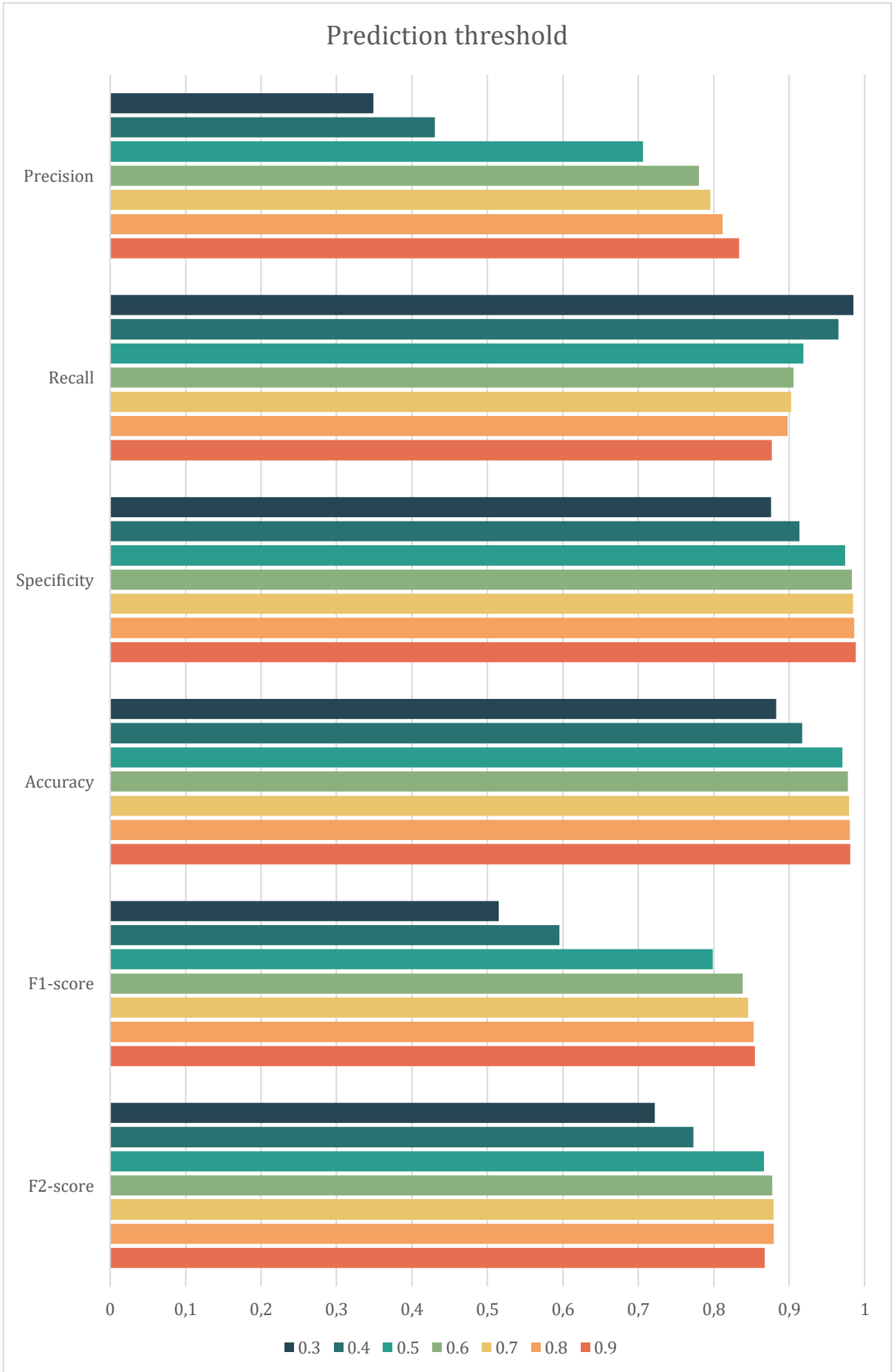


Σχήμα 32: Αποτελέσματα πειράματος συνδυασμών μοντέλων

5.4.6 Κατώφλια πρόβλεψης

Το τελευταίο πείραμα περιλαμβάνει τον έλεγχο των κατωφλιών πρόβλεψης. Το πιο συνηθισμένο κατώφλι πρόβλεψης είναι το 0.5, το οποίο χρησιμοποιείται και έως τώρα στην πειραματική διαδικασία. Με τη χρήση αυτού, οι πιθανότητες ουσιαστικά στρογγυλοποιούνται για να μετατραπούν σε προβλέψεις 0 ή 1, με όποια πιθανότητα μεγαλύτερη ή ίση του 0.5 να ανήκει στην κλάση 1, ενώ τις υπόλοιπες στην κλάση 0. Ωστόσο, η επιλογή διαφορετικού κατωφλίου μπορεί να αποφέρει πολύ καλύτερα και έμπιστα αποτελέσματα, καθώς αποτελεί τεχνική για καλύτερη γενίκευση του συστήματος. Το πείραμα αυτό διεξάχθηκε για κατώφλια από 0.3 έως και 0.9 με βήμα 0.1 στο συνδυασμό μοντέλων Decision Tree και LightGBM που φάνηκε να έχει τα καλύτερα αποτελέσματα. Επίσης, χρησιμοποιούνται τα συμπεράσματα των προηγούμενων πειραμάτων, δηλαδή χρησιμοποιήθηκε Random under sampling σε όλο το σύνολο δεδομένων, χωρίς χρονικούς περιορισμούς. Τα αποτελέσματα παρουσιάζονται παρακάτω, στο σχήμα 33.

Από το σχήμα, παρατηρούμε ότι τα κατώφλια 0.3 και 0.4, ενώ παρουσιάζουν μεγαλύτερο recall από τα υπόλοιπα, έχουν σημαντική διαφορά στις υπόλοιπες μετρικές. Κατώφλια μεγαλύτερα του 0.5 φαίνονται να έχουν καλύτερη επίδοση, με καλύτερο, λαμβάνοντας υπόψη τη γενική εικόνα, να αναδεικνύεται το 0.8.



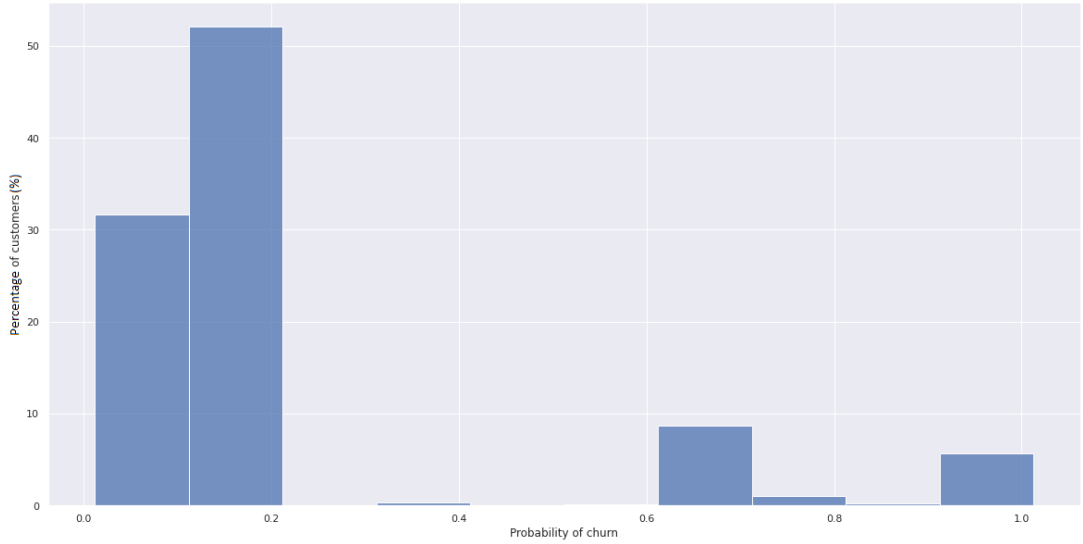
Σχήμα 33: Αποτελέσματα πειράματος κατωφλιών πρόβλεψης

5.5 Αποτελέσματα Plug & Play

Χρησιμοποιώντας τις πληροφορίες που αντλήσαμε από την πειραματική διαδικασία που αναλύθηκε προηγουμένως, δημιουργούμε το πλαίσιο Plug & Play, δηλαδή την προτεινόμενη μεθοδολογία αυτής της διπλωματικής εργασίας που περιεγράφηκε εκτενώς στο προηγούμενο κεφάλαιο. Σε αυτήν την παράγραφο θα αναλύσουμε τα αποτελέσματα που προέκυψαν από πειράματα που διεξάχθηκαν για τον έλεγχο λειτουργίας του Plug & Play. Χρησιμοποιήθηκαν δύο σύνολα δεδομένων: αυτό που χρησιμοποιήθηκε και στην πειραματική διαδικασία, δηλαδή τα δεδομένα χρήσης της υπηρεσίας streaming μουσικής που προσφέρει η KKBox Inc., και ένα νέο σύνολο δεδομένων που παράχθηκαν από την IBM και αφορά το churn των πελατών μιας πλασματικής εταιρείας τηλεπικοινωνιών. Το συγκεκριμένο data set μπορεί να βρεθεί στον ακόλουθο σύνδεσμο https://www.kaggle.com/datasets/blastchar/telco-customer-churn.

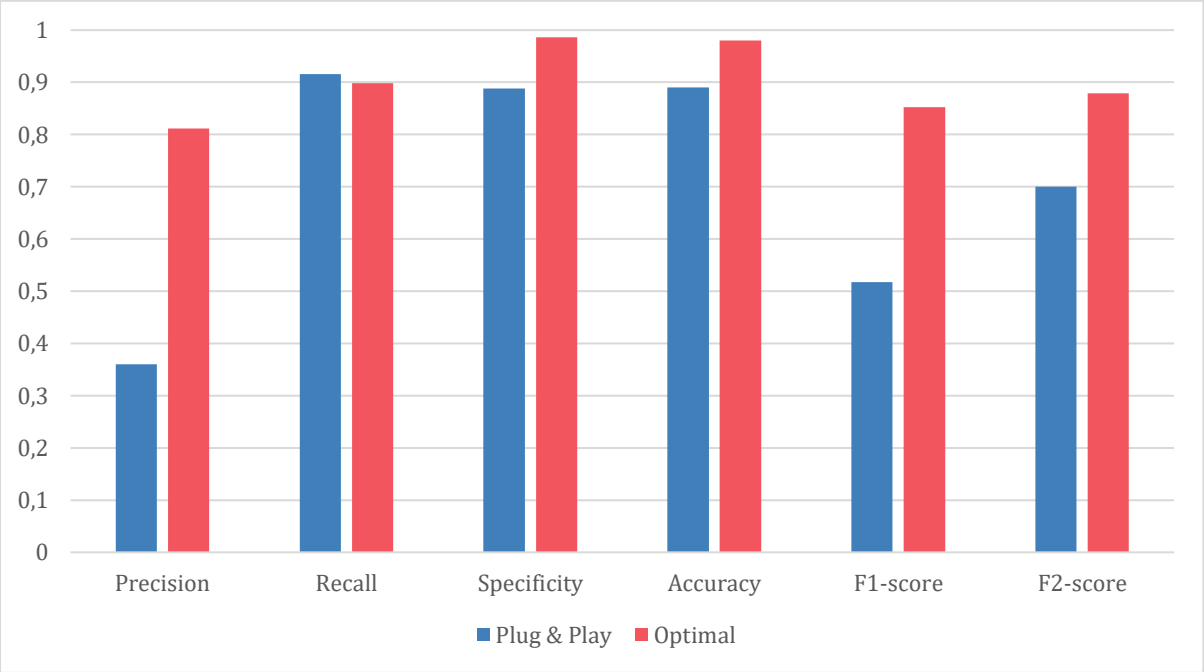
5.5.1 KKBox's Churn Prediction Challenge

Αρχικά, για τον έλεγχο λειτουργίας της προτεινόμενης μεθοδολογίας, διεξάχθηκε ένα πείραμα με το ίδιο σύνολο δεδομένων που χρησιμοποιήθηκε στην πειραματική διαδικασία. Έτσι, μπορούμε να συγκρίνουμε απευθείας τα αποτελέσματα και να εξάγουμε συμπεράσματα για την επίδοση του πλαισίου. Ως πρώτο μέτρο σύγκρισης, θα αναλύσουμε τις μετρικές που προέκυψαν από την εκτέλεση, με τις βέλτιστες μετρικές που προέκυψαν κατά την πειραματική διαδικασία. Σημειώνουμε ότι το βέλτιστο μοντέλο της πειραματικής διαδικασίας αναδείχθηκε ο συνδυασμός μεταξύ Decision Tree και LightGBM με κατώφλι πιθανοτήτων ίσο με 0.8, ενώ στο Plug & Play προέκυψε ότι το βέλτιστο μοντέλο στο συγκεκριμένο πείραμα είναι ταξινομητής Decision Tree με κατώφλι πιθανοτήτων ίσο με 0.2. Το ιστόγραμμα για τις προβλέψεις των πιθανοτήτων που παρήγαγε το μοντέλο του προτεινόμενου πλαισίου φαίνεται παρακάτω.



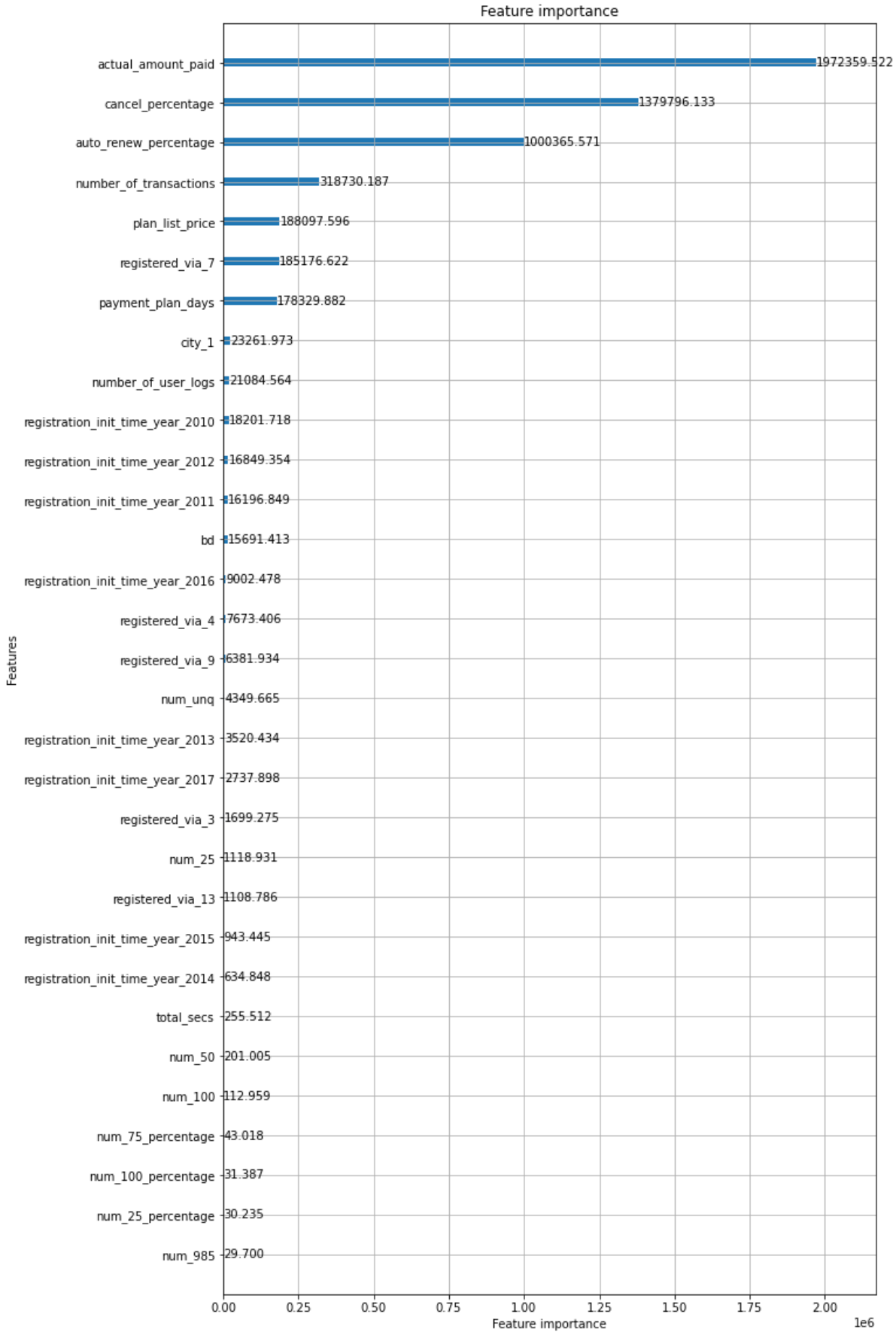
Σχήμα 34: Ιστόγραμμα πιθανοτήτων πειράματος Plug & Play με σύνολο δεδομένων το KKBox's Churn Prediction Challenge

Από το παραπάνω διάγραμμα είναι ξεκάθαρος ο λόγος επιλογής του κατωφλίου 0.2 ως βέλτιστος. Βλέπουμε ότι πάνω από 80% έχουν πιθανότητα churn λιγότερη από 0.2, ενώ τα επόμενα μεγαλύτερα σύνολα είναι αυτά μεταξύ των πιθανοτήτων 0.6 και 0.7, και 0.9 και 1. Παρακάτω παρουσιάζονται τα αποτελέσματα των μετρικών του πειράματος, σε σύγκριση αυτών που προέκυψαν στην πειραματική διαδικασία.



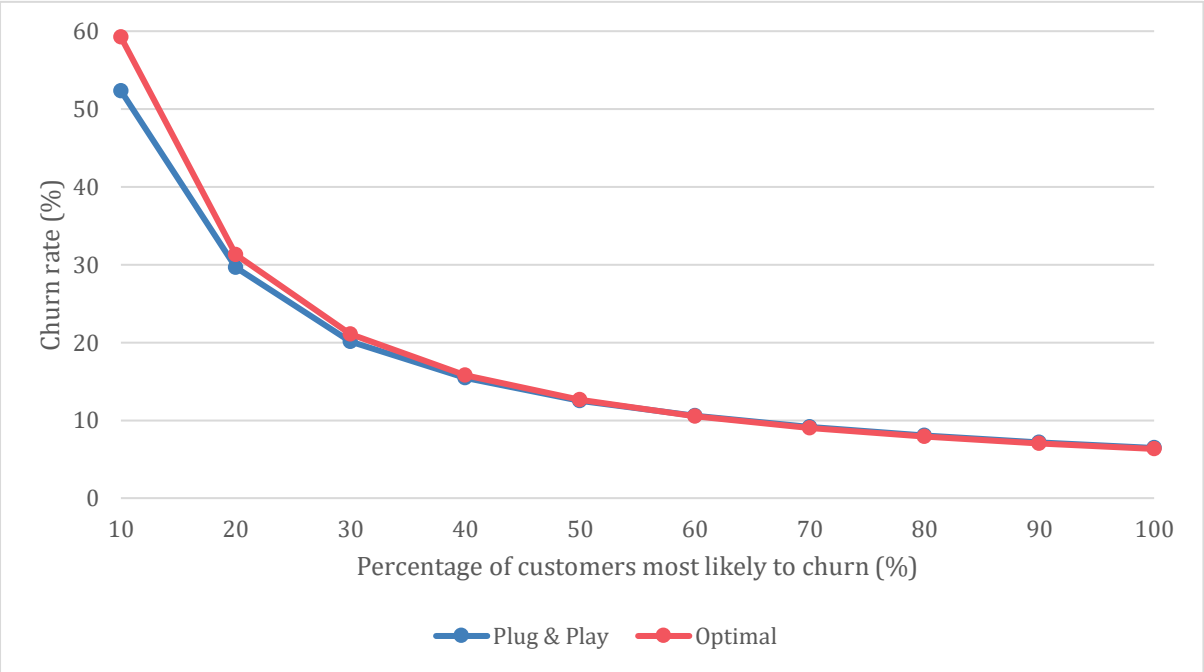
Σχήμα 35: Αποτελέσματα πειράματος Plug & Play με σύνολο δεδομένων το KKBox's Churn Prediction Challenge

Παρατηρούμε ότι το Plug & Play είναι σημαντικά χειρότερο στο precision και, κατά συνέπεια, στο F1-score από το βέλτιστο μοντέλο, γεγονός που αναμέναμε λόγω της μη δυνατότητας μέγιστης βελτιστοποίησης της προεπεξεργασίας των δεδομένων πριν την εκπαίδευσή τους. Αυτή η βελτιστοποίηση περιλαμβάνει κυρίως τη δημιουργία χαρακτηριστικών που προσφέρουν πολύ σημαντικότερη πληροφορία στα μοντέλα και τους επιτρέπει να εκπαιδεύονται πιο αποδοτικά. Για παράδειγμα, το "membership_duration" είχε αποδειχθεί το σημαντικότερο χαρακτηριστικό κατά την πειραματική διαδικασία, αλλά στο πείραμα αυτό δεν υπήρχε στη διάθεση των μοντέλων και, έτσι, εδώ έχουμε ως σημαντικότερο χαρακτηριστικό το συνολικό ποσό που έχει ξοδέψει ο χρήστης. Παρακάτω επισυνάπτεται ο πίνακας που παράγει το μοντέλο LightGBM, αν και δεν επιλέχθηκε ως βέλτιστο στο συγκεκριμένο πείραμα, που αναδεικνύει τη σημασία των χαρακτηριστικών κατά την εκπαίδευση.



Σχήμα 36: Αξία των χαρακτηριστικών για την εκπαίδευση του LightGBM στο πείραμα του Plug & Play με σύνολο δεδομένων το KKBox's Churn Prediction Challenge

Παρά την κακή επίδοση στη μετρική του precision, στο σχήμα 35 βλέπουμε ότι στις υπόλοιπες μετρικές το προτεινόμενο πλαίσιο φτάνει αρκετά κοντά και δίνει μια καλή πρώτη εικόνα για τις τιμές που πρέπει να αναμένουμε. Συγκρίνοντας και τις γραφικές παραστάσεις του churn rate και του ποσοστού πελατών στο σχήμα 37, βλέπουμε ότι δεν υπάρχει σημαντική διαφορά μεταξύ τους. Άρα, το προτεινόμενο πλαίσιο δίνει μια ορθή εικόνα των πραγμάτων, με την οποία μπορεί να αποφασιστεί το κατάλληλο ποσοστό των πελατών στο οποίο αξίζει να επενδυθούν χρήματα για τη διατήρηση των πελατών στην υπηρεσία και την πρόληψη του churn.



Σχήμα 37: Γραφική παράσταση churn rate – ποσοστό πελατών στο πείραμα Plug & Play με σύνολο δεδομένων το KKBox's Churn Prediction Challenge

5.5.2 Telco Customer Churn

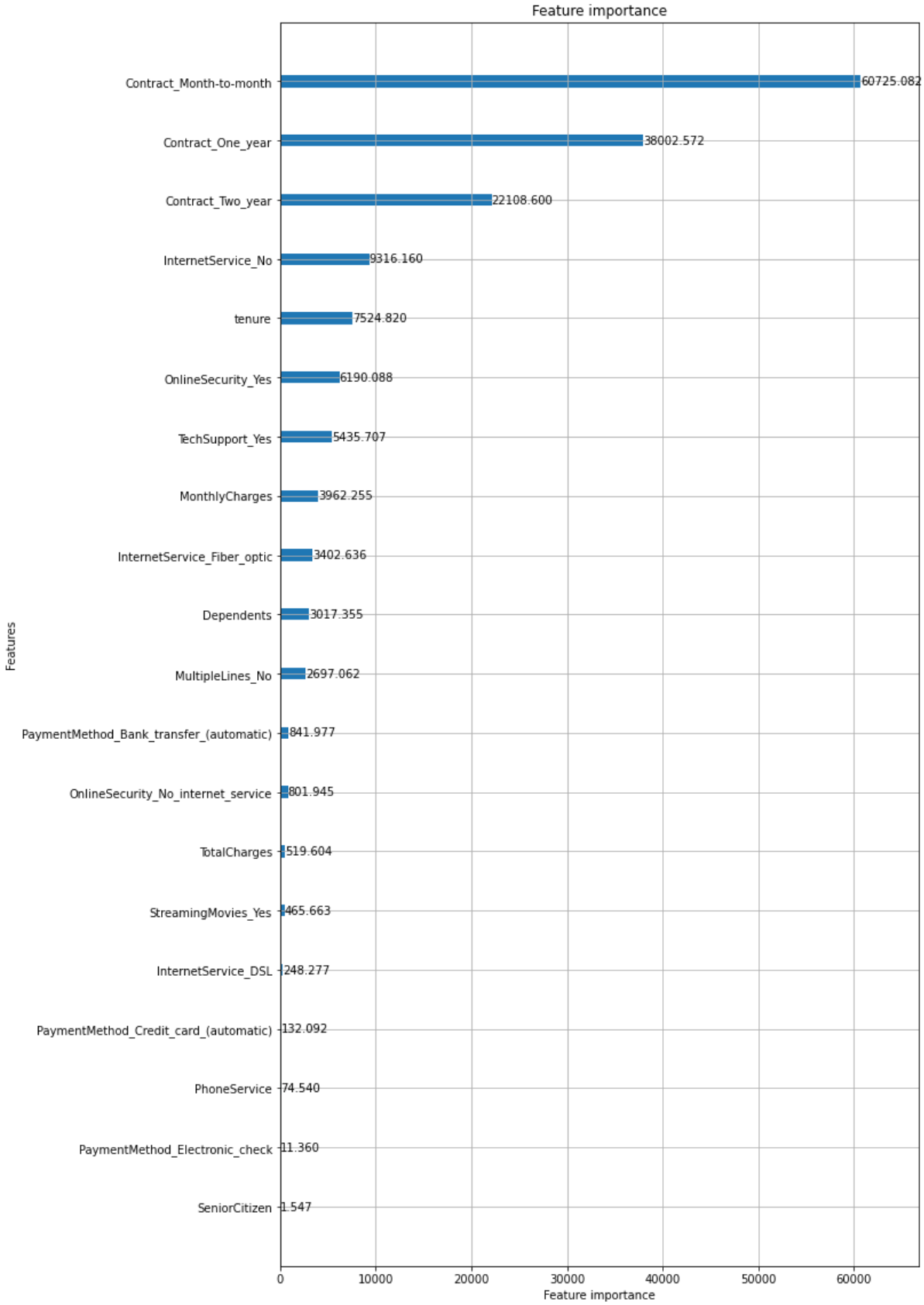
Αυτό το σύνολο δεδομένων αναπτύχθηκε από την IBM Business Analytics Community και περιέχει πληροφορίες μιας πλασματικής εταιρείας τηλεπικοινωνιών που παρείχε υπηρεσίες τηλεφωνίας και διαδικτύου σε 7,043 πελάτες στην Καλιφόρνια το Q3 του 2017. Παρουσιάζουμε στον παρακάτω πίνακα συνοπτικά τα χαρακτηριστικά του συνόλου.

Telco Customer Churn	
Χαρακτηριστικό	Περιγραφή
customerID	μοναδικό αναγνωριστικό χρήστη
gender	φύλο του χρήστη
SeniorCitizen	αν ο χρήστης είναι ηλικιωμένος
Partner	αν ο χρήστης έχει σύντροφο
Dependents	αν ο χρήστης έχει προστατευόμενα μέλη
tenure	αριθμός μηνών που ο χρήστης έχει παραμείνει με την εταιρεία

PhoneService	αν ο χρήστης έχει υπηρεσία τηλεφώνου
MultipleLines	αν ο χρήστης έχει πολλές γραμμές τηλεφωνίας
InternetService	τύπος υπηρεσίας διαδικτύου του χρήστη
OnlineSecurity	αν ο χρήστης έχει διαδικτυακή ασφάλεια
OnlineBackup	αν ο χρήστης έχει διαδικτυακό backup
DeviceProtection	αν ο χρήστης έχει προστασία συσκευών
TechSupport	αν ο χρήστης έχει τεχνική υποστήριξη
StreamingTV	αν ο χρήστης έχει streaming τηλεόρασης
StreamingMovies	αν ο χρήστης έχει streaming ταινιών
Contract	τύπος συμβολαίου του χρήστη
PaperlessBilling	αν ο χρήστης έχει αποϋλοποιημένη χρέωση
PaymentMethod	μέθοδος πληρωμής που χρησιμοποιεί ο χρήστης
MonthlyCharges	μηνιαίο ποσό που χρεώνεται ο χρήστης
TotalCharges	συνολικό ποσό χρέωσης του χρήστη
Churn	label για το churn

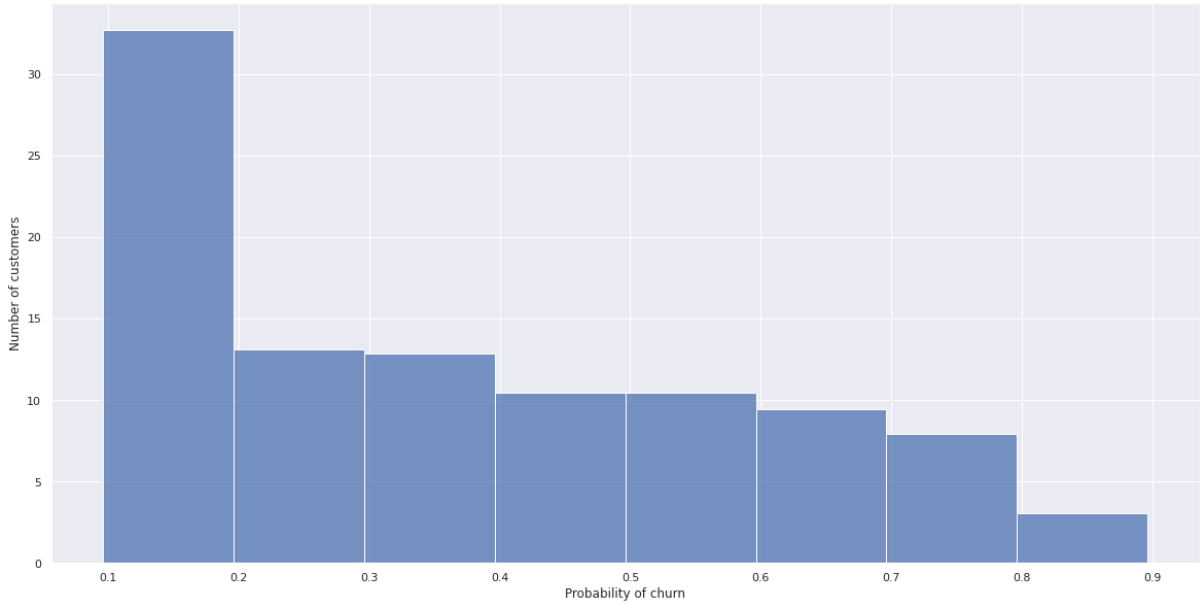
Πίνακας 11: Χαρακτηριστικά Telco Customer Churn

Τα περισσότερα χαρακτηριστικά που αφορούν τις υπηρεσίες που παρέχονται στο χρήστη λαμβάνουν κατηγορηματικές τιμές, καθώς δεν έχουν ως απάντηση μόνο «ναι» ή «όχι». Για παράδειγμα, το χαρακτηριστικό “MultipleLines” λαμβάνει τις τιμές “Yes”, “No”, “No phone service”. Για το λόγο αυτό, με την εισαγωγή τους στο Plug & Play, πολλά χαρακτηριστικά γίνονται one-hot encoded. Ύστερα, τα δεδομένα διαχωρίζονται σε train και test set, και γίνονται oversampled με SMOTE καθώς ήταν μη ισορροπημένα και το μέγεθός τους δεν υπερέβαινε το όριο που έχουμε θέσει για undersampling των δεδομένων. Ακολουθεί η εκπαίδευση των μοντέλων και η επιλογή του βέλτιστου συνδυασμού. Στο συγκεκριμένο πείραμα, το βέλτιστο μοντέλο προέκυψε με συνδυασμό των ταξινομητών Random Forest και LightGBM. Αξιοσημείωτη είναι η γραφική παράσταση που παράγει ο ταξινομητής LightGBM για την αξία των χαρακτηριστικών κατά την εκπαίδευσή του, όπως φαίνεται στο παρακάτω σχήμα. Βλέπουμε ότι τα σημαντικότερα χαρακτηριστικά είναι τα one-hot encoded χαρακτηριστικά που δημιουργήθηκαν από το “Contract”, δηλαδή τον τύπο συμβολαίου που έχει ο χρήστης. Είναι λογικό για χρήστες που έχουν μηνιαίο συμβόλαιο να έχουν αυξημένο churn rate, και για το λόγο αυτό το χαρακτηριστικό αυτό αποτέλεσε βασικό στοιχείο της εκπαίδευσης των μοντέλων.



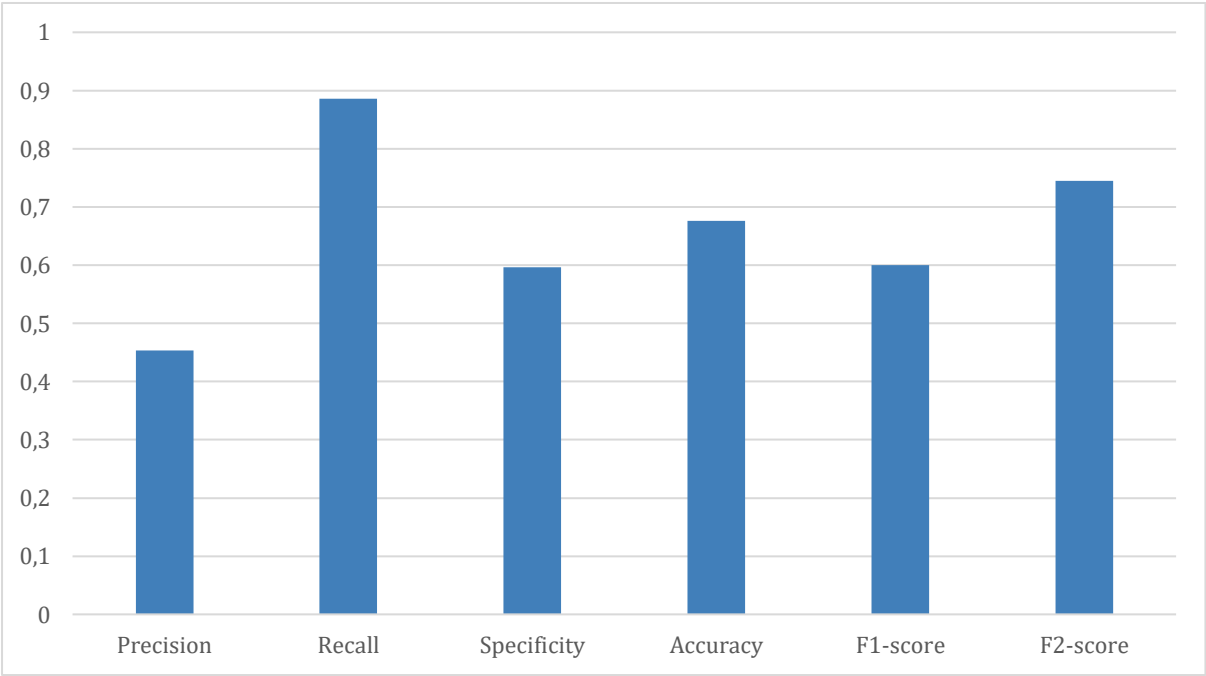
Σχήμα 38: Αξία των χαρακτηριστικών για την εκπαίδευση του LightGBM στο πείραμα του Plug & Play με σύνολο δεδομένων το Telco Customer Churn

Στη συνέχεια του πειράματος, έχουμε την επιλογή του βέλτιστου κατωφλίου πιθανοτήτων για το μοντέλο. Οι προβλέψεις που παρήγαγε το μοντέλο όσον αφορά τις πιθανότητες παρουσιάζονται στο παρακάτω διάγραμμα.

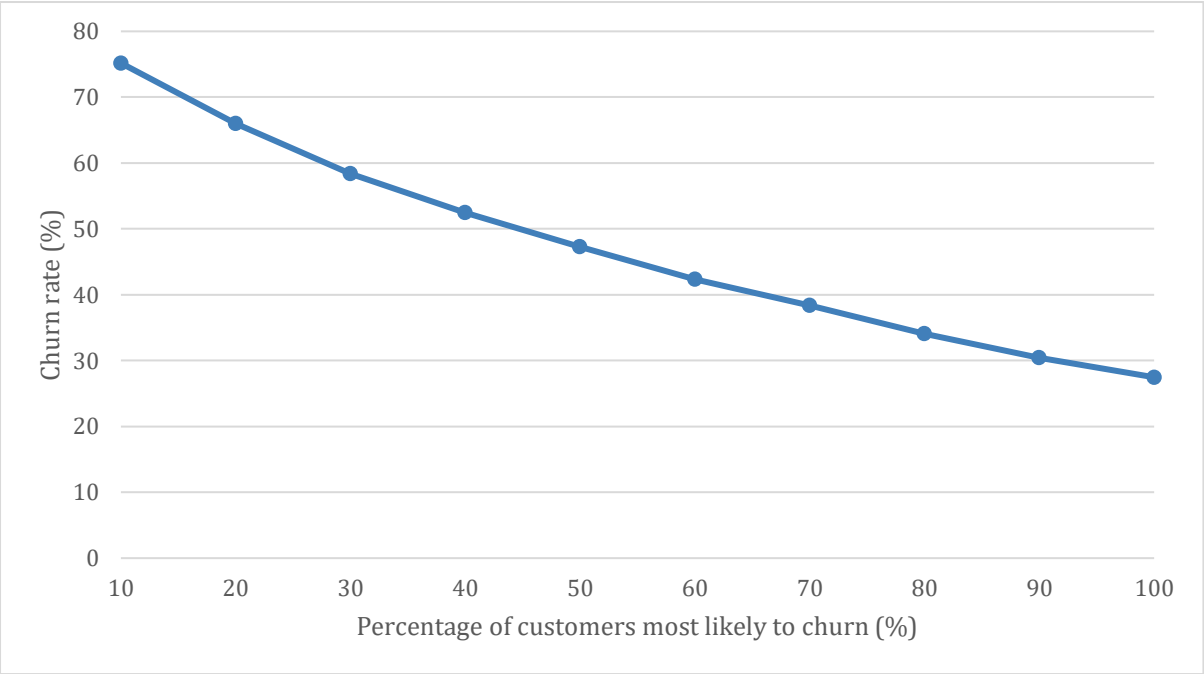


Σχήμα 39: Ιστόγραμμα πιθανοτήτων πειράματος Plug & Play με σύνολο δεδομένων το Telco Customer Churn

Βλέπουμε μεγάλη συγκέντρωση στο σύνολο πιθανοτήτων μεταξύ 0.1 και 0.2 και μετά μια σταδιακή μείωση όσο αυξάνεται η πιθανότητα. Το βέλτιστο κατώφλι βρέθηκε ίσο με 0.3. Τα αποτελέσματα που προέκυψαν από τη διεξαγωγή του πειράματος αυτού, καθώς και η γραφική παράσταση του churn rate βάσει ποσοστού πελατών, παρουσιάζονται στα παρακάτω σχήματα.



Σχήμα 40: Αποτελέσματα πειράματος Plug & Play με σύνολο δεδομένων το Telco Customer Churn



Σχήμα 41: Γραφική παράσταση churn rate – ποσοστό πελατών στο πείραμα Plug & Play με σύνολο δεδομένων το Telco Customer Churn

Από τις παραπάνω γραφικές παραστάσεις συμπεραίνουμε ότι το μοντέλο έχει ικανοποιητική επίδοση. Αξιοσημείωτη είναι η χαμηλή μετρική του precision, όπως παρατηρήθηκε και στο σύνολο δεδομένων που χρησιμοποιήθηκε στις προηγούμενες παραγράφους. Έτσι, επηρεάζεται σημαντικά και το F1-score, ωστόσο παρατηρούμε ότι το υψηλό recall που σημειώνει το μοντέλο συνεπάγεται ικανοποιητικό αποτέλεσμα και στο F2-score δεδομένης της μη βελτιστοποίησης των δεδομένων πριν την εκπαίδευση, που είναι και η μετρική στην οποία δίνουμε το μεγαλύτερο βάρος. Στη δεύτερη γραφική είναι εμφανής η πτώση του churn rate βάσει του ποσοστού των πελατών κάθε φορά, και, έτσι, αποδεικνύεται έμπιστη για την απόφαση του κατάλληλου ποσοστού πελατών στο οποίο αξίζει να επενδυθούν χρήματα για τη διατήρηση των πελατών στην υπηρεσία και την πρόληψη του churn της εταιρείας.

Κεφάλαιο 6. Συμπεράσματα και Προεκτάσεις

6.1 Συμπεράσματα

Η εφαρμογή της μηχανικής μάθησης και των νευρωνικών δικτύων σε προβλήματα ταξινόμησης είναι ένας κλάδος ιδιαίτερα διαδεδομένος τα τελευταία χρόνια. Αυτό οφείλεται κατά κύριο λόγο στη μεγάλη επιτυχία που έχουν οι τεχνικές αυτές στην επίλυση αυτών των προβλημάτων, καθώς και στο απεριόριστο εύρος τομέων στους οποίους εφαρμόζονται οι τεχνικές αυτές. Είναι, λοιπόν, λογικό επόμενο η προσπάθεια εφαρμογής της μηχανικής μάθησης και των νευρωνικών δικτύων στον κλάδο της απώλειας πελατών, με σκοπό την αξιοποίηση της ακρίβειας και ταχύτητας των αποτελεσμάτων που ο άνθρωπος δε μπορεί να συναγωνιστεί. Με μια πρώτη ματιά στα αποτελέσματα της πειραματικής διαδικασίας και της προτεινόμενης μεθοδολογίας γίνεται εύκολα σαφές ότι οι τεχνικές αυτές μπορούν να αποτελέσουν ιδιαίτερα αποδοτικό μέσο παραγωγής προβλέψεων για την απώλεια πελατών μιας εταιρείας, με τη βασική προϋπόθεση της κατάλληλης προεπεξεργασίας των δεδομένων και ρύθμισης συγκεκριμένων παραμέτρων.

Τα αποτελέσματα της παρούσας διπλωματικής εργασίας επιβεβαιώνουν την παραπάνω θέση και παρουσιάζουν ενδιαφέρον ως προς την πρόβλεψη του customer churn μιας εταιρείας. Συγκεκριμένα:

- Έγινε σαφής η ανωτερότητα των μοντέλων που βασίζονται στα δέντρα αποφάσεων στα προβλήματα ταξινόμησης. Πιο συγκεκριμένα, οι ταξινομητές Decision Tree, Random Forest, Gradient Boosting, και LightGBM που χρησιμοποιήθηκαν κατά την πειραματική διαδικασία είχαν καλύτερη επίδοση σε όλες τις μετρικές που εξετάστηκαν σε σχέση με τους ταξινομητές Logistic Regression και Gaussian Naive Bayes, ενώ ο δεύτερος από τους οποίους σημείωσε ικανοποιητικά αποτελέσματα για specificity και accuracy αλλά σημαντικά χειρότερα για τις υπόλοιπες μετρικές. Έτσι, επιβεβαιώνουμε ότι τα μοντέλα δέντρων αποφάσεων είναι καλύτερα σχεδιασμένα για τέτοια προβλήματα ταξινόμησης, λόγω του τρόπου λειτουργίας τους, που εξηγείται εκτενώς στο τρίτο κεφάλαιο της εργασίας.
- Αναδείχθηκε η ανωτερότητα της εφαρμογής oversampling σε ανισόρροπο σύνολο δεδομένων εκπαίδευσης, καθώς μέσα από τα πειράματα που διεξάχθηκαν προέκυψε ότι η τεχνική oversampling SMOTE που χρησιμοποιήθηκε σημειώνει σημαντικά καλύτερη απόδοση ως προς τη μετρική precision σε σχέση με τη μέθοδο Random under sampling. Αυτό ήταν αναμενόμενο, καθώς εφαρμόζοντας κάποια τεχνική under sampling αφαιρούμε παραδείγματα από το σύνολο δεδομένων που θα χρησιμοποιούνταν για την εκπαίδευση του μοντέλου και, συνεπώς, το μοντέλο αυτό δε μπορεί να επιτύχει τη βέλτιστη γενίκευση. Ωστόσο, είναι σημαντικό να σημειωθεί το γεγονός ότι η τεχνική του oversampling είναι υπολογιστικά απαιτητική, ειδικά αν το σύνολο δεδομένων είναι ήδη μεγάλο, δηλαδή περιέχει πολλά παραδείγματα (γραμμές) και χαρακτηριστικά (στήλες). Έτσι, κατά τη δημιουργία οποιουδήποτε μοντέλου πρόβλεψης με σύνολο

δεδομένων που δεν είναι ισορροπημένο μεταξύ των κλάσεών του, πρέπει κανείς να σκεφτεί το trade-off μεταξύ σχετικά καλύτερης επίδοσης που προσφέρει το oversampling και χαμηλότερου χρόνου εκπαίδευσης που προσφέρει το under sampling.

- Έγινε σαφής η υπεροχή της δημιουργίας νέων χαρακτηριστικών, βασισμένων στα χαρακτηριστικά του συνόλου δεδομένων που χρησιμοποιείται, για τη βέλτιστη απόδοση των μοντέλων πρόβλεψης όσον αφορά τα προβλήματα απώλειας πελατών. Πιο συγκεκριμένα, είδαμε ότι τα αποτελέσματα της πειραματικής διαδικασίας που διεξάχθηκε στην παρούσα διπλωματική εργασία είναι καλύτερα από αυτά που προέκυψαν εφαρμόζοντας το προτεινόμενο πλαίσιο στο ίδιο σύνολο δεδομένων. Ο κυριότερος λόγος για το φαινόμενο αυτό, πέρα από την τυχαιότητα των πειραμάτων όσον αφορά τα παραδείγματα που θα αφαιρέσει ο Random under sampler και το διαχωρισμό των δεδομένων σε train και test set, είναι ότι κατά την πειραματική διαδικασία δημιουργήθηκαν χαρακτηριστικά τα οποία κρίθηκαν ως τα σημαντικότερα για την εκπαίδευση των μοντέλων. Μέσα από τη γραφική παράσταση της σημαντικότητας των χαρακτηριστικών που μπορεί να παραγάγει ο ταξινομητής LightGBM, φάνηκε ότι το χαρακτηριστικό που προσφέρει το μεγαλύτερο information gain στο μοντέλο (διπλάσιο σε σχέση με το δεύτερο στη σειρά) είναι χαρακτηριστικό που δημιουργήσαμε χρησιμοποιώντας ήδη υπάρχοντα και εκφράζει τη διάρκεια της συνδρομής του χρήστη σε μέρες. Γενικεύοντας το φαινόμενο αυτό, συμπεραίνουμε ότι οι προσπάθειες αυτοματοποίησης προσφέρουν μια πολύ καλή προσέγγιση των αποτελεσμάτων του προβλήματος ταξινόμησης, ωστόσο για τη βέλτιστη απόδοση των μοντέλων απαιτείται η ειδικευμένη εξέταση και επεξεργασία των δεδομένων, δημιουργώντας νέα χαρακτηριστικά που προσφέρουν ουσιαστικότερη πληροφορία στα μοντέλα πρόβλεψης.
- Αναδείχθηκε ο τύπος δεδομένων του πελάτη που σχετίζεται πιο άμεσα με το churn και το επηρεάζει περισσότερο. Πιο συγκεκριμένα, παρατηρήθηκε ότι τα δημογραφικά δεδομένα και τα δεδομένα χρήσης της υπηρεσίας δεν ήταν τόσο σημαντικά για την εκπαίδευση των μοντέλων πρόβλεψης και, κατά συνέπεια, την ταξινόμηση του πελάτη ως churner ή όχι. Αντιθέτως, τα δεδομένα που αφορούν τις συναλλαγές και τη συνδρομή του χρήστη αποδείχθηκαν τα σημαντικότερα στα σύνολα δεδομένων που εξετάστηκαν κατά τη διάρκεια της πειραματικής διαδικασίας. Ωστόσο, κανείς θα ανέμενε τα δεδομένα χρήσης της υπηρεσίας από τον πελάτη να είναι άμεσα συνδεδεμένα με το ενδεχόμενο να απομακρυνθεί από την εταιρεία, καθώς αν ο πελάτης δε χρησιμοποιεί την εφαρμογή αρκετά θα ήταν λογικό επόμενο να αναθεωρήσει τη συνδρομή του σε αυτή και πιθανώς να τη διακόψει. Ένας λόγος για τον οποίο πιθανώς να συμβαίνει το αντίθετο είναι ότι οι χρήστες που χρησιμοποιούν την υπηρεσία για πολύ καιρό έχουν μικρότερη πιθανότητα να αλλάξουν πάροχο όσον αφορά την υπηρεσία αυτή, καθώς μπορεί να έχουν συνηθίσει σε αυτή και να μην επιθυμούν να αφιερώσουν το χρόνο για τη διαδικασία απεγγραφής και εξοικείωσης με νέα πλατφόρμα. Έτσι, δικαιολογείται η υψηλή αξία που απεύθυνε ο ταξινομητής LightGBM στα χαρακτηριστικά “membership_duration” του συνόλου δεδομένων του KKBox και “tenure” του Telco Customer Churn. Ταυτόχρονα, άλλος λόγος για την εξήγηση του φαινομένου

αυτού είναι η πιθανή παραμέληση του πελάτη για τη συνδρομή του στην υπηρεσία, είτε έχοντας υπογράψει συμβόλαιο για μεγάλο χρονικό διάστημα και χάνοντας το ενδιαφέρον του για την υπηρεσία, είτε έχοντας ενεργή την αυτόματη ανανέωση της μηνιαίας συνδρομής επιθυμώντας να έχει διαθέσιμη την επιλογή χρήσης της υπηρεσίας χωρίς απαραίτητα να τη χρησιμοποιεί σε μεγάλο βαθμό σε σχέση με άλλους χρήστες. Με τον τρόπο αυτό, εξηγείται η αξία που δόθηκε στα χαρακτηριστικά "uses_auto_renew" και τα one-hot encoded "Contract" στα δύο προαναφερθέντα σύνολα δεδομένων, αντίστοιχα.

Επίσης, από το σύνολο της μελέτης μπορούμε να συμπεράνουμε τα εξής για την εφαρμογή της προτεινόμενης μεθοδολογίας της παρούσας διπλωματικής από μια εταιρεία ή έναν οργανισμό:

- Δίνει τη δυνατότητα άμεσης μείωσης του churn rate της εταιρείας ή οργανισμού, μέσω της σωστής διαχείρισης των αποτελεσμάτων του προτεινόμενου πλαισίου, με την προσφορά εκπτώσεων ή κάποιου αντίστοιχου πλεονεκτήματος στους χρήστες τους οποίους τα μοντέλα προέβλεψαν ότι έχουν τη μεγαλύτερη πιθανότητα να κάνουν churn. Η μείωση του churn rate είναι αδιαμφισβήτητα σημαντική για μια εταιρεία ή έναν οργανισμό, ιδιαίτερα αν πρόκειται για παροχή συνδρομητικής υπηρεσίας. Αυτό ισχύει καθώς είναι γεγονός ότι η απόκτηση νέων πελατών αποτελεί μεγαλύτερη επένδυση από τη διατήρηση των ήδη υπαρχόντων. Με την απομάκρυνση ενός πελάτη από την υπηρεσία, η επένδυση που συνδέεται με την αρχική απόκτησή του χάνεται μαζί με τον πελάτη. Έτσι, η προσφορά κινήτρων για την παραμονή των πελατών στην υπηρεσία είναι σημαντική. Ταυτόχρονα, ένας πελάτης που ήδη χρησιμοποιεί την υπηρεσία που παρέχεται από την εταιρεία ή τον οργανισμό είναι πιο πιθανό να αγοράσει από την ίδια εταιρεία, δεδομένου ότι μπορεί να έχει αναπτύξει μια σχέση εμπιστοσύνης με αυτή ή να παρουσιάζει αδράνεια όσον αφορά την αλλαγή σε άλλο πάροχο αντίστοιχης υπηρεσίας.
- Δίνει τη δυνατότητα έμμεσης μείωσης του churn rate της εταιρείας ή οργανισμού, από την ανάλυση των αποτελεσμάτων και πληροφοριών που παρέχει το προτεινόμενο πλαίσιο με στόχο τη βελτίωση της προσφερόμενης υπηρεσίας. Η ανάλυση του customer churn μέσω της μηχανικής μάθησης προσφέρει στην εταιρεία ή τον οργανισμό, πέρα από ακριβείς προβλέψεις για τις πιθανότητες των χρηστών να κάνουν churn, τη δυνατότητα να κατανοήσουν τη συμπεριφορά των πελατών σε περιβάλλον χρήσης της υπηρεσίας. Μπορούν να αναλύσουν τη σημασία των χαρακτηριστικών των προϊόντων και υπηρεσιών που προσφέρουν και να εντοπίσουν αυτά που είναι άμεσα συνδεδεμένα με το churn. Με τον τρόπο αυτό, ανακαλύπτουν τις προτιμήσεις των χρηστών και τις αδυναμίες αυτών που προσφέρουν, πληροφορία που μπορούν να χρησιμοποιήσουν για να βελτιώσουν τις υπηρεσίες τους και να αποκτήσουν ένα ανταγωνιστικό πλεονέκτημα στην αγορά. Προσφέροντας ένα καλύτερο πακέτο σε σχέση με τον ανταγωνισμό, θα είναι πιο εύκολο για την εταιρεία ή τον οργανισμό να διατηρήσει τους ήδη υπάρχοντες πελάτες, αλλά και να ξεχωρίσει στην αγορά προσελκύοντας νέους.

6.2 Προεκτάσεις

Παρά την επιτυχία της προτεινόμενης μεθοδολογίας, λόγω των περιορισμών χρόνου και υπολογιστικών πόρων που περιλαμβάνονται στη διεξαγωγή μιας διπλωματικής εργασίας, υπάρχουν αρκετές βελτιώσεις που θα μπορούσε κανείς να εφαρμόσει με σκοπό την περαιτέρω αύξηση της απόδοσης του μοντέλου. Κάποιες από αυτές αποτελούν οι εξής:

- Επέκταση των χώρων αναζήτησης των υπερπαραμέτρων που χρησιμοποιούνται από τα μοντέλα πρόβλεψης και προσθήκη αναζήτησης βέλτιστης τιμής νέων υπερπαραμέτρων. Με τον τρόπο αυτό, σε συνδυασμό με την αύξηση των επαναλήψεων για τις οποίες εκτελείται η αναζήτηση, τα μοντέλα πρόβλεψης θα μπορούν να επιτύχουν βελτιωμένη απόδοση και, συνεπώς, καλύτερες τελικές προβλέψεις.
- Εξέταση περισσότερων μοντέλων πρόβλεψης για τον καλύτερο προσδιορισμό αυτών που λειτουργούν καλύτερα για το συγκεκριμένο πρόβλημα ταξινόμησης. Για παράδειγμα, θα μπορούσαν να χρησιμοποιηθούν ταξινομητές όπως AdaBoost (Adaptive Boosting) και XGBoost (eXtreme Gradient Boosting) που αποτελούν επεκτάσεις της τεχνικής boosting και είναι βασισμένοι στα δέντρα αποφάσεων, στα οποία βασίζονται και οι ταξινομητές που φάνηκαν να έχουν τις καλύτερες επιδόσεις κατά την πειραματική διαδικασία. Παράλληλα, θα μπορούσαν να εξεταστούν και τεχνικές νευρωνικών δικτύων όπως Support Vector Machines και perceptron πολλαπλών επιπέδων.
- Εφαρμογή της πειραματικής διαδικασίας σε περισσότερα σύνολα δεδομένων με ποικίλα χαρακτηριστικά. Με τον τρόπο αυτό, θα μπορέσουμε να αυτοματοποιήσουμε και να προσθέσουμε νέα στάδια στην προεπεξεργασία των δεδομένων του προτεινόμενου πλαισίου, με απώτερο σκοπό την καλύτερη γενίκευση των μοντέλων πρόβλεψης και, συνεπώς, τη βελτιωμένη απόδοση του συστήματος.
- Εξέταση περισσότερων τεχνικών oversampling και under sampling εντός της πειραματικής διαδικασίας. Για το oversampling, αυτές μπορεί να περιλαμβάνουν την τεχνική ADASYN (Adaptive Synthetic sampling) ή παραλλαγές του SMOTE που χρησιμοποιήθηκε στην παρούσα διπλωματική εργασία, όπως Borderline, KMeans, και SVM SMOTE. Παράλληλα, για το under sampling θα μπορούσε να εξεταστεί η λειτουργία τεχνικών όπως σύνδεσμοι Tomek ή κάποιον αλγόριθμο που βασίζεται σε nearest neighbors, για παράδειγμα condensed ή edited nearest neighbors.
- Ένταξη μεθοδολογίας στο προτεινόμενο πλαίσιο που επιλύει το πρόβλημα μεγιστοποίησης του κέρδους που περιεγράφηκε στο προηγούμενο κεφάλαιο. Με την επιλογή ενός περιορισμένου ποσοστού πελατών με τις υψηλότερες πιθανότητες να κάνουν churn σύμφωνα με το μοντέλο πρόβλεψης, το churn rate του υποσυνόλου αναμένεται να είναι πολύ υψηλότερο από το churn rate στο σύνολο του δείγματος. Έτσι, το πρόβλημα ανάγεται στην επιλογή του βέλτιστου ποσοστού πελατών για τη μεγιστοποίηση του κέρδους της εταιρείας, καθώς η επιλογή μικρότερου ποσοστού περιλαμβάνει στατιστικά περισσότερες απώλειες

πελατών ενώ η επιλογή μεγαλύτερου ποσοστού συνεπάγεται περισσότερα έξοδα σε ενέργειες μάρκετινγκ (προσφορές, εκπτώσεις) για τη διατήρηση πελατών που πιθανώς να μην έκαναν churn.

Βιβλιογραφία

[1] García, D.L., Nebot, À. & Vellido, A. (2016). Intelligent data analysis approaches to churn as a business problem: a survey. Knowledge and Information Systems, 51(3), p. 719-774

[2] Richeldi, M. & Perrucci, A. (2002). Churn Analysis Case Study. Telecom Italia Lab, 3-6

[3] Ahn, J., Hwang, J., Kim, D., Choi, H. & Kang, S. (2020). A Survey on Churn Analysis in Various Business Domains. IEEE Access, 8, p. 220816-220839

[4] Jain, H., Khunteta, A. & Srivastava, S. (2020). Telecom churn prediction and used techniques, datasets and performance measures: a review. Telecommunication Systems, 76, p. 613-630

[5] Mutanen, T. (2006). Customer churn analysis – a case study

[6] Naz, N.A., Shoaib, U. & Sarfraz, M.S. (2018). A Review on Customer Churn Prediction Data Mining Modeling Techniques. Indian Journal of Science and Technology, 11(27)

[7] Umayaparvathi, V. & Iyakutti, K. (2016). A Survey on Customer Churn Prediction in Telecom Industry: Datasets, Methods and Metrics. International Research Journal of Engineering and Technology (IRJET), 3(4)

[8] The Lift Curve: Unveiled, https://towardsdatascience.com/the-lift-curve-unveiled-998851147871, accessed on [2022-08-01]

[9] Alpaydin, E. (2014). Introduction to Machine Learning. The MIT Press. Third Edition

[10] McCulloch, W.S. & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 5

[11] Ivakhnenko, A. & Lapa, V. (1967). Cybernetics and forecasting techniques. American Elsevier Publishing Company

[12] Becker, S. (1991). Unsupervised Learning Procedures for Neural Networks. International Journal of Neural Systems, 1&2, p. 17-33

[13] Mitchell, T.M. (1997). Machine Learning. McGraw Hill

[14] Kotsiantis, S.B. (2007). Supervised Machine Learning: A Review of Classification Techniques. Informatica, 31, p. 249-268

[15] Cord, M. & Cunningham, P. (2008). Machine Learning Techniques for Multimedia. Springer

[16] Bishop, C.M. (2006). Pattern Recognition and Machine Learning. Springer

[17] Haykin, S. (2010). Νευρωνικά Δίκτυα και Μηχανική Μάθηση. Εκδόσεις Παπασωτηρίου. Τρίτη Έκδοση

[18] Liu, Q. & Wu, Y. (2012). Supervised Learning. Encyclopedia of the Sciences of Learning, p. 3243–3245

[19] Hammoudeh, A. (2018). A Concise Introduction to Reinforcement Learning

[20] Wang, H., Zariphopoulou, T. & Yu Zhou, X. (2019). Exploration versus exploitation in reinforcement learning: a stochastic control approach. arXiv preprint arXiv:1812.01552

[21] Dridi, S. (2021). Unsupervised Learning - A Systematic Literature Review

[22] Nwankpa, C.E., Ijomah, W., Gachagan, A. & Marshall, S. (2018). Activation Functions: Comparison of Trends in Practice and Research for Deep Learning. arXiv preprint arXiv:1811.03378

[23] Sazli, M.H. (2006). A Brief Review of Feed-Forward Neural Networks. Communications Faculty Of Science University of Ankara, 50(1), p. 11-17

[24] Schmidt, R.M. (2019). Recurrent Neural Networks (RNNs): A gentle Introduction and Overview. arXiv preprint arXiv:1912.05911

[25] Jurafsky, D. & Martin, J.H. (2021). Speech and Language Processing. Prentice Hall

[26] Logistic regression, https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression, accessed on [2022-08-23]

[27] Gaussian Naive Bayes, https://scikit-learn.org/stable/modules/naive_bayes.html#gaussian-naive-bayes, accessed on [2022-08-24]

[28] Bustamante, C., Garrido, L. & Soto, R. (2006). Comparing Fuzzy Naive Bayes and Gaussian Naive Bayes for Decision Making in RoboCup 3D. Mexican International Conference on Artificial Intelligence, Apizaco, Mexico, 13–17 November 2006, p. 237-247

[29] Sammut, C. & Webb, G.I. (2010). Encyclopedia of Machine Learning. Springer

[30] Decision Trees, https://scikit-learn.org/stable/modules/tree.html, accessed on [2022-08-25]

[31] Breiman, L. (2001). Random Forests. Machine Learning, 45, p. 5-32

[32] Subramaniam, P. & Kaur, M. (2019). Review of Security in Mobile Edge Computing with Deep Learning. 2019 Advances in Science and Engineering Technology International Conferences (ASET)

[33] Baturynska, I. & Martinsen, K. (2021). Prediction of geometry deviations in additive manufactured parts: comparison of linear regression with machine learning algorithms. Journal of Intelligent Manufacturing, 32

[34] Features, https://lightgbm.readthedocs.io/en/latest/Features.html, accessed on [2022-08-27]

[35] Shi, H. (2007). Best-first Decision Tree Learning

[36] Building a One Hot Encoding Layer with TensorFlow https://towardsdatascience.com/building-a-one-hot-encoding-layer-with-tensorflow-f907d686bf39, accessed on [2022-08-30]

[37] Cross-validation: evaluating estimator performance, https://scikit-learn.org/stable/modules/cross_validation.html, accessed on [2022-08-31]

[38] Hasan, M. (2007). Customer Churn: The Stealth Enemy. Sigillum Corporation

[39] Bloemer, J.M.M., Brijs, T., Vanhoof, K. & Swinnen, G. (2002). Comparing complete and partial classification for identifying customers at risk. International Journal of Research in Marketing, 20, p. 117-131

[40] Buckinx, W. & Van den Poel, D. (2004). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. European Journal of Operational Research 164, p. 252-268

[41] Bauer, H.H., Hammerschmidt, M. & Braehler, M. (2004). The Customer Lifetime Value Concept and its Contribution to Corporate Valuation

[42] Au, W.-H., Chan, K.C.C. & Yao, X. (2003). A Novel Evolutionary Data Mining Algorithm With Applications to Churn Prediction. IEEE Transactions on Evolutionary Computation, 7(6)

[43] Ahmed, A. & Linen, D.M. (2017). A review and analysis of churn prediction methods for customer retention in telecom industries. 4th International Conference on Advanced Computing and Communication Systems (ICACCS), p. 1-7

[44] Powers, D.M.W. (2011). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, 2 (1), p. 37-63

[45] Most Common Types of Machine Learning Problems, https://vitalflux.com/most-common-types-machine-learning-problems/, accessed on [2022-08-31]

[46] Russell, S.J. & Norvig, P. (2010). Artificial Intelligence: A Modern Approach. Prentice Hall. Third Edition

[47] Yang, Z.R. & Yang, Z. (2014). Comprehensive Biomedical Physics. Elsevier, p. 1

[48] Winston, P.H. (1992). Artificial Intelligence. Addison-Wesley Publishing Company. Third Edition

[49] Devroye, L., Gyorfi, L. & Lugosi, G. (1996). A probabilistic theory of pattern recognition. Springer

[50] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 31st Conference on Neural Information Processing Systems (NIPS 2017)