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

Μεθοδολογία πρόβλεψης της συμπεριφοράς καταναλωτών: Αγορά νέων προιόντων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανδρέας Αλέξανδρος Στ. Δελής

Επιβλέπων : Βασίλειος Ασημακόπουλος
Καθηγητής Ε.Μ.Π.

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

Αθήνα, Ιούνιος 2023

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

Μεθοδολογία πρόβλεψης της συμπεριφοράς καταναλωτών: Αγορά νέων προιόντων

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανδρέας Αλέξανδρος Στ. Δελής

Επιβλέπων : Βασίλειος Ασημακόπουλος
Καθηγητής Ε.Μ.Π.

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

Εγκρίθηκε από την τριμελή επιτροπή την _/_/2023

……	……	……
Βασίλειος	Ψαρράς	Δημήτριος
Ασημακόπουλος	Ιωάννης	Ασκούνης

Αθήνα, Ιούνιος 2023

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

.....
Ανδρέας Αλέξανδρος Δελής
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Ηλεκτρονικών Υπολογιστών

Ευχαριστίες

Η διπλωματική αυτή εργασία εκπονήθηκε στα πλαίσια των ερευνητικών δραστηριοτήτων της Μονάδας Προβλέψεων και Στρατηγικής κατά το ακαδημαϊκό έτος 2022 – 2023.

Αρχικά θέλω να ευχαριστήσω τον καθηγητή Βασίλειο Ασημακόπουλο για την ευκαιρία και την εμπιστοσύνη που μου έδειξε να εκπονήσω τη διπλωματική μου εργασία στη Μονάδα Προβλέψεων και Στρατηγικής, καθώς επίσης και για τη δυνατότητα που μου προσέφερε να ασχοληθώ σε βάθος με το αντικείμενο της πρόβλεψης της συμπεριφοράς καταναλωτών και της δημιουργίας συστημάτων προτάσεων με χρήση πρότυπων τεχνικών και μηχανικής μάθησης. Θα ήθελα ακόμα να ευχαριστήσω τον Καθηγητή κ. Ψαρρά Ιωάννη και τον Καθηγητή κ. Ασκούνη Δημήτριο για τη συμμετοχή τους στην επιτροπή εξέτασης της εργασίας.

Στη συνέχεια θα ήθελα και οφείλω να ευχαριστήσω θερμά τον ερευνητικό συνεργάτη και υποψήφιο Διδάκτορα Ε.Μ.Π. Αρτέμιο-Ανάργυρο Σεμένογλου, ο οποίος με καθοδηγούσε και με συμβούλευε καθ΄ όλη τη διάρκεια της εκπόνησης της εργασίας μου, με μεγάλη επιτυχία.

Με την παρουσίαση της διπλωματικής μου εργασίας και κλείνοντας έτσι το κεφάλαιο της τριτοβάθμιας εκπαίδευσης, δράττομαι της ευκαιρίας να αναβιώσω έστω και απο μνήμης όλες εκείνες τις στιγμές που με διαμόρφωσαν και φέρω μέσα μου. Το μεγαλύτερο μερίδιο επαίνων αξίζει και αποδίδω στην οικογένεια μου, τους γονείς μου Σταύρο και Λίλυ-Μαρία και τον δίδυμο αδερφό μου Χρίστο για την αμέριστη αγάπη, φροντίδα και καθοδήγηση τους. Χωρίς αυτούς, την στήριξη τους σε κάθε δυσκολία και την παρουσία τους σε κάθε στιγμή, η ζωή μου αυτά τα χρόνια θα είχε ένα μεγάλο ασυμπλήρωτο κενό. Για αυτό και τους ευχαριστώ βαθύτατα.

Τέλος δε θα μπορούσα να μην αναφερθώ στην πολύτιμη φιλία που σύναψα κατά τη διάρκεια των σπουδών μου και που ελπίζω να με ακολουθεί ως το πολύ μακρινό μέλλον, με τον φίλο και συμφοιτητή μου Θάνο, μαζί με τον οποίο μοχθήσαμε, ξενυχτήσαμε, διαφωνήσαμε και τελικώς υπερκεράσαμε κάθε πρόκληση ρίχνοντας την αυλαία μαζί.

Ανδρέας Αλέξανδρος Δελής

Αθήνα, Ιούνιος 2023.

Περίληψη

Η ανάγκη για έγκαιρες και ακριβείς προβλέψεις της συμπεριφοράς των καταναλωτών καθώς και η ερμηνεία και αξιολόγηση αυτών, αποτελεί πρόκληση για κάθε εταιρεία ή οργανισμό που επιθυμεί να ικανοποιεί τους πελάτες του αλλά και να προσελκύει νέους καταναλωτές. Η παρούσα διπλωματική εργασία έχει ως στόχο την μελέτη και ανάπτυξη μιας ευρύτερης μεθολογίας επίτευξης του σκοπού αυτού. Δίνοντας ιδιαίτερη βαρύτητα σε τεχνικές μηχανικής μάθησης αλλά και σε αλγορίθμους collaborative-filtering, διερευνούμε τους τρόπους σύστασης ενός άρτιου εργαλείου, το οποίο θα μπορεί να χρησιμοποιηθεί για να παράγει προβλέψεις και να πραγματοποιεί προτάσεις σε υφιστάμενους ή νεοεισερχόμενους πελάτες μιας επιχείρησης σχετικά με τις μελλοντικές ανάγκες και προτιμήσεις τους. Ιδιαίτερα τα τελευταία χρόνια όπου ο όγκος των δεδομένων και της πληροφορίας παρουσιάζει εκθετική αύξηση, τεχνικές όπως αυτές που θα αναλυθούν στη συνέχεια, συνιστούν ένα ιδιαίτερα σημαντικό εργαλείο που μπορεί να χρησιμοποιηθεί για την στοχευμένη χάραξη προωθητικών ενεργειών αλλά και την ανάδειξη νέων δυνατοτήτων και αναγκών.

Η βασική ιδέα της προτεινόμενης μεθοδολογίας είναι η αυτοματοποίηση της διαδικασίας παραγωγής προβλέψεων από σύνολα δεδομένων που αφορούν τη συμπεριφορά πελατών. Το προτεινόμενο πλαίσιο δέχεται ως είσοδο ένα σύνολο δεδομένων και παράγει ως έξοδο τις προβλέψεις για αυτά, καθώς και γραφικές παραστάσεις και μετρικές για την αξιολόγησή τους. Η μεθοδολογία αυτή περιλαμβάνει τέσσερα βήματα. Το πρώτο αφορά στην προεπεξεργασία των δεδομένων εισόδου και το μετασχηματισμό αυτών σε μορφή κατάλληλη για την καλύτερη εκπαίδευση των μοντέλων και αλγορίθμων πρόβλεψης. Στο δεύτερο βήμα πραγματοποιείται εκτενής διερεύνηση των τιμών των υπερπαραμέτρων των μοντέλων, που συνιστούν τον καλύτερο συνδυασμό και αποσκοπούν στη βελτιστοποίηση των παραγόμενων προβλέψεων. Στο τρίτο στάδιο, εξετάζονται και δοκιμάζονται οι συνδυασμοί που αναπτύχθηκαν για το εκάστοτε μοντέλο, ενώ καταγράφονται και επιλέγονται ως βέλτιστα τα μοντέλα με τα καλύτερα αποτελέσματα για τις διάφορες μετρικές επίδοσης. Τέλος, στο τέταρτο βήμα, αναλύεται και αξιολογείται το

συνολικό μοντέλο που προέκυψε, με χρήση των παραγόμενων γραφικών παραστάσεων και μετρικών.

Για την ανάπτυξη της μεθοδολογίας αυτής διεξήχθη μια εκτενής πειραματική διαδικασία, εξετάζοντας κάθε φορά διαφορετικές μεταβλητές και παραμέτρους του συστήματος. Για τα πειράματα χρησιμοποιήθηκαν δεδομένα του ομίλου επιχειρήσεων Santander Group ο οποίος αποτελεί μια Ισπανική πολυεθνική εταιρεία οικονομικών υπηρεσίων. Όλα τα δεδομένα προήλθαν από τον τραπεζικό τομέα του ομίλου και την Santander Bank, που περιλαμβάνει τα δεδομένα των χρηστών για τις ποικίλες υπηρεσίες που προσφέρει η ομώνυμη τράπεζα.

Λέξεις κλειδιά:

Μηχανική μάθηση, Πρόβλεψη συμπεριφοράς πελατών, Πρόβλεψη πιθανοτήτων, Συστήματα προτάσεων.

Abstract

The need for timely and accurate predictions of consumer behavior, as well as their interpretation and evaluation, is a challenge for every company or organization that aims to satisfy its customers and attract new consumers. This thesis aims to study and develop a broader methodology to achieve this goal. With special emphasis on machine learning techniques and collaborative filtering algorithms, we explore ways to recommend a comprehensive tool that can be used to generate predictions and make recommendations to existing or potential customers of a business regarding their future needs and preferences. Particularly in recent years, where the volume of data and information has experienced exponential growth, techniques such as those to be analyzed subsequently represent a particularly significant tool that can be used for targeted promotional actions, as well as the identification of new opportunities and needs.

The main idea of the proposed methodology is the automation of the prediction generation process from datasets concerning customer behavior. The proposed framework takes a dataset as input and produces predictions for it, as well as graphical representations and metrics for their evaluation. This methodology includes four steps. The first step involves preprocessing the input data and transforming it into a suitable format for the optimal training of prediction models and algorithms. In the second step, an extensive exploration of the hyperparameter values of the models is performed, aiming to optimize the generated predictions. The third step involves examining and testing the combinations developed for each model, while recording and selecting the models with the best results for various performance metrics. Finally, in the fourth step, the overall model that emerged is analyzed and evaluated using the generated graphical representations and metrics.

To develop this methodology, an extensive experimental process was conducted, examining different variables and system parameters each time. For the experiments, data from the Santander Group, a Spanish multinational financial services company, were used.

All the data originated from the banking sector of the group and Santander Bank, which includes user data for the various services offered by the bank.

Keywords:

Machine learning, Customer behavior prediction, Probability prediction, Recommendation systems.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Κεφάλαιο 1: Εισαγωγή

1.1 Αντικείμενο της εργασίας

Εν μέσω της ραγδαίας τεχνολογικής ανάπτυξης και ψηφιακής επανάστασης, η παγκοσμιοποίηση και η ανταγωνιστική φύση των αγορών έχουν δημιουργήσει μια σειρά προκλήσεων για τις επιχειρήσεις οι οποίες έρχονται συνεχώς αντιμέτωπες με την ανάγκη της εγκαθίδρυσής τους στη διαρκώς εναλλασσόμενη αγορά.

Η σύγχρονη πραγματικότητα υπαγορεύει πως μία από αυτές τις προκλήσεις είναι η ανάγκη πρόβλεψης της συμπεριφοράς των καταναλωτών σχετικά με τις αγορές προιόντων και υπηρεσιών.

Η πρόβλεψη αυτή, καθίσταται καθοριστική αν αναλογιστούμε πως η ικανότητα των επιχειρήσεων να προοικονομήσουν τη συμπεριφορά των καταναλωτών και τις αλλαγές στις προτιμήσεις και τις αγοραστικές τους συνήθειες επηρεάζει άμεσα τόσο το κέρδος όσο και την βιωσιμότητα τους. Μια αποτελεσματική πρόβλεψη μπορεί να βοηθήσει τις επιχειρήσεις να προσαρμοστούν στις αλλαγές της αγοράς, να προσφέρουν προϊόντα και υπηρεσίες που θα ανταποκρίνονται στις ανάγκες των καταναλωτών και να δημιουργήσουν ή να διατηρήσουν ένα ανταγωνιστικό πλεονέκτημα, ανιχνεύοντας κενά σημεία και διερευνώντας την περαιτέρω ανάπτυξη και εξέλιξη των προτιμόμενων απο το κοινό προϊόντων ή υπηρεσιών μέσω της καινοτομίας. Πέραν ωστόσο του προφανούς οικονομικού κινήτρου, μέσω της πρόβλεψης, επιτυγχάνεται η καλύτερη αξιοποίηση του χρόνου των καταναλωτών, οι οποίοι με τη σειρά τους δε χρειάζεται πλέον να αναλώνονται στην αναζήτηση τρόπων κάλυψης των αναγκών τους. Καλλιεργείται ένα κλίμα ασφάλειας το οποίο προσλαμβάνει ο καταναλωτής καθώς βλέπει τις ανάγκες του να γίνονται αντιληπτές και κατανοητές και εγκαθιδρύεται έτσι σχέση ασφάλειας και εμπιστοσύνης. Μέσω της πρόβλεψης χρήσης ή μη, μιας υπηρεσίας ή προιόντος, ένας οργανισμός είναι σε θέση να γνωρίζει και εμμέσως να καθορίζει την ζήτηση των προιόντων αυτών.

Η εκτίμηση της πιθανότητας αγοράς κάποιου προϊόντος ή χρήσης μιας υπηρεσίας από έναν καταναλωτή, μπορεί να οδηγήσει στην λήψη πληρέστερων αποφάσεων και την χάραξη πιο αποτελεσματικών στρατηγικών σε τομείς όπως το μαρκετινγκ, η εξυπηρέτηση πελατών (customer care) και η εφοδιαστική αλυσίδα (supply chain).

Για να επιτύχει αυτή τη πρόβλεψη, μια επιχείρηση καλείται να χρησιμοποιήσει εργαλεία παρατήρησης, ανάλυσης και πρόβλεψης της συμπεροφοράς πελατών. Μια εμπεριστατωμένη μέθοδος πρόβλεψης της πιθανής συμπεριφοράς του 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, και ελαχιστοποιούνται οι πιθανότητες έκφρασης τυχόν δυσαρέσκειας από τις παρεχόμενες υπηρεσίες. Μέσα από την ανάλυση της συμπεριφοράς των πελατών, μια επιχείρηση μπορεί να κατανοήσει τους λόγους για τους οποίους οι πελάτες εκφράζουν προτιμήσεις. 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 υφιστάμενους, αλλά και στοχεύοντας σε νέους πελάτες, εκ των προτέρων.

Η διαδικασία ανάλυσης και πρόβλεψης της συμπεριφοράς των καταναλωτών βασίζεται στη συλλογή επαρκών δεδομένων που αφορούν τους χρήστες των προσφερόμενων υπηρεσιών. Τα δεδομένα αυτά περιλαμβάνουν δημογραφικά δεδομένα, δεδομένα συναλλαγών, και δεδομένα καταγραφής δραστηριότητας και χρήσης των παρεχόμενων υπηρεσιών και προϊόντων. Μέχρι τώρα, χρησιμοποιούνταν μέθοδοι όπως η γραμμική παλινδρόμηση ή η απλή συσχέτιση και ομαδοποίηση των χρηστών 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 θα αναλύσουμε.

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 και οι αλγόριθμοι που χρησιμοποιούνται συνθέτουν 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 αλγόριθμοι γραμμικής παλινδρόμησης, μοντέλα βασισμένα 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, αλλά και παραδοσιακές τεχνικές συγκρίσης και υπολογισμού αποστάσεων για συνεργατικό φιλτράρισμα μεταξύ των δειγμάτων. Επιπλέον εξετάζεται η σημασία της συστηματικής μεθόδευσης αυτής της πρόβλεψης της συμπεριφοράς ενός πελάτη, καθώς και η ερμηνεία και αξιολόγηση των αποτελεσμάτων που προκύπτουν. Η προτεινόμενη μεθοδολογία λαμβάνει ως είσοδο ένα σύνολο δεδομένων που αφορά τη συμπεριφόρα καταναλωτών σε βάθος 18 μηνών, 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. Τέλος, η μεθοδολογία παράγει ως έξοδο τις προβλέψεις για τα εισαγόμενα δεδομένα, καθώς και γραφικές παραστάσεις και μετρικές για την αξιολόγησή τους.

Για την ανάπτυξη της μεθοδολογίας αυτής διεξήχθη μια εκτενής προεπεξεργασία των δεδομένων καθώς και πολλαπλά πειράματα που αφορούν τη βελτιστοποίηση των εισαγόμενων δεδομένων και τη διαδικασία εκπαίδευσης των μοντέλων πρόβλεψης, εξετάζοντας κάθε φορά διαφορετικές μεταβλητές και παραμέτρους του συστήματος. Για τα πειράματα χρησιμοποιήθηκαν δεδομένα του ομίλου επιχειρήσεων Santander Group ο οποίος αποτελεί μια Ισπανική πολυεθνική εταιρεία οικονομικών υπηρεσίων. Όλα τα

δεδομένα προήλθαν από τον τραπεζικό τομέα του ομίλου και την Santander Bank, που περιλαμβάνει τα δεδομένα των χρηστών για τις ποικίλες υπηρεσίες που προσφέρει η ομώνυμη τράπεζα.

1.2 Οργάνωση της εργασίας

Στο δεύτερο κεφάλαιο της παρούσας διπλωματικής εργασίας γίνεται μια εισαγωγή στην συμπεριφορά των καταναλωτών για την ανίχνευση και τον προσδιορισμό της πιθανότητας αγοράς νέων προϊόντων και υπηρεσιών. Αρχικά, προσδιορίζεται ο ορισμός της συμπεριφοράς των καταναλωτών γενικότερα και αναλύεται ο λόγος για τον οποίο απασχολεί τις σύγχρονες επιχειρήσεις. Στη συνέχεια, επιχειρείται μια προσέγγιση στην απόδοση του ορισμού της πρόβλεψης της συμπεριφοράς τους, σχετικά δηλαδή με το αν, εν δυνάμει, θα προβούν σε αγορά ή μη νέων προιόντων, και περιγράφεται ο τρόπος με τον οποίο η πρόβλεψη εντάσσεται στο ευρύτερο πλαίσιο της ανάλυσης της συμπεριφοράς τους, καθώς και η αξία που προσδίδει η εκτενής μελέτη και έρευνα του αντικειμένου της τόσο γενικά όσο και ειδικά, σε χρηματοοικονομικά σύνολα δεδομένων. Ύστερα, παρουσιάζονται διαφορετικές μέθοδοι για την πρόβλεψη της συμπεριφοράς πελατών, καθώς και μερικές μετρικές για την αξιολόγηση των προβλέψεων αυτών.

Το τρίτο κεφάλαιο αποτελεί εισαγωγή στη μηχανική μάθηση και τις τεχνικές προβλέψεων που βασίζονται στην ομοιότητα που παρουσιάζουν μεταξύ τους οι διάφοροι καταναλωτές . Γίνεται, αρχικά, μια ανασκόπηση των κατηγοριών της μηχανικής μάθησης και, έπειτα, παρατίθεται η λειτουργία των τεχνικών πρόβλεψης που χρησιμοποιούν μετρικές όπως οι απόσταση ή η ομοιότητα μεταξύ των εξεταζόμενων δειγμάτων. Στη συνέχεια, επεξηγείται η λειτουργία των μοντέλων πρόβλεψης που χρησιμοποιούνται στην παρούσα εργασία, δηλαδή του ταξινομητή Logistic Regression και των μοντέλων που βασίζονται στα δέντρα αποφάσεων και τέλος αναλύεται ο τρόπος λειτουργίας παραγωγής προβλέψεων μέσω συνεργατικού φιλτραρίσματος, των οποίων η δομή και η μέθοδος αναλύεται εκτενώς.

Στο τέταρτο κεφάλαιο παρουσιάζεται αναλυτικά η προτεινόμενη μεθοδολογία αυτοματοποίησης της διαδικασίας παραγωγής αποτελεσμάτων για την πρόβλεψη της συμπεριφοράς των πελατών μιας εταιρείας ή ενός οργανισμού, σχετικά με την πρόθεση τους να αγοράσουν ένα νέο για αυτούς προιόν ή υπηρεσία. Η μεθοδολογία περιλαμβάνει την προεπεξεργασία των δεδομένων εισόδου, την ανάπτυξη των μοντέλων πρόβλεψης, την επιλογή των υπερπαραμέτρων των μοντέλων πρόβλεψης που δίνουν τα βέλτιστα αποτελέσματα, και την εξαγωγή των προβλέψεων της πιθανότητας αγοράς νέων προιόντων 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.

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

Κεφάλαιο 2: Πρόβλεψη συμπεριφοράς πελατών (Customer behavior forecast)

2.1 Εισαγωγή

Εν μέσω συνεχούς και εκθετικής αύξησης της πληροφορίας και των δεδομένων στους περισσότερους τομείς της καθημερινότητας μας, η ανάγκη για στόχευση του ενδιαφέροντος των επιχειρήσεων στην πρόβλεψη της συμπεριφοράς των πελατών τους αποτελεί φλέγον ζήτημα. Οι επιχειρήσεις αντιλαμβάνονται ότι η απλή συλλογή δεδομένων δεν αρκεί πλέον για να εντοπίσουν πλήρως τις προτιμήσεις και τις ανάγκες των πελατών τους. Αντίθετα, πρέπει να επιδείξουν την ικανότητα να αναλύουν αυτά τα δεδομένα και να αντλήσουν πολύτιμες πληροφορίες και συστάσεις από αυτά. Με την ανάδυση της τεχνητής νοημοσύνης και των εργαλείων ανάλυσης, οι περισσότεροι οργανισμοί και επιχειρήσεις μπορούν πλέον να εξάγουν πρότυπα και τάσεις από τα δεδομένα, προβλέποντας με μεγαλύτερη ακρίβεια τις προτιμήσεις και τις συνήθειες των πελατών τους. Αυτή η δυνατότητα πρόβλεψης, τους επιτρέπει να προσαρμόζονται και να προσφέρουν εξατομικευμένες λύσεις, αυξάνοντας έτσι την πιθανότητα επιτυχίας του σκοπού των επιχειρήσεων καθώς και την ικανοποίηση των πελατών τους.

2.2 Ορισμός της συμπεριφοράς των καταναλωτών

Με τον όρο «συμπεριφορά των καταναλωτών» αναφέρόμαστε στις δράσεις, τις αντιδράσεις και τις επιλογές στις οποίες προβαίνουν οι καταναλωτές κατά τη διάρκεια της αγοράς προϊόντων ή υπηρεσιών. Επιχειρώντας να δώσουμε έναν πλήρη ορισμό της «συμπεριφοράς των καταναλωτών», θα λέγαμε πως περιλαμβάνει τον τρόπο με τον οποίο οι διάφοροι πελάτες μιας επιχείρησης ή οργανισμού, αντιλαμβάνονται, αξιολογούν και

αντιδρούν σε μια πρόταση ή προσφορά, καθώς και την επιρροή που έχουν οι προσωπικές τους ανάγκες, προτιμήσεις, αξίες και συνήθειες στην αγοραστική τους συμπεριφορά.

Οι συμπεριφορές των καταναλωτών μπορούν να ταξινομηθούν σε διάφορες κατηγορίες, οι οποίες συνοψίζονται στις ακόλουθες:

Συμπεριφορά αγοράς: Αυτή η κατηγορία περιλαμβάνει τις ενέργειες των καταναλωτών κατά την αγορά προϊόντων ή υπηρεσιών. Περιλαμβάνει τη λήψη αποφάσεων αγοράς, την αναζήτηση πληροφοριών, τη σύγκριση τιμών, την επιλογή προμηθευτών και την ολοκλήρωση της αγοράς.

Συμπεριφορά κατανάλωσης: Αναφέρεται στον τρόπο με τον οποίο οι καταναλωτές αξιοποιούν και καταναλώνουν τα αγορασθέντα προϊόντα ή υπηρεσίες. Αυτή η συμπεριφορά περιλαμβάνει τη συχνότητα και την ποσότητα της κατανάλωσης, την αναζήτηση νέων εμπειριών και την αξιοποίηση των προϊόντων με βάση τις προσωπικές προτιμήσεις και συνήθειες.

Συμπεριφορά επικοινωνίας: Αυτή η κατηγορία, αναφέρεται στην αλληλεπίδραση των καταναλωτών με τις επιχειρήσεις και τους άλλους καταναλωτές. Περιλαμβάνει τις προτιμήσεις επικοινωνίας, τη συμμετοχή σε κοινότητες καταναλωτών, την αξιολόγηση και την ανταλλαγή απόψεων για προϊόντα ή υπηρεσίες.

Συμπεριφορά απόφασης: Πρόκειται για τις διαδικασίες και τους παράγοντες που επηρεάζουν τις αποφάσεις αγοράς των καταναλωτών. Περιλαμβάνει την επίδραση της διαφήμισης, των προσωπικών συστάσεων, των κοινωνικών παραγόντων και των προηγούμενων εμπειριών στις αγοραστικές αποφάσεις.

Συνολικά, η κατανόηση της συμπεριφοράς των καταναλωτών είναι κρίσιμη για τις επιχειρήσεις, καθώς τους επιτρέπει να αντιληφθούν και να προβλέψουν τις ανάγκες και τις προτιμήσεις των πελατών τους. Μέσω αναλύσεων και εφαρμογής στρατηγικών μάρκετινγκ, οι επιχειρήσεις μπορούν να προσαρμόζονται και να παρέχουν εξατομικευμένες προσφορές και εμπειρίες, βελτιώνοντας έτσι την ικανοποίηση των πελατών και ενισχύοντας την ανταγωνιστικότητά τους στην αγορά.

2.3 Ορισμός της πρόβλεψης της συμπεριφοράς των καταναλωτών

Επιδιώκοντας να αποδόσουμε τον ορισμό της πρόβλεψης της συμπεριφοράς των καταναλωτών, θα λέγαμε πως πρόκειται ουσιαστικά για ένα σύνολο μεθοδευμένων και στοχευμένων ενεργειών, κατά τις οποίες χρησιμοποιώντας ένα σύνολο διαθέσιμων δεδομένων και μια μεθοδολογία, επιδιώκουμε να εξάγουμε μια πρόβλεψη βάσει πιθανοτήτων σχετικά με το αν και ποιά προϊόντα ή υπηρεσίες πρόκειται να αγοράσουν ή και να πάψουν να χρησιμοποιούν οι πελάτες μιας επιχείρησης ή οργανισμού. Συνήθως, η αρμοδιότητα για την πρόβλεψη της συμπεριφοράς των καταναλωτών ανήκει σε ένα τμήμα που ασχολείται με το μάρκετινγκ, όπως το τμήμα ανάλυσης μάρκετινγκ (marketing analytics) ή το τμήμα διαχείρισης πελατών (customer management). Σήμερα με τη ραγδαία ανάπτυξη της τεχνολογίας και της τεχνητής νοημοσύνης, με βεβαιότητα θα μπορούσαμε να πούμε πως ιδιαίτερα καίρια είναι η συνεισφορά τόσο του data analytics τομέα όσο και των αρμόδιων μηχανικών συστημάτων προτάσεων και μηχανικής μάθησης στη δουλειά των οποίων βασίζεται όλη η αναπτυχθείσα μεθοδολογία για την εκάστοτε εταιρεία.

Η πρόβλεψη της συμπεριφοράς των πελατών μιας εταιρείας είναι πολύ σημαντική και έχει αποδειχθεί ότι συνδέεται με ουσιαστικά οφέλη και κέρδη για τις επιχειρήσεις. Μερικά ενδιαφέροντα στατιστικά και use cases που αποδεικνύουν τη σημασία της πρόβλεψης της συμπεριφοράς των πελατών είναι τα εξής:

Αύξηση εσόδων: Σύμφωνα με μια έρευνα της McKinsey, οι εταιρείες που χρησιμοποιούν δεδομένα και αναλύσεις για την πρόβλεψη της συμπεριφοράς των πελατών έχουν πιθανότητα 126% υψηλότερης ανταπόκρισης στις καμπάνιες μάρκετινγκ και 85% υψηλότερη πιθανότητα αύξησης των εσόδων.

Προσαρμοστικότητα στις ανάγκες των πελατών: Μια έρευνα της Evergage δείχνει ότι το 88% των εμπορικών επιχειρήσεων πιστεύουν ότι η προσαρμογή της εμπειρίας των πελατών βάσει της πρόβλεψης της συμπεριφοράς τους έχει αυξήσει την απόδοση των πωλήσεων, την ευχαρίστηση των πελατών και την επικύρωση των πωλήσεων.

Ένα ενδιαφέρον παράδειγμα είναι αυτό του Netflix, σύμφωνα με το οποίο, το Netflix χρησιμοποιεί δεδομένα και αναλύσεις για να προβλέψει τις προτιμήσεις των χρηστών και να προτείνει εξατομικευμένες ταινίες και σειρές. Σύμφωνα με μια μελέτη της Gartner, το Netflix εκτιμά ότι το 75% των προβολών των χρηστών προέρχεται από προτάσεις που βασίζονται στην πρόβλεψη της συμπεριφοράς τους. Αυτό έχει συμβάλει στην αύξηση του αριθμού των συνδρομητών και τη διατήρηση της πιστότητάς τους.

Μελετώντας αντίστοιχα πραγματικά use cases σε τομείς χρηματοοικονομικής διαχείρισης, έχουμε τη χρήση της πρόβλεψης της συμπεριφοράς των πελατών για την πρόβλεψη της αγοράς προϊόντων και υπηρεσιών των τραπεζών. Με βάση τα δεδομένα συμπεριφοράς των πελατών, οι τράπεζες μπορούν να αναγνωρίσουν τα μοτίβα και τις τάσεις των πελατών τους και να προβλέψουν τις ανάγκες τους σε χρηματοοικονομικά προϊόντα και υπηρεσίες.

Ένα παράδειγμα είναι η πρόβλεψη των πελατών που ενδέχεται να αναζητήσουν δάνεια για την αγορά ακινήτου. Με τη χρήση δεδομένων όπως η ιστορική συμπεριφορά των πελατών, οι τράπεζες μπορούν να αναγνωρίσουν τα χαρακτηριστικά των πελατών που συνήθως ζητούν δάνεια για αγορά ακινήτου, όπως η ηλικία, το εισόδημα, ο τόπος διαμονής και οι οικονομικές συνήθειες. Με βάση αυτή την ανάλυση, γίνεται στόχευση σε πελάτες που πληρούν αυτά τα χαρακτηριστικά και τους προσφέρουν ειδικές προσφορές για δάνεια ακινήτων, αυξάνοντας έτσι τις πιθανότητες επιτυχίας της πώλησης και ικανοποίησης των πελατών.

Η πρόβλεψη της συμπεριφοράς των πελατών στον τραπεζικό τομέα μπορεί επίσης να χρησιμοποιηθεί για την αναγνώριση ανεπιθύμητων συναλλαγών ή απάτης. Με την ανάλυση των δεδομένων συναλλαγών και την αναγνώριση μοτίβων που υποδεικνύουν ανορθόδοξη συμπεριφορά, οι τράπεζες δύνανται άμεσα να αναλάβουν δράση για τον περιορισμό του κινδύνου και την προστασία των πελατών τους συμβάλλοντας στην αύξηση της εμπιστοσύνης, στη μείωση της απώλειας τους και φυσικά στην αύξηση της κερδοφορίας της ίδιας της τράπεζας.

Συμπερασματικά μέσω πληθώρας στατιστικών στοιχείων και παραδειγμάτων, αποδεικνύεται ότι η πρόβλεψη της συμπεριφοράς των πελατών έχει σημαντικό όφελος για

τις επιχειρήσεις και μπορεί να οδηγήσει σε αυξημένη ανταπόκριση, αύξηση εσόδων και βελτίωση της εμπειρίας των πελατών.

2.4 Μέθοδοι πρόβλεψης συμπεριφοράς των καταναλωτών

Η πρόβλεψη της συμπεριφοράς των πελατών σχετικά με την αγορά προϊόντων μπορεί να απαιτεί την υιοθέτηση ενός συνολικού πλαισίου και ακολουθίας βημάτων. Οι κύριες διαδικασίες που μπορούν να ακολουθηθούν περιλαμβάνουν τα εξής:

1. Εντοπισμός και συλλογή δεδομένων: Το πρώτο βήμα είναι η συλλογή δεδομένων σχετικά με τους πελάτες, τις αγορές τους και άλλους σχετικούς παράγοντες. Αυτά τα δεδομένα μπορεί να περιλαμβάνουν πληροφορίες όπως ηλικία, φύλο, τοποθεσία, ιστορικό αγορών, προτιμήσεις και συμπεριφορά αγορών.
2. Προεπεξεργασία των δεδομένων: Στη συνέχεια, τα δεδομένα υπόκεινται σε καθαρισμό και προετοιμασία. Αυτό περιλαμβάνει την αφαίρεση ανωμαλιών, τη συμπλήρωση κενών τιμών, την ανάλυση των δεδομένων και τη δημιουργία χαρακτηριστικών που είναι χρήσιμα για την πρόβλεψη.
3. Ανάπτυξη μοντέλου πρόβλεψης: Έπειτα, αναπτύσσεται ένα μοντέλο πρόβλεψης με τη χρήση μηχανικής μάθησης ή άλλων αλγορίθμων πρόβλεψης. Το μοντέλο αξιοποιεί τα δεδομένα που συλλέχθηκαν και προετοιμάστηκαν για να προβλέψει την πιθανή συμπεριφορά των πελατών σε σχέση με την αγορά προϊόντων.
4. Αξιολόγηση και βελτίωση: Το τελευταίο βήμα περιλαμβάνει την αξιολόγηση του μοντέλου πρόβλεψης και τη βελτίωσή του. Αυτό περιλαμβάνει τη σύγκριση των προβλέψεων με τα πραγματικά αποτελέσματα, την αναγνώριση αδυναμιών και την προσαρμογή του μοντέλου για καλύτερες προβλέψεις στο μέλλον.

Αυτά είναι γενικά τα βήματα που ακολουθούνται για την πρόβλεψη της συμπεριφοράς των πελατών σχετικά με την αγορά προϊόντων. Κάθε οργανισμός ή επιχείρηση μπορεί να έχει τις δικές της προσαρμογές και εφαρμογές ανάλογα με τις ανάγκες και τους στόχους της.

Τα δεδομένα που συλλέγονται για την πραγμάτωση των προβλέψεων συνήθως περιλαμβάνουν:

- Δημογραφικά δεδομένα (π.χ. ηλικία, φύλο, χώρα/πόλη)

- Ιστορικά αρχεία καταγραφής της χρήσης των υπηρεσιών της επειχείρησης (π.χ. χρήση πιστωτικής κάρτας, λήψη δανείου κ.λ.π. αν πρόκειται για μια τράπεζα)

- Το ιδιαίτερα εμπιστευτικά δεδομένα των πελατών σε χρονικό ορίζοντα βάθους της μελετης (παρελθοντικό εισόδημα, κατάσταση υγείας κ.λ.π.)

Για την πρόβλεψη της συμπεριφοράς της αγοράς ή χρήσης μιας υπηρεσίας ή προιόντος, μερικές από τις πιο δημοφιλείς μεθόδους που χρησιμοποιούνται παρουσιάζονται παρακάτω.

Regression Analysis

Το μοντέλο της δυαδικής λογιστικής παλινδρόμησης (Logistic Regression) είναι ένα αλγόριθμος μηχανικής μάθησης που χρησιμοποιείται για την πρόβλεψη πιθανοτήτων και την κατηγοριοποίηση (clustering). Χρησιμοποιείται ευρέως σε προβλήματα όπου η μεταβλητή εξόδου είναι δυαδική, δηλαδή έχει δύο πιθανές κατηγορίες.

Το μοντέλο λειτουργεί με βάση την εκτίμηση των πιθανοτήτων των κατηγοριών και την ανάθεση μιας προβλεπόμενης κατηγορίας με βάση αυτές τις πιθανότητες. Η βασική ιδέα είναι η χρήση ενός συναρτησιακού μοντέλου που περιγράφει την πιθανότητα ενός γεγονότος να ανήκει σε μια από τις δύο κατηγορίες. Η συνάρτηση που χρησιμοποιείται είναι η λογιστική (logistic) ή σιγμοειδής (sigmoid) συνάρτηση, που περιορίζει τις πιθανότητες μεταξύ του 0 και 1.

Για να εκπαιδευθεί το μοντέλο, χρησιμοποιούνται δεδομένα, όπου οι κατηγορίες των δειγμάτων είναι γνωστές. Το μοντέλο προσαρμόζει τις παραμέτρους του ώστε να προσεγγίσει καλύτερα τις πραγματικές πιθανότητες της κάθε κατηγορίας και μετά την εκπαίδευση, μπορεί να χρησιμοποιηθεί για να προβλέψει την κατηγορία ενός νέου δείγματος, βασιζόμενο στις χαρακτηριστικές παραμέτρους του. Ο αλγόριθμος εφαρμόζει τη σιγμοειδή συνάρτηση στις παραμέτρους του δείγματος και το κατηγοριοποιεί με βάση ένα κατώφλι (threshold) που έχει προηγουμένως καθοριστεί.

Το μοντέλο της δυαδικής λογιστικής παλινδρόμησης είναι απλό και ευέλικτο, και μπορεί να χρησιμοποιηθεί για προβλέψεις σε πολλούς τομείς, συμπεριλαμβανομένου και του τραπεζικού τομέα. Μπορεί να εφαρμοστεί για πρόβλεψη αγοράς προϊόντων ή υπηρεσιών τραπεζών, πρόβλεψη της πιθανότητας πληρωμής δανείων, ανίχνευση απάτης και πολλά άλλα.

Χρησιμοποιείται σε περιπτώσεις όπου η εξαρτημένη μεταβλητή, η οποία στην περίπτωσή μας είναι η αγορά ενός προιόντος, 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 στη συγκεκριμένη περίπτωση της δικής μας εργασίας αντιστοιχεί στην απόκτηση του προιόντος ή όχι, βάσει συνεχών ή/και κατηγορικών μεταβλητών. Το τυπικό linear regression model είναι το εξής:

$$P(X) = \alpha + \beta X, \text{ όπου } X = (x1, x2, \ldots, xn)$$

Για να περιορίσουμε την πιθανότητα εντός του διαστήματος [0, 1] και να διαφοροποιείται μονοτονικά βάσει του X, απαιτείται όπως προείπαμε η χρήση συναρτήσεων πέρα των γραμμικών. Μια συνάρτηση που πληροί τις προϋποθέσεις αυτές είναι η προαναφερθείσα σιγμοειδής συνάρτηση:

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

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:

$$P(X) = \alpha + b1x1 + b2x2 + \cdots + bnxn$$

και ο αντίστοιχος τύπος στο logistic regression μοντέλο είναι:

$$Q(X) = 1 - P(X) = \frac{1}{1+e^{-(\alpha+b1x1+b2x2+\cdots+bnxn)}}$$

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

Collaborative filtering

Το collaborative filtering (συνεργατικό φιλτράρισμα ή διήθηση) είναι μια τεχνική που χρησιμοποιείται στον τομέα της πρόβλεψης συμπεριφοράς των καταναλωτών και των προτάσεων προϊόντων και υπηρεσιών. Στην ουσία, αποσκοπεί στην πρόβλεψη των προτιμήσεων ενός ατόμου βασιζόμενο στις προτιμήσεις ή τη συμπεριφορά παρόμοιων ατόμων.

Ο collaborative filtering αλγόριθμος βασίζεται στην ανάλυση των παρατηρήσεων και των συνήθειών των χρηστών-πελατών. Υπάρχουν δύο βασικές προσεγγίσεις:

Το user-based collaborative filtering και το item-based collaborative filtering. Αυτές οι δύο παραλλαγές διαφέρουν ουσιαστικά στον τρόπο με τον οποίο πραγματοποιείται η πρόβλεψη.

Το user-based collaborative filtering βασίζεται στην αναζήτηση πελατών με παρόμοιες προτιμήσεις με έναν συγκεκριμένο πελάτη. Με βάση αυτές τις παρόμοιες προτιμήσεις, μπορούν να προβλεφθούν οι προτιμήσεις του συγκεκριμένου καταναλωτή για αντικείμενα που παρόμοιοι καταναλωτές δείχνουν να διαθέτουν ή να έχουν χρησιμοποιήσει στο παρελθόν.

Από την άλλη πλευρά, το item-based collaborative filtering επικεντρώνεται στην ανάλυση παρόμοιων αντικειμένων. Βασίζεται στην υπόθεση ότι αν ένας πελάτης αξιολογήσει θετικά ένα συγκεκριμένο προϊόν ή υπηρεσία, τότε θα έχει προτίμηση και για παρόμοια προιόντα.

Και οι δύο παραλλαγές έχουν τα πλεονεκτήματά τους και χρησιμοποιούνται ανάλογα κατά περίπτωση. Η επιλογή μεταξύ user-based και item-based collaborative filtering εξαρτάται από τον τρόπο που θέλουμε να γίνει η πρόβλεψη των προτιμήσεων των πελατών, ενώ περαιτέρω παράγοντες που επηρεάζουν την επιλογή έκαστου είναι οι εξής:

Δεδομένα: Ο τύπος των δεδομένων που έχουμε στη διάθεσή μας μπορεί να επηρεάσει την επιλογή. Συγκεκριμένα, αν έχουμε πληροφορίες για τις αγορές των χρηστών για προϊόντα, τότε το user-based collaborative filtering μπορεί να είναι πιο κατάλληλο. Αντίθετα, αν έχουμε πληροφορίες για την αξιολόγηση των αντικειμένων από τους χρήστες, τότε το item-based collaborative filtering μπορεί να είναι προτιμητέο.

Διαστάσεις: Ο αριθμός των διαστάσεων των δεδομένων μπορεί επίσης να επηρεάσει την επιλογή. Συγκεκριμένα, αν ο αριθμός των χρηστών είναι μεγάλος σε σχέση με τον αριθμό των αντικειμένων, τότε το user-based collaborative filtering μπορεί να είναι πιο αποδοτικό. Αντίθετα, αν ο αριθμός των αντικειμένων είναι μεγαλύτερος, τότε το item-based collaborative filtering μπορεί να είναι πιο κατάλληλο.

Αλγόριθμος: Ο τρόπος που θέλουμε να γίνει η πρόβλεψη μπορεί να επηρεάσει την επιλογή. Ο user-based collaborative filtering αλγόριθμος βασίζεται στην ομοιότητα μεταξύ

των χρηστών, ενώ ο item-based collaborative filtering αντίστοιχα, βασίζεται στην ομοιότητα μεταξύ των αντικειμένων. Καθένας από αυτούς τους αλγορίθμους μπορεί να παρέχει διαφορετικά αποτελέσματα και να είναι πιο κατάλληλος για συγκεκριμένες προτιμήσεις και απαιτήσεις.

Συνολικά, η επιλογή μεταξύ user-based και item-based collaborative filtering εξαρτάται από τη φύση των δεδομένων, τις διαστάσεις των δεδομένων και τον αλγόριθμο που επιθυμούμε να χρησιμοποιήσουμε για την πρόβλεψη των προτιμήσεων των καταναλωτών.

Ένα επιπλέον σημείο που πρέπει να αναφερθεί εδώ, είναι ο υπολογισμός των αποστάσεων μεταξύ των πελατών ή των αντικειμένων. Στους collaborative filtering αλγόριθμους, τείνουν να χρησιμοποιούνται διάφορες μετρικές ομοιότητας ανάλογα το πεδίο εφαρμογής. Οι πιο συνηθισμένες μετρικές περιλαμβάνουν:

Ευκλείδεια Απόσταση (Euclidean Distance): Υπολογίζει την ευκλείδεια απόσταση μεταξύ δύο σημείων σε ένα πολυδιάστατο χώρο. Αυτή η μετρική χρησιμοποιείται για τον υπολογισμό της ομοιότητας μεταξύ χρηστών ή αντικειμένων βάσει των χαρακτηριστικών τους.

Συνημιτονοειδής Ομοιότητα (Cosine Similarity): Υπολογίζει την ομοιότητα μεταξύ δύο διανυσμάτων βάσει της μεταξύ τους γωνίας. Αυτή η μετρική χρησιμοποιείται κυρίως για την αναπαράσταση των αξιολογήσεων των χρηστών για αντικείμενα ή των χαρακτηριστικών των αντικειμένων.

Jaccard Similarity: Η μετρική αυτή χρησιμοποιείται συνήθως σε περιπτώσεις όπου έχουμε να κάνουμε με σύνολα δεδομένων, και θέλουμε να μετρήσουμε την ομοιότητα μεταξύ των συνόλων αυτών. Συγκεκριμένα, η Jaccard Similarity είναι χρήσιμη για την αξιολόγηση της ομοιότητας μεταξύ δύο συνόλων με βάση τα κοινά και τα μη κοινά στοιχεία τους.

Machine learning

Η πιο διαδεδομένη μέθοδος πρόβλεψης της συμπεριφοράς των καταναλωτών σε σχέση με την αγορά ή μη ενός προιόντος ή υπηρεσίας, είναι η χρήση κάποιου μοντέλου μηχανικής μάθησης. Από τις πιο συνηθισμένες επιλογές είναι τα δέντρα απόφασης, στα οποία, οι κόμβοι του δέντρου δημιουργούνται βάσει των τιμών και της σημασίας 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.

Η Μηχανική Μάθηση, καθώς και τα διάφορα μοντέλα που χρησιμοποιούνται, αναλύονται εκτενώς στο επόμενο κεφάλαιο της εργασίας.

2.5 Μετρικές αξιολόγησης των προβλέψεων

Προκειμένου να αξιολογήσουμε αλλά και να κατανοήσουμε σε βάθος τα αποτελέσματα των προβλέψεων που παράγουν οι διάφορες μέθοδοι και μοντέλα πρόβλεψης, απαιτείται να ορίσουμε κάποιες κατάλληλες μετρικές. Η πλειονότητα των μετρικών που θα ορίσουμε βασίζεται στις έννοιες των true/false positive και true/false negative των οποίων η σημασία ακολουθεί:

- True Positive (TP): το μοντέλο προβλέπει ορθώς ότι ο πελάτης αγοράζει το προϊόν
- True Negative (TN): το μοντέλο προβλέπει ορθώς ότι ο πελάτης δεν αγοράζει το προϊόν
- False Positive (FP): το μοντέλο προβλέπει λανθασμένα ότι ο πελάτης αγοράζει το προϊόν
- False Negative (FN): το μοντέλο προβλέπει λανθασμένα ότι ο πελάτης δεν αγοράζει το προϊόν

Οι μετρικές που χρησιμοποιούνται για την αξιολόγηση των μοντέλων παρουσιάζονται παρακάτω:

Confusion Matrix

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

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

Πίνακας 1: Confusion Matrix 2x2

Accuracy

Το accuracy (ακρίβεια) είναι μία μετρική που χρησιμοποιείται για να αξιολογήσει την απόδοση ενός μοντέλου δυαδικής πρόβλεψης. Αναφέρεται στο ποσοστό των σωστών προβλέψεων που πραγματοποίησε το μοντέλο συνολικά, σε σχέση με τον συνολικό αριθμό των προβλέψεων και ορίζεται απο τη σχέση:

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

Πρόκειται για το ποσοστό των σωστών προβλέψεων (TP και TN) συνολικά, δηλαδή το άθροισμα των πελατών που ορθώς προβλέψαμε οτι αγόρασαν το προϊόν και ομοίως αυτών που ορθώς προβλέψαμε οτι δεν αγόρασαν το προϊόν, δια τον συνολικό αριθμό των προβλέψεων (TP, TN, FP και FN). Η ακρίβεια είναι ένας από τους πιο βασικούς και προφανείς τρόπους να μετρήσουμε την απόδοση ενός μοντέλου, αλλά μπορεί να παρουσιάζει προβλήματα σε περιπτώσεις μη ισορροπημένων κατηγοριών.

Precision

Το precision (ακρίβεια) είναι μία μετρική που αξιολογεί την ακρίβεια των θετικών προβλέψεων ενός μοντέλου δυαδικής πρόβλεψης (κατηγοριοποίησης). Αναφέρεται στο ποσοστό των σωστών θετικών προβλέψεων σε σχέση με τον συνολικό αριθμό των θετικών προβλέψεων που έγιναν.
Ο ορισμός του precision είναι ο εξής:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Πρόκειται για το ποσοστό των σωστών θετικών προβλέψεων (TP) δηλαδή των πελατών που ορθώς προβλέφθηκε ότι αγόρασαν το προϊόν, σε σχέση με τον συνολικό

αριθμό των θετικών προβλέψεων που έγιναν (TP και FP). Το precision εστιάζει στην ακρίβεια των προβλέψεων για μια συγκεκριμένη κατηγορία και δείχνει πόσο συχνά οι θετικές προβλέψεις είναι πραγματικά σωστές. Αυξημένη τιμή του precision υποδηλώνει μικρό ποσοστό false positives, δηλαδή μικρή πιθανότητα να προβλέψουμε λανθασμένα μια αρνητική περίπτωση ως θετική.

Recall

Το recall (sensitivity) είναι μία μετρική που αξιολογεί την ικανότητα ενός μοντέλου κατηγοριοποίησης να εντοπίζει όλες τις θετικές περιπτώσεις. Αναφέρεται στο ποσοστό των σωστών θετικών προβλέψεων σε σχέση με τον συνολικό αριθμό των πραγματικών θετικών περιπτώσεων δηλαδή το άθροισμα όσων πελατών ορθώς προβλέφθηκε ότι θα αγοράσουν ένα προϊον και εκείνων που λανθασμένα προφλέφθηκε ότι δε θα αγοράσουν το προϊόν.

Ο ορισμός του recall δίνεται ως εξής:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Πρόκειται για το ποσοστό των σωστών θετικών προβλέψεων (TP) σε σχέση με τον συνολικό αριθμό των πραγματικών θετικών περιπτώσεων (TP και FN). Το recall εστιάζει στην ικανότητα του μοντέλου να εντοπίζει όσο το δυνατόν περισσότερες από τις πραγματικές θετικές περιπτώσεις. Μεγαλύτερη τιμή του recall υποδηλώνει μικρό ποσοστό false negatives, δηλαδή την πιθανότητα να αποτύχει στο να προβλέψει μια θετική περίπτωση.

F1-score

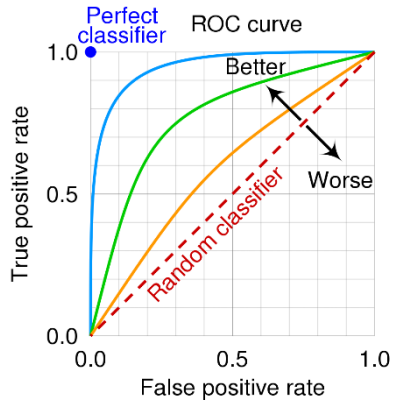
Το F1-score είναι ο αρμονικός μέσος μεταξύ precision και recall. Έτσι, ορίζεται ως:

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

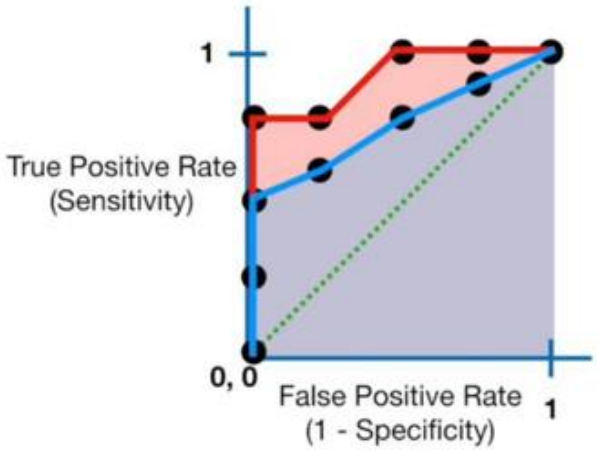
Ο συνδυασμός των δύο παραπάνω μετρικών δίνει μια καλή γενική εικόνα των αποτελεσμάτων του μοντέλου. Γενικεύεται σε Fβ-score για την πρόσθεση επιπλέον παράγοντα βάρους στο precision ή στο recall:

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

ROC (Receiver Operating Characteristic) curve & AUC (Area Under the receiver operating characteristic Curve)

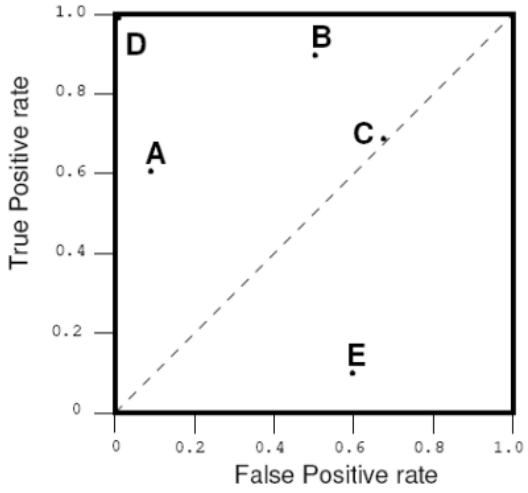


Σχήμα 1: Διαφορετικά μοντέλα με τις αντίστοιχες ROC curves που παράγουν



Σχήμα 2: Δύο διαφορετικά μοντέλα με τις αντίστοιχες ROC curves & AUC που παράγουν

Η ROC καμπύλη απεικονίζει τον ρυθμό των true positives (TPR) έναντι του ρυθμού των false positives (FPR) για διάφορα κατώφλια (thresholds) πρόβλεψης. Το AUC είναι το εμβαδόν μεταξύ ενός ROC γραφήματος και του άξονα των x. Παίρνει τιμές από 0 έως 1 και ένα μοντέλο με AUC ίσο με 1 έχει τέλεια δυνατότητα διάκρισης μεταξύ των θετικών και αρνητικών περιπτώσεων δηλάδη της αγοράς ή όχι ενός προϊόντος στην περίπτωση μας, ενώ ένα μοντέλο με AUC ίσο με 0.5 δεν έχει δυνατότητα ακριβούς διάκρισης και λειτουργεί τυχαία. Συνεπώς μεγαλύτερες τιμές υποδηλώνουν καλύτερη απόδοση και μεγαλύτερη ικανότητα διάκρισης μεταξύ των κατηγοριών. Στην ακόλουθη εικόνα βλέπουμε το γράφημα μια καμπύλης ROC όπου τα πέντε σημεία A, B, C, D και E, αποτελούν πέντε διαφορετικούς ταξινομητές (μοντέλα πρόβλεψης) όπου το καθένα έχει διαφορετικό threshold για την κατηγοριοποίηση ενός πελάτη σε αυτούς που προβαίνουν σε αγορά ενός προϊόντος ή όχι.



Σχήμα 3: Παράδειγμα 5 διακριτών ταξινομητών

Κάποια σημεία πάνω στον ROC χώρο παρουσιάζουν ιδιαίτερο ενδιαφέρον. Αρχικά το χαμηλότερο σημείο από τα αριστερά, το (0,0) αντιπροσωπεύει την στρατηγική του να μην εκδίδεται ποτέ μία θετική ταξινόμηση (αγορά προϊόντος). Αυτό σημαίνει ότι στην περίπτωση αυτή ένας ταξινομητής δεν παράγει ούτε ψευδώς θετικά αποτελέσματα αλλά ούτε και αληθώς θετικά. Η αντίθετη περίπτωση, κατά την οποία, ανεξαρτήτως συνθηκών παράγονται μόνο θετικές ταξινομήσεις, αντιπροσωπεύεται από το ανώτερο από τα δεξιά σημείο, (1,1) για το οποίο ο ταξινομητής έχει threshold 0. Το σημείο (0,1) αντιπροσωπεύει την τέλεια ταξινόμηση. Το σημείο D στην παραπάνω γραφική παράσταση (σχήμα 1) αντιστοιχεί στην ιδανική αυτή περίπτωση. Ανεπίσημα, ένα σημείο στον ROC χώρο είναι καλύτερο από ένα άλλο εφόσον βρίσκεται περισσότερο βορειοδυτικά από εκείνο, πράγμα που σημαίνει ότι (FPR χαμηλότερο, TPRυψηλότερο, ή και τα δύο). Οι ταξινομητές που εμφανίζονται στην αριστερή πλευρά του ROC γραφήματος, κοντά στον άξονα των x, μπορούν να θεωρηθούν πιο συντηρητικοί καθώς αυτοί πραγματοποιούν θετικές ταξινομήσεις μόνο όταν διαθέτουν ισχυρές ενδείξεις. Αυτό έχει σαν αποτέλεσμα να δίνουν λίγα ψευδώς θετικά αποτελέσματα. Η τακτική αυτή όμως έχει ως συνέπεια ένα χαμηλό TPR, δηλαδή περιορισμένο αριθμό αληθώς θετικών αποτελεσμάτων. Από την άλλη πλευρά, οι ταξινομητές που βρίσκονται στην άνω δεξιά πλευρά του ROC γραφήματος μπορούν να θεωρηθούν πιο φιλελεύθεροι. Αυτοί πραγματοποιούν θετικές ταξινομήσεις με ασθενέστερα

στοιχεία με αποτέλεσμα να ταξινομούν σχεδόν όλα τα θετικά περιστατικά σωστά, αλλά συγχρόνως έχουν και υψηλό ποσοστό ψευδώς θετικών ταξινομήσεων. Στο σχήμα 3, το σημείο Α είναι περισσότερο συντηρητικό από το Β. Επειδή σε πολλούς τομείς, στον πραγματικό κόσμο επικρατούν μεγάλοι αριθμοί αρνητικών περιστατικών, η περιοχή της αριστερής πλευράς του ROC γραφήματος παρουσιάζει ιδιαίτερο ενδιαφέρον.

Κεφάλαιο 3: Μέθοδοι προβλέψεων και Μηχανική μάθηση

3.1 Εισαγωγή

Οι μέθοδοι πρόβλεψης αποτελούν ένα σύνολο αλγορίθμων και τεχνικών που χρησιμοποιούνται για να προβλέψουν τη μελλοντική κατάσταση ενός συστήματος ή μιας διαδικασίας. Τα πεδία στα οποία βρίσκουν εφαρμογή ολοένα και πληθαίνουν, με μερικά απο αυτά να είναι η οικονομία, η βιομηχανία, η υγεία και η επιστήμη των υπολογιστών. Σκοπός των διαφόρων μεθόδων πρόβλεψης, είναι να συμβάλλουν στην ορθότερη λήψη αποφάσεων και να βελτιώσουν την απόδοση του συστήματος για το οποίο εφαρμόζονται. Οι κύριες κατηγορίες στις οποίες διαχωρίζονται και κατατάσσονται, είναι συνήθως οι ακόλουθες δύο:

1. Μέθοδοι μηχανικής μάθησης
2. Κλασσικές στατιστικές μέθοδοι

Οι κλασσικές στατιστικές μέθοδοι πρόβλεψης βασίζονται σε αρχές στατιστικής, στη θεωρία της πιθανότητας και στην ανάλυση ιστορικών δεδομένων, καθώς και σε έννοιες όπως η παλαιότητα και η τάση ενώ οι μη στατιστικές μέθοδοι μηχανικής μάθησης βασιζονται σε αρχές τεχνητής νοημοσύνης εφαρμόζοντας αλγορίθμους που επιτρέπουν στο σύστημα να "μάθει" από τα δεδομένα και να προβλέψει μελλοντικά παραδείγματα.

Στο κεφάλαιο αυτό, εξετάζονται διάφορες κλασσικές μέθοδοι προβλέψεων και μηχανικής μάθησης, καθώς και οι αρχές που τις διέπουν. Θα εξετάσουμε τα πλεονεκτήματα και τις προκλήσεις κάθε μεθόδου και θα δούμε πώς μπορούν να εφαρμοστούν σε διάφορους τομείς, προσδίδοντας σημαντική αξία στη λήψη αποφάσεων και στην πρόβλεψη μελλοντικών καταστάσεων. Για το σκοπό αυτό ξεκινάμε το κεφάλαιο αυτό με μια συνοπτική ανάλυση της έννοιας της μηχανικής μάθησης και των διάφορων κατηγοριών στις οποίες διαχωρίζεται.

3.2 Κατηγορίες μηχανικής μάθησης

Η μηχανική μάθηση αποτελεί έναν υποτομέα της τεχνητής νοημοσύνης που ασχολείται με την ανάπτυξη αλγορίθμων και μοντέλων που επιτρέπουν στα συστήματα να "μάθουν" από τα δεδομένα και να προβλέπουν ή να προσαρμόζονται αυτόματα χωρίς να χρειάζεται να προγραμματίζονται για κάθε πιθανή κατάσταση που ενδέχεται να προκύψει. Η δημιουργία της Μηχανικής Μάθησης προήλθε από την ανάγκη για αυτόνομα συστήματα που μπορούν να αντιληφθούν και να εκτελέσουν καθήκοντα χωρίς να προγραμματιστούν απόλυτα για κάθε πιθανή περίπτωση. Με την αύξηση του όγκου των διαθέσιμων δεδομένων και την ανάπτυξη ισχυρών υπολογιστικών συστημάτων, κατέστη δυνατό να αξιοποιηθούν οι αλγόριθμοι μηχανικής μάθησης για να αντληθούν αυτόματα συμπεράσματα και προβλέψεις από τα διάφορα δεδομένα και πληροφορίες εξασφαλίζοντας έτσι νέους σύγχρονους τρόπους σντιμετώπισης πληθώρας τύπων προβλημάτων σε διάφορους τομείς.

Οι κύριοι τύποι προβλημάτων που λύνονται με τη χρήση αλγορίθμων μηχανικής μάθησης περιλαμβάνουν:

- Προβλήματα Ταξινόμησης (Classification): Σε αυτά τα προβλήματα, το μοντέλο μαθαίνει να αντιστοιχίζει εισόδους σε συγκεκριμένες κατηγορίες ή ετικέτες. Για παράδειγμα, αναγνώριση εικόνας ως σκύλος ή γάτα, αναγνώριση ηλεκτρονικών μηνυμάτων ως spam ή όχι.
- Προβλήματα Παλινδρόμησης (Regression): Σε αυτά τα προβλήματα, το μοντέλο μαθαίνει να προβλέπει μια συνεχή αριθμητική τιμή ή ένα σύνολο αριθμητικών τιμών. Για παράδειγμα, πρόβλεψη της τιμής ενός ακινήτου βάσει χαρακτηριστικών του, πρόβλεψη της ποσότητας πωλήσεων ενός προϊόντος βάσει διαφημίσεων και παραμέτρων πελατείας.
- Συσταδοποίηση (Clustering): Σε αυτό τον τύπο προβλημάτων, το μοντέλο επιδιώκει να ομαδοποιήσει τα δεδομένα σε ομάδες (συστάδες) βάσει των χαρακτηριστικών τους. Συχνές εφαρμογές περιλαμβάνουν την ομαδοποίηση πελατών για στρατηγικές μάρκετινγκ και την ομαδοποίηση εικόνων ή κειμένου για ανάκτηση πληροφοριών.
- Συστάσεις-προτάσεις (Recommendation): Σε αυτόν τον τύπο προβλημάτων, το μοντέλο προτείνει αντικείμενα ή προϊόντα που θα ενδιέφεραν έναν χρήστη βάσει

των προηγούμενων προτιμήσεών του. Παραδείγματα περιλαμβάνουν προτάσεις για ταινίες, μουσική, προϊόντα αγορών και κοινωνικά δίκτυα.

- 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.

Αυτοί είναι μόνο μερικοί από τους τύπους προβλημάτων που μπορούν να λυθούν με τη χρήση αλγορίθμων μηχανικής μάθησης. Η ευελιξία και η δυνατότητα προσαρμογής των αλγορίθμων μηχανικής μάθησης τους καθιστούν ιδιαίτερα χρήσιμους σε πολλούς τομείς και εφαρμογές.

Αναλυτικότερα οι διάφορες καηγορίες μάθησης παρουσιάζονται παρακάτω.

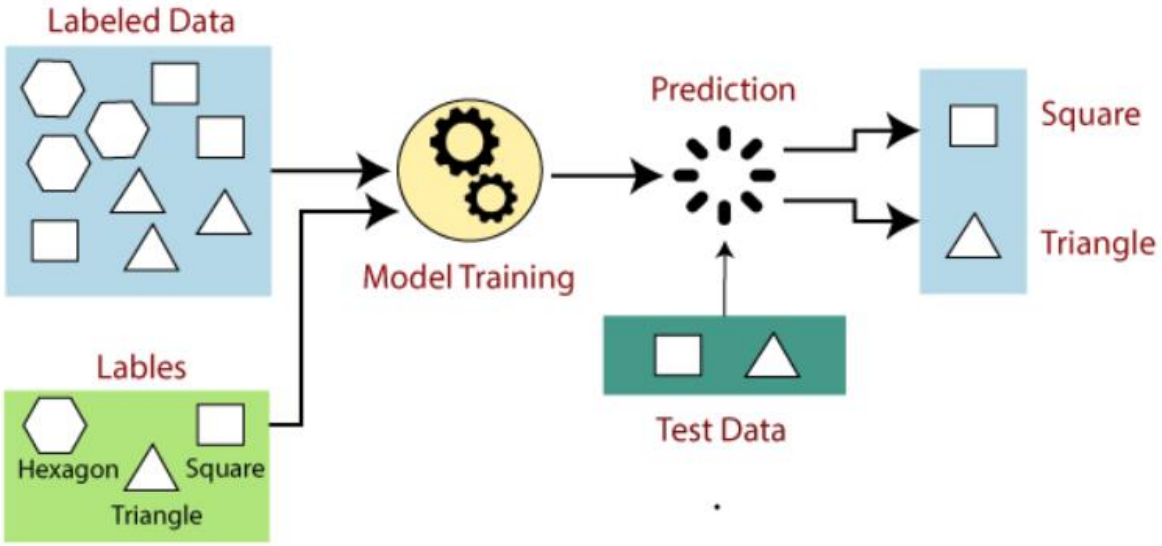
3.2.1 Επιβλεπόμενη μάθηση (Supervised learning)

Η επιβλεπόμενη μάθηση ή αλλιώς μάθηση με εκπαιδευτή, είναι μια κατηγορία μηχανικής μάθησης που χρησιμοποιείται για να εκπαιδεύσει μοντέλα από ένα σύνολο δεδομένων τα οποία περιλαμβάνουν ετικέτες (labeled data). Ένα μοντέλο που ανήκει σε αυτήν την κατηγορία, λαμβάνει ως είσοδο δεδομένα κάθε ένα εκ των οποίων συνοδεύεται από την ετικέτα του και αντικειμενικός σκοπός του μοντέλου είναι να παράξει την ετικέτα νέων δεδομένων που θα λάβει ώς είσοδο αφού πρώτα εκπαιδευτεί. Ανάλογα με τη μορφή των ετικετών που επιθυμούμε να παράξουμε, η οποία είναι ρητά συνδεδεμένη και με τη φύση του προς λύση προβλήματος διακρίνουμε διάφορες κατηγορίες προβλημάτων.

Ουσιαστικά μπορούμε να πούμε πως 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. Ένα τέτοιο παράδειγμα είναι και η πρόβλεψη αγοράς ή όχι ενός προϊόντος ή υπηρεσίας, και για αυτό το είδος μάθησης που χρησιμοποιήθηκε για τη διεκπεραίωση της παρούσας εργασίας είναι το συγκεκριμένο όπως θα εξηγηθεί και στα επόμενα κεφάλαια. Από την άλλη μεριά, αν η έξοδος λαμβάνει συνεχείς τιμές, η αντιστοίχιση οδηγεί σε παλινδρόμηση (regression). Τα προβλήματα regression αφορούν τη δυνατότητα πρόβλεψης αριθμών, όπως η τιμή μιας μετοχής ή η θέση ενός αυτοκινήτου.

Η έννοια της επιβλεπόμενης μάθησης πρωτοεμφανίστηκε στα τέλη του 1950 και τα αρχές του 1960 και αποτέλεσε σημαντική πρόοδο στο πεδίο της τεχνητής νοημοσύνης. Σήμερα, η επιβλεπόμενη μάθηση χρησιμοποιείται συχνά στους τομείς των υπολογιστών, των βάσεων δεδομένων, των συστημάτων πληροφορικής, της πρόβλεψης ζήτησης κ.α..

Σχηματικά η έννοια της επιβλεπόμενης μάθησης μπορέι να γλινει αντιληπτή, αν λάβουμε υπόψη μας την ακόλουθη απεικόνιση.



Σχήμα 4: Επιβλεπόμενη μάθηση [1]

Στο παραπάνω διάγραμμα, το μοντέλο τροφοδοτείται απο ένα σύνολο δεδομένων για τα οποία γνωρίζουμε και γνωστοποιούμε στο μοντέλο τις αντίστοιχες ετικέτες τους. Με αυτόν τον τρόπο εκπαδεύουμε ουσιαστικά το μοντέλο για κάθε σχήμα.

- Αν το δοσμένο σχήμα έχει 4 πλευρές και όλες οι πλευρές είναι ίσες, τότε θα ονομάζεται (=θα έχει την ετικέτα) **τετράγωνο**.
- Αν το δοσμένο σχήμα έχει 3 πλευρές, τότε θα έχει την ετικέτα **τρίγωνο**.
- Αν το δοσμένο σχήμα έχει 6 πλευρές οι οποίες είναι και μεταξύ τους ίσες, τότε θα ονομάζεται **εξάγωνο**.
- κ.ο.κ.

Έχοντας ολοκληρώσει την εκπαίδευση του μοντέλου, προχωράμε σε έλεγχο του μοντέλου μέσω των test data ελέγχοντας αν ο αντικειμενικός σκοπός του μοντέλου που εκπαιδεύσαμε (κατηγοριοποίηση σχημάτων) πραγματοποιείται επιτυχώς. Τροφοδοτούμε δηλαδή το μοντέλο με σχήματα των οποίων την ετικέτα γνωρίζουμε ήδη και έτσι λαμβάνουμε μια εικόνα σχετικά με την απόδοση του. Μέσω αυτής της διαδικασίας το

μοντέλο καθίσταται ικανό να προβλέψει όλα τα πιθανά σχήματα που θα λάβει ως είσοδο, λαμβάνοντας υπόψην του το πλήθος των πλευρών του εισηγμένου σχήματος.

3.2.2 Μη επιβλεπόμενη μάθηση (Unsupervised learning)

Μία δεύτερη κατηγορία μηχανικής μάθησης είναι αυτή των αλγορίθμων μη επιβλεπόμενης μάθησης (unsupervised learning). Αντίθετα με την προηγούμενη κατηγορία μοντέλων επιβλεπόμενης μάθησης στην οποία η εκπαίδευση πραγματοποιείται χρησιμοποιώντας ετικετοποιημένα δεδομένα, σε αυτήν την κατηγορία, το υπό εκπαίδευση μοντέλο, «μαθαίνει», χρησιμοποιώντας δεδομένα τα οποία δεν συνοδεύονται από κάποια γνωστή ετικέτα (πληροφορία). Κύριος λόγος ανάπτυξης της συγκεκριμένης κατηγορίας, είναι πως σε πραγματικές περιπτώσεις μπορεί να μην έχουμε στη διάθεση μας τέτοιου είδους πληροφορίες που να συνοδεύουν τα δεδομένα μας.

Βασικός στόχος αυτής της κατηγορίας αλγορίθμων μηχανικής μάθησης, είναι να ανακαλυφθούν δομές, μοτίβα ή πρότυπα στα δεδομένα δίχως την άμεση διάθεση προηγούμενης πληροφορίας.

Όπως υποδηλώνει και το όνομά της κατηγορίας αυτής, η μη επιβλεπόμενη μάθηση είναι μια τεχνική μηχανικής μάθησης στην οποία τα μοντέλα δεν εκπαιδεύονται με χρήση train δεδομένων. Αντίθετα, ανακαλύπτουν από μόνα τους τα κρυμμένα μοτίβα και τις πληροφορίες από τα δεδομένα που δίνονται. Μια τέτοια διαδικασία, μπορεί να παρομοιαστεί με τη διαδικασίας συμπερασμού γνώσης που λαμβάνει χώρα στον ανθρώπινο εγκέφαλο κατά την εκμάθηση νέων πραγμάτων.

Η μη επιβλεπόμενη μάθηση δεν μπορεί να εφαρμοστεί απευθείας σε ένα πρόβλημα παλινδρόμησης ή ταξινόμησης, διότι, αντίθετα με την επιβλεπόμενη μάθηση, έχουμε τα δεδομένα εισόδου αλλά όχι τα αντίστοιχα δεδομένα εξόδου. Ο στόχος της μη επιβλεπόμενης μάθησης είναι να βρεί την υποκείμενη δομή του συνόλου δεδομένων, να ομαδοποιήσει αυτά τα δεδομένα βάσει ομοιοτήτων και να αναπαραστήσει αυτό το σύνολο δεδομένων με μια συνεταγμένη μορφή.

Ας εξετάσουμε το ακόλουθο παράδειγμα:

Ας υποθέσουμε ότι ένας αλγόριθμος μη επιβλεπόμενης μάθησης λαμβάνει ώς είσοδο, ένα σύνολο εικόνων που περιέχουν διάφορες ράτσες γατών και σκύλων. Αντικειμενικός στόχος του υπό εξέταση αλγορίθμου, είναι να αναγνωρίσει τα χαρακτηριστικά της εικόνας μόνος του και να κατατάξει στην κατάλληλη κατηγορία το υπό εξέταση δείγμα. Ο αλγόριθμος δεν έχει προηγουμένως εκπαιδευτεί πάνω στο συγκεκριμένο σύνολο δεδομένων, πράγμα που σημαίνει ότι δεν έχει καμία ιδέα για τα χαρακτηριστικά του. Ένας αλγόριθμος μη επιβλεπόμενης μάθησης θα εκτελέσει αυτήν την εργασία διαχωρίζοντας το σύνολο των εικόνων σε ομάδες με βάση τις ομοιότητες μεταξύ τους.

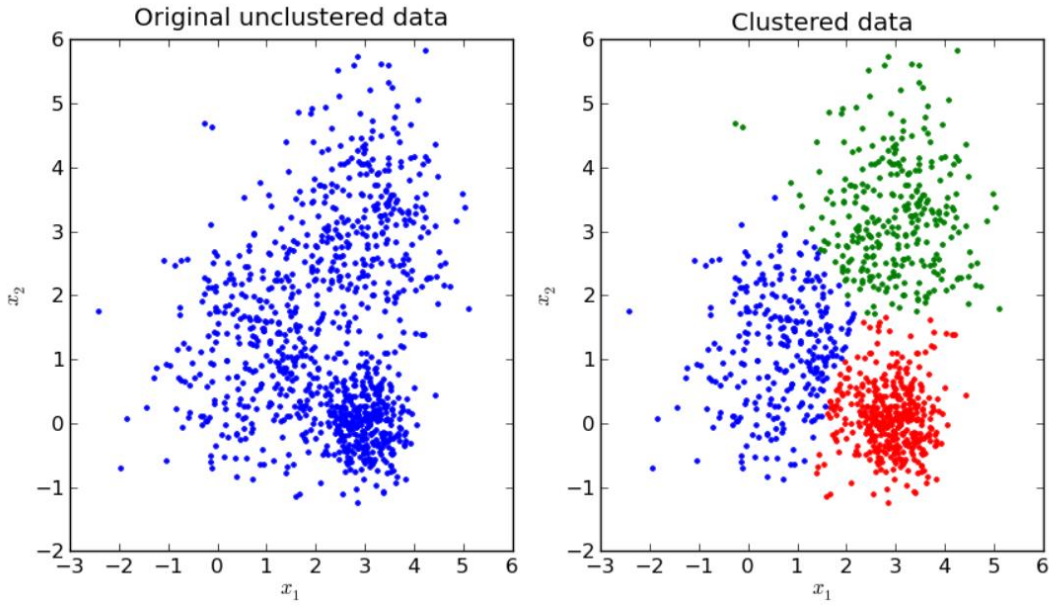
Ορισμένοι βασικοί λόγοι που αναδεικνύουν τη σημασία αλγορίθμων αυτής της κατηγορίας, είναι η χρησιμότητα που προσδίδουν στην αναζήτηση χρήσιμων πληροφοριών απο ένα σύνολο δεδομένων, η μεγάλη ομοιότητα που παρουσιάζουν με την ανθρώπινη διαδικασία αποκόμισης γνώσης, η ικανότητα τους να «μαθαίνουν» δίχως την ρητή σύνδεση δεδομενων με ετικέτες και τέλος το γεγονός ότι στον πραγματικό κόσμο δεν διαθέτουμε πάντα δεδομένα τα οποία συνοδεύονται από μία αντιστοιχη έξοδο.

Παρακάτω παρουσιάζονται συνοπτικά οι κύριες κατηγορίες εργασιών για τις οποίες η χρήση μη επιβλεπόμενης μηχανικής μάθησης χρησιμοποιείται με ιδιαίτερα υψηλή συχνότητα και επιτυχία.

Ομαδοποίηση (Clustering):

Στην ομαδοποίηση, τα δεδομένα εισόδου διαιρούνται αυτόματα σε ομάδες ή κατηγορίες με βάση τις ομοιότητες που παρουσιάζουν μεταξύ τους. Ο στόχος είναι να επιτευχθεί η ανακάλυψη των δομών και των μοτίβων που υπάρχουν στα δεδομένα χωρίς να υπάρχει προκαθορισμένη γνώση για τον αριθμό ή τον τύπο των ομάδων. Χαρακτηριστικό παράδειγμα αποτελεί η ομαδοποίηση παρόμοιων ταινιών με σκοπό να προταθούν στους χρήστες μιας διαδικτυακής πλατφόρμας μετάδοσης εριεχομένου τηλεοπτικών σειρών και ταινιών. Οι αλγόριθμοι που είναι υπεύθυνοι για το clustering, επεξεργάζονται τα δεδομένα και βρίσκουν, αν υπάρχουν, clusters (ομάδες-κλάσεις) μεταξύ αυτών. Ένα cluster είναι, συνεπώς, μια συλλογή αντικειμένων που είναι όμοια μεταξύ τους και ανόμοια με

αντικείμενα που ανήκουν σε άλλα clusters. Στην ακόλουθη εικόνα, φαίνεται πως απο ένα γενικό σύνολο μη ομαδοποιημένων δεδομένων, καταλήγουμε να έχουμε ομάδες στις οποίες διαχωρίζονται.



Σχήμα 5: Clustering Συνόλου δεδομένων [2]

Αξίζει εδώ να σημειωθεί πως μπορούμε να προσαρμόσουμε έναν αλγόριθμο clustering με τρόπο τέτοιο ούτως ώστε να πραγματοποιεί εύρεση συγκεκριμένου αριθμού clusters, γεγονός που επιτρέπει την αλλαγή του βαθμού ανάλυσης των ομάδων αυτών.
Τέλος σημειώνεται πως ανάλογα με τον τρόπο που ομαδοποιούνται τα δεδομένα, υπάρχουν πολλοί τρόποι clustering. Κύριοι εξ αυτών είναι:

Τμηματοποίηση (Partitioning):

Η τμηματοποίηση είναι η πιο βασική και διαδεδομένη μέθοδος ομαδοποίησης. Σε αυτήν την τεχνική, τα δεδομένα διαιρούνται σε μη-επικαλυπτόμενα σύνολα ομάδων, γνωστά και ως κατηγορίες ή clusters. Οι ομάδες διαχωρίζονται με βάση τις ομοιότητες μεταξύ των δεδομένων, όπου τα δεδομένα που είναι πιο κοντά μεταξύ τους θα ανήκουν στην ίδια ομάδα. Ο στόχος είναι να ελαχιστοποιηθεί η διαφορά εντός των ομάδων και να

μεγιστοποιηθεί η διαφορά μεταξύ των ομάδων. Ένας δημοφιλής αλγόριθμος τμηματοποίησης είναι ο K-Means.

Ιεραρχική (Hierarchical) Ομαδοποίηση:

Η ιεραρχική ομαδοποίηση αναπαριστά την ομαδοποίηση σε μορφή δένδρου (ιεραρχίας). Στην αρχή, κάθε δείγμα θεωρείται ως ένα ξεχωριστό δένδρο. Στη συνέχεια, τα δένδρα συγχωνεύονται με βάση την ομοιότητα, δημιουργώντας υποδένδρα. Οι διαδικασίες συγχώνευσης συνεχίζονται μέχρι να δημιουργηθεί ένα μεγάλο δένδρο που αντιπροσωπεύει την ολική ιεραρχία. Η ιεραρχική ομαδοποίηση μπορεί να είναι συσσωματική (agglomerative) ή διαχωριστική (divisive).

Επικαλυπτόμενη (Overlapping) Ομαδοποίηση:

Η επικαλυπτόμενη ομαδοποίηση επιτρέπει σε ένα δείγμα να ανήκει σε περισσότερες από μία ομάδες. Αντίθετα με τις προηγούμενες τεχνικές ομαδοποίησης που χρησιμοποιούν μη-επικαλυπτόμενα σύνολα, εδώ υπάρχει μερική επικάλυψη μεταξύ των ομάδων. Αυτό επιτρέπει την πιο ευέλικτη αναπαράσταση της πολυπλοκότητας των δεδομένων και την αναγνώριση περιοχών που μοιράζονται χαρακτηριστικά μεταξύ διαφορετικών ομάδων.

Πιθανολογική (Probabilistic) Ομαδοποίηση:

Η πιθανολογική ομαδοποίηση χρησιμοποιεί πιθανότητες για να αναθέσει κάθε δείγμα σε μία ομάδα. Αυτή η τεχνική βασίζεται στην υπόθεση ότι τα δεδομένα προέρχονται από κατανομές πιθανοτήτων. Ο αλγόριθμος εκτιμά τις πιθανότητες ομαδοποίησης για κάθε δείγμα, με βάση τα χαρακτηριστικά του. Ένας γνωστός αλγόριθμος πιθανολογικής ομαδοποίησης είναι ο Gaussian Mixture Models (GMM).

Κάθε μία από αυτές τις τεχνικές ομαδοποίησης έχει τα δικά της πλεονεκτήματα και περιορισμούς και επιλέγεται ανάλογα με τη φύση των δεδομένων και τους στόχους του προβλήματος που επιθυμούμε να επιλύσουμε.

Ανιχνευση ανωμαλιών (Anomaly Detection):

Μια άλλη κατηγορία που τοποθετείται κάτω απο την ομπρέλα της μη επιβλεπόμενης μηχανικής μάθησης, είναι το anomaly detection. Η ανίχνευση ανωμαλιών ασχολείται με την αναγνώριση ατυπιών, εξαιρέσεων ή ανωμαλιών στα δεδομένα. Ο στόχος είναι να εντοπιστούν παρατηρήσεις που διαφέρουν από τον αναμενόμενο κανόνα ή πρότυπο, και μπορεί να είναι χρήσιμη για την ανίχνευση απάτης, την παρακολούθηση δικτύων και την ανίχνευση προβλημάτων σε συστήματα. Με την ανίχνευση ανωμαλιών, εντοπίζονται δεδομένα που αποτελούν outliers σε ένα dataset. Για παράδειγμα, ένα outlier σε ένα δίκτυο μπορεί να σημαίνει ότι το χακαρισμένο δίκτυο στέλνει ευαίσθητο περιεχόμενο σε έναν μη εξουσιοδοτημένο εξυπηρετητή ενώ σε ένα σύστημα μέτρησης θερμοκρασίας, outlier μπορεί να αποτελέσει μια ασυνήθιστα υψηλή τιμή.

Διασταυρούμενη ανάλυση (Association Rule Learning):

Η διασταυρούμενη ανάλυση αποσκοπεί στην εύρεση συσχετίσεων και πρότυπων μεταξύ των διαφορετικών μεταβλητών σε μια συλλογή δεδομένων. Συχνά χρησιμοποιείται για την ανακάλυψη κρυμμένων σχέσεων και την εξαγωγή ενδιαφέροντων κανόνων από μεγάλα σύνολα δεδομένων, όπως σε ηλεκτρονικές αγορές για την πρόβλεψη συμπεριφοράς αγοραστών. 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.

Διάσπαση (Dimensionality Reduction):

Η διάσπαση ασχολείται με τη μείωση των διαστάσεων ενός συνόλου δεδομένων, διατηρώντας ωστόσο όλη τη σημαντική πληροφορία που εμπεριέχει. Συχνά χρησιμοποιείται για την αποτύπωση των δεδομένων σε μια χαμηλότερη διάσταση χώρου για ευκολότερη ανάλυση και οπτικοποίηση, καθώς και για την απομάκρυνση θορύβου και

την εξαγωγή των κυρίαρχων χαρακτηριστικών των δεδομένων. Αποσκοπεί στην απλοποίηση των δεδομένων, εξαλείφοντας ή συμπιέζοντας τα χαρακτηριστικά που παρουσιάζουν μικρή πληροφορία ή υψηλή συσχέτιση. Με τον όρο διαστάσεις, αναφερόμαστε στον αριθμό των χαρακτηριστικών που χρησιμοποιούνται για να περιγράψουν ένα σύνολο δεδομένων. Σε πραγματικά προβλήματα, τα δεδομένα μπορεί να έχουν πολύ μεγάλο αριθμό χαρακτηριστικών, κάτι που μπορεί να οδηγήσει σε προβλήματα όπως overfitting του μοντέλου, αυξημένος χρόνος εκτέλεσης και δυσκολία στην ανάλυση και οπτικοποίηση των δεδομένων. Υπάρχουν διάφορες μέθοδοι διάσπασης που χρησιμοποιούνται στην πράξη. Δύο από τις πιο κοινές είναι η ανάλυση κυρίας συνιστώσας (Principal Component Analysis - PCA) και η Linear Discriminant Analysis – LDA στις οποίες όμως δε θα αναφερθούμε περαιτέρω.

3.2.3 Ενισχυτική Μάθηση (Reinforcement Learning)

Η ενισχυτική μάθηση (reinforcement learning) είναι μια κατηγορία μηχανικής μάθησης κατά την οποία ένας πράκτορας (αλγόριθμος) προσπαθεί να μάθει πώς να προσαρμόζεται και να λαμβάνει αποφάσεις σε ένα περιβάλλον με βάση την αλληλεπίδρασή του με αυτό. Αυτή η προσέγγιση είναι εμπνευσμένη από τον τρόπο με τον οποίο μαθαίνουν και λαμβάνουν αποφάσεις οι άνθρωποι μέσω μιας διαδικασίας δοκιμασίας, σφάλματος και ανταμοιβής.

Στην ενισχυτική μάθηση, ο αλγόριθμος παίρνει αποφάσεις δεδομένου ενός περιβάλλοντος και εκτελεί δράσεις. Μετά από κάθε δράση, το περιβάλλον παρέχει μια ανταμοιβή ή αποτίμηση (reward ή αντίθετα penalty) που αντιπροσωπεύει τον βαθμό επίτευξης του στόχου. Ο στόχος του αλγορίθμου είναι να μάθει να προσαρμόζεται κατάλληλα για να επιτύχει υψηλή απόδοση, μεγιστοποιώντας την ανταμοιβή που λαμβάνει από το περιβάλλον.

Για να επιτύχει αυτόν τον στόχο, ο ενισχυτικός αλγόριθμος χρησιμοποιεί μια στρατηγική που συνδυάζει την εξερεύνηση και την εκμάθηση. Στην αρχή, όταν ο αλγόριθμος δεν έχει πολλές πληροφορίες για το περιβάλλον, πρέπει να εξερευνήσει διάφορες δράσεις

για να ανακαλύψει την βέλτιστη στρατηγική. Καθώς αποκτά περισσότερη εμπειρία, αρχίζει να εκτελεί δράσεις που έχουν αποδειχθεί ότι παράγουν μεγαλύτερη ανταμοιβή. Κύρια συστατικά ενός συστήματος ενισχυτικής μάθησης αποτελούν τα ακόλουθα:

- 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.

Οι reinforcement learning αλγόριθμοι συχνά βασίζονται σε μοντέλα που περιγράφουν το περιβάλλον και την αλληλεπίδραση τους με αυτό. Τα μοντέλα μπορούν να χρησιμοποιηθούν για να προβλέψουν τις ανταμοιβές ή τις καταστάσεις που προκύπτουν από μια συγκεκριμένη δράση, επιτρέποντας στον αλγόριθμο να λάβει πιο ενημερωμένες αποφάσεις. Οι αλγόριθμοι ενισχυτικής μάθησης σχετίζονται με τους αλγορίθμους που χρησιμοποιούνται στο δυναμικό προγραμματισμό για τη λύση προβλημάτων βελτιστοποίησης. Δεδομένου ότι 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 καταστάσεων 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 μιας greedy στρατηγικής και της συνεχής εξερέυνησης προκειμένου να επιτύχει ενδεχομένως μακροπρόθεσμα οφέλη.
Αν ο αλγόριθμος επικεντρωθεί υπερβολικά στην εξερεύνηση, μπορεί να καταλήξει να δαπανά πολύ χρόνο σε αναποτελεσματικές δράσεις και να μην εκμεταλλεύεται πλήρως τις γνώσεις που έχει αποκτήσει. Αντίθετα, αν επικεντρωθεί υπερβολικά στην εκμάθηση, μπορεί να μην ανακαλύψει αποτελεσματικές στρατηγικές και να μένει περιορισμένος στην αρχική του απόδοση.

Ένα σημαντικό κριτήριο για την επίλυση αυτού του προβλήματος είναι η χρήση κατάλληλων αλγορίθμων εξερεύνησης-εκμάθησης, όπως ο επιλεκτικός καταναλωτής (επιλέγει τις δράσεις βάσει ενός προκαθορισμένου κριτηρίου) ή ο αλγόριθμος UCB (Upper Confidence Bound) που συνδυάζει την εξερεύνηση και την εκμάθηση.

Επιπλέον, η ρύθμιση των παραμέτρων του αλγορίθμου, όπως ο ρυθμός εξερεύνησης και οι ανταμοιβές, απαιτεί προσοχή και δοκιμή. Οι παράμετροι αυτές πρέπει να επιλεγούν με προσοχή έτσι ώστε να επιτυγχάνεται η ισορροπία μεταξύ εξερεύνησης και εκμάθησης, καθώς και η βελτιστοποίηση της απόδοσης του αλγορίθμου.

Η επιλογή και η προεπεξεργασία των χαρακτηριστικών του περιβάλλοντος μπορεί επίσης να επηρεάσει την απόδοση του αλγορίθμου ενισχυτικής μάθησης. Σωστές επιλογές και προεπεξεργασία μπορούν να βοηθήσουν τον αλγόριθμο να ανακαλύψει αποτελεσματικές στρατηγικές και να βελτιώσει την απόδοσή του.

Τέλος, οι εφαρμογές της ενισχυτικής μάθησης καλύπτουν πολλούς τομείς, όπως την ρομποτική, την αυτοματοποίηση, τα παιχνίδια και την αποφυγή συγκρούσεων. Ένα παράδειγμα είναι ο αλγόριθμος Deep Q-Network (DQN), που μαθαίνει να παίζει παιχνίδια χρησιμοποιώντας ένα νευρωνικό δίκτυο για να εκτιμήσει την αξία κάθε δράσης σε μια συγκεκριμένη κατάσταση.

3.3 Logistic Regression Classifier

Ο λογιστικός (logistic) ταξινομητής (classifier) είναι ένας αλγόριθμος μηχανικής μάθησης που παρά το γεγονός ότι ονομάζεται logistic **regression,** χρησιμοποιείται για την **ταξινόμηση** δεδομένων σε δύο κατηγορίες. Ανήκει στην κατηγορία της επιβλεπόμενης μάθησης, καθώς απαιτεί ετικετοποιημένα δεδομένα (labeled data) για την εκπαίδευσή του, ενώ η βασική ιδέα πίσω από την εφαρμογή του, είναι η χρήση ενός λογαριθμικού μοντέλου στατιστικής ανάλυσης, με σκοπό να προβλεφθεί η πιθανότητα ενός δείγματος να ανήκει σε μία από δύο κατηγορίες (0 ή 1), λαμβάνοντας υπόψιν προηγούμενες παρατηρήσεις ενός συνόλου δεδομένων.

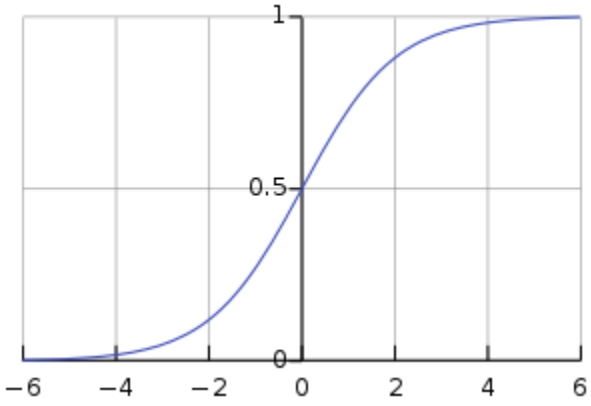
Ένα μοντέλο λογιστικής παλινδρόμησης προβλέπει την τιμή μιας εξαρτημένης μεταβλητής, αναλύοντας τη σχέση μεταξύ μίας ή περισσότερων ανεξάρτητων μεταβλητών. Για παράδειγμα, η λογιστική παλινδρόμηση μπορεί να χρησιμοποιηθεί για να προβλέψει εάν ένας πολιτικός υποψήφιος θα κερδίσει ή θα χάσει μια εκλογική αναμέτρηση ή εάν ένας μαθητής λυκείου θα γίνει δεκτός ή όχι σε ένα συγκεκριμένο πανεπιστήμιο. Στα πλαίσια της παρούσας διπλωματικής εργασίας, το συγκεκριμένο μοντέλο θα χρησιμοποιηθεί όπως θα δούμε και αργότερα, προκειμένου να προβλέψουμε την πιθανότητα ένας πελάτης μιας τράπεζας να προβεί σε αγορά ενός προϊόντος ή υπηρεσίας της τράπεζας στο διάστημα των επόμενων μηνών από τη στιγμή της μελέτης. Αυτά τα δυαδικά αποτελέσματα επιτρέπουν απλές αποφάσεις μεταξύ δύο εναλλακτικών.

Επιπλέον, ένα μοντέλο logistic regression, μπορεί να λαμβάνει υπόψη πολλά κριτήρια ή αλλιώς μεταβλητές εισόδου (διάνυσμα εισόδου). Στην περίπτωση της αποδοχής σε ένα πανεπιστήμιο, η λογιστική συνάρτηση μπορεί να λαμβάνει υπόψη παράγοντες όπως ο μέσος όρος βαθμολογίας του μαθητή, η βαθμολογία SAT και ο αριθμός των εξωσχολικών δραστηριοτήτων. Βασιζόμενο σε ιστορικά δεδομένα για προηγούμενα αποτελέσματα που αφορούν τις ίδιες μεταβλητές εισόδου, αξιολογεί νέα περιστατικά ως προς την πιθανότητα να ανήκουν σε μία από τις δύο κατηγορίες αποτελέσματος. Ο τρόπος με τον οποίο επιτυγχάνεται αυτό εξηγείται μέσω της μελέτης του μαθηματικού υποβάθρου της *λογιστικής-σιγμοειδούς συνάρτησης*:

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

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

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 παράστασης, που παρουσιάζεται στο παρακάτω σχήμα και μοιάζει με «S».



Σχήμα 6: Σιγμοειδής συνάρτηση [3]

Όπως φαίνεται και στο παραπάνω διάγραμμα, η συνάρτηση λαμβάνει τιμές από ένα συνεχές πεδίο τιμών, στο διάστημα [0, 1]. Αξίζει εδώ να σημειώσουμε πως μία πολύ σημαντική απόφαση που πρέπει να ληφθεί κατα την εφαρμογή της συνάρτησης, είναι να οριστεί το κατώφλι (threshold) άνω του οποίου γίνεται η εναλλαγή της προβλεπόμενης τιμής απο 0 σε 1 για τις περιπτώσεις τις δυαδικής ταξινόμησης. Είθισται αυτό το κατώφλι

να επιλέγεται ως το 0.5 δηλαδή το σημείο για το οπόιο φ(υ) = $\frac{1}{1+e^{-\alpha\upsilon}}$ = 0.5 → $e^{-\alpha\upsilon}$= 1 → *αυ=0.* Ανάλογα την περίπτωση μελέτης ωστόσο αυτό μπορεί να αλλάζει κατά βούληση.

Οποιοσδήποτε πρόκειται να χρησιμοποιήσει αυτή τη μέθοδο προκειμένου να προβλέψει ένα αποτέλεσμα, θα πρέπει ωστόσο να λαμβάνει υπόψιν μερικές προϋποθέσεις κατά τη χρήση της λογιστικής παλινδρόμησης. Αρχικά, οι μεταβλητές πρέπει να είναι ανεξάρτητες μεταξύ τους. Για παράδειγμα, ο ταχυδρομικός κώδικας και το φύλο ενός ανθρώπου μπορούν να χρησιμοποιηθούν σε ένα μοντέλο, αλλά ο ταχυδρομικός κώδικας και η πολιτεία όχι.

Άλλες λιγότερο εμφανείς συσχετίσεις μεταξύ των μεταβλητών εισόδου μπορεί να χαθούν όταν η λογιστική παλινδρόμηση χρησιμοποιείται ως σημείο αναφοράς για περίπλοκες εφαρμογές μηχανικής μάθησης. Για παράδειγμα, οι data scientists μπορεί να πρέπει αφιερώσουν σημαντική προσπάθεια για να διασφαλίσουν ότι μεταβλητές που σχετίζονται με διακρίσεις, όπως το φύλο και η εθνικότητα, δεν θα πρέπει να συμπεριλαμβάνονται στο μοντέλο. Ωστόσο, αυτές μπορεί να ενσωματωθούν κατά λάθος στον αλγόριθμο μέσω μεταβλητών που δεν θεωρούνταν συσχετισμένες, όπως ο ταχυδρομικός κώδικας, η σχολή ή τα χόμπι.

Είναι επίσης σημαντικό η σχέση μεταξύ των μεταβλητών και του αποτελέσματος να υφίσταται μέσω λογαριθμικών πιθανοτήτων, που είναι λίγο πιο ευέλικτες έναντι μιας γραμμικής σχέσης.

Μια άλλη προϋπόθεση με τη λογιστική παλινδρόμηση είναι ότι κάθε μεταβλητή πρέπει να μπορεί να αναπαρασταθεί χρησιμοποιώντας δυαδικές κατηγορίες, όπως αρσενικό/θηλυκό, αγορά/πώληση. Επιπλέον απαιτείται ειδικός χειρισμός σε περιπτώσεις αναπαράστασης κατηγοριών με περισσότερες από δύο κλάσεις. Μπορεί, για παράδειγμα, να μετατρέψουμε μία κατηγορία με τρείς διαφορετικές εθνικότητες, σε τρεις ξεχωριστές μεταβλητές, όπου καθεμία καθορίζει εάν ένα άτομο προέρχεται από αυτή τη χώρα ή όχι. Η τεχνική αυτή ονομάζεται one-hot-encoding και θα αναλυθεί περαιτέρω αργότερα κατά την παρουσίαση του δικού μας πεδίου εφαρμογής της συγκεκριμένης μεθόδου.

Η λογιστική παλινδρόμηση (Logistic regression) αποτελεί στην ουσία ένα μοντέλο ταξινόμησης των τιμών μιας μεταβλητής απόκρισης Υ με βάση τη θεωρία των πιθανοτήτων. Στο μοντέλο αυτό όπου η μεταβλητή Υ συνήθως έχει δυαδικό χαρακτήρα (λαμβάνει δύο

τιμές) στοχεύεται η πρόβλεψη της έκβασης αυτής από ένα πλήθος προβλεπτικών μεταβλητών που μπορεί να είναι ονομαστικές, τακτικές ή ποσοτικές. Η σημαντικότερη διαφοροποίηση μεταξύ λογιστικής και γραμμικής παλινδρόμησης βασίζεται στη φύση της επιλεγμένης μεταβλητής απόκρισης, η οποία στην μεν πρώτη μπορεί να είναι κατηγορική, (τακτική ή ονομαστική, στη δε δεύτερη αποκλειστικά ποσοτική. Ενώ κατά την κλασική γραμμική παλινδρόμηση η εκτίμηση των παραμέτρων α και bi γίνεται με τη μέθοδο των ελάχιστων τετραγώνων, κατά τη λογιστική παλινδρόμηση η εκτίμηση των παραμέτρων γίνεται με τη μέθοδο του λόγου πιθανοφάνειας (μέθοδος συνήθως εφαρμοζόμενη στα γενικευμένα γραμμικά υποδείγματα), δηλαδή επιλέγονται οι πιο πιθανοφανείς τιμές των παραμέτρων, προκειμένου να οδηγήσουν στα παρατηρούμενα αποτελέσματα. Ως επακόλουθο, η πρώτη παραδέχεται την ύπαρξη ομοιογένειας (ομοσκεδαστικότητας) στα υπολείμματα των αποκρίσεων ενώ στη δεύτερη αναπτύσσεται πάντα ετεροσκεδαστικότητα σε κάθε προβλεπόμενη τιμή εξαιτίας του μεταβαλλόμενου ποσοστού διακύμανσης που αναλογεί σε αυτήν.

Διακρίνονται τρεις τύποι λογιστικής παλινδρόμησης ανάλογα με την ιδιαίτερη φύση της εξαρτημένης κατηγορικής μεταβλητής η οποία μπορεί να είναι:

1. Δίτιμη ή δυαδική ή διχοτομική (binary) ή διμερής εξαρτημένη μεταβλητή. Συνίσταται από δύο κατηγορίες, όπως π.χ. είναι οι εκβάσεις επιτυχία/αποτυχία, ΝΑΙ/ΟΧΙ, γεγονός απόν/παρόν.

2. Τακτική (ordinal) μεταβλητή. Η εξαρτημένη μεταβλητή συνίσταται από τρεις ή περισσότερες κατηγορίες μεταξύ των οποίων ισχύει η έννοια της ανισότητας, όπως π.χ. σε μια ερώτηση της κλίμακας διαφωνώ καθόλου, λίγο, μέτρια, αρκετά, πολύ, στην κατάταξη ενός στρώματος υλικού ως λεπτού, μεσαίου, παχέος.

3. Ονομαστική (Nominal) ή πολυωνυμική (polynomial) ή πολυχοτομική (polychotomus) ή κατηγορική αδιαβάθμητη (non-ordered categorical) ή πολυμερής μεταβλητή απόκρισης. Περιέχει τρεις ή περισσότερες κατηγορίες χωρίς κάποια φυσική διαβάθμιση, όπως π.χ. ο χαρακτηρισμός ενός τροφίμου ως τραγανού, μαλακού, εύθρυπτου ή του χρώματος αντικειμένων ως ερυθρού, πράσινου, κίτρινου κτλ.

Η λογιστική παλινδρόμηση επινοήθηκε ως εναλλακτική επιλογή της γραμμικής διακριτικής ανάλυσης για την ταξινόμηση των στοιχείων (ονομαστικών ή τακτικών) της

εξαρτημένης, με ευρεία απήχηση σε πολλά διαφορετικά επιστημονικά πεδία και κυρίως στην ιατρική και τις κοινωνικές επιστήμες.
Χαρακτηριστικά, χρησιμοποιείται στην πρόβλεψη της:

- εμφάνισης ή μη μιας νόσου (π.χ. διαβήτης) από ένα σύνολο διαφορετικών χαρακτηριστικών του πάσχοντος ατόμου (ηλικία, φύλο, αιματολογικά, ηλεκτροκαρδιογράφημα κτλ.)
- επιλογής ενός πολιτικού κόμματος με βάση την καταγραφή των δημογραφικών στοιχείων των πολιτών, όπως είναι η ηλικία, φύλο, φυλή, τόπος διαμονής, εισόδημα, προηγούμενη ψηφοφορία
- πιθανότητας αποτυχίας μιας διεργασίας παραγωγής προϊόντος σε ένα εργοστάσιο τροφίμων
- πρόβλεψη της πρόθεσης αγοράς ενός αγαθού από έναν καταναλωτή (έρευνα αγοράς)
- πιθανότητας αθέτησης από δανειολήπτη της αποπληρωμής του δανείου του

Η κατανόηση των όρων και μαθηματικών τύπων που συνοδεύουν τη μελέτη της λογιστικής παλινδρόμησης αποτελεί κυριολεκτικά πρόκληση για τον απλό επιστήμονα. Ως εκ τούτου, στο κεφάλαιο αυτό γίνεται αναφορά μόνο των αναγκαίων τεχνικών εφαρμογής της λογιστικής παλινδρόμησης με φειδωλή χρήση των μαθηματικών τύπων.

Στη γλώσσα της στατιστικής, η λογιστική παλινδρόμηση χρησιμοποιείται για την πρόβλεψη της πιθανότητας εμφάνισης ενός γεγονότος προσαρμόζοντας τα δεδομένα της μελέτης στην εξίσωση της λογιστικής καμπύλης όπως αυτή παρουσιάζεται στο σχήμα 6. Η καμπύλη αυτή έχει σιγμοειδή μορφή και χαρακτηρίζεται από ένα στάδιο εκθετικής ανάπτυξης στο οποίο ο ρυθμός αύξησης επιβραδύνεται βαθμιαία και περατώνεται στο ασυμπτωτικό στάδιο κορεσμού της ανάπτυξης (η ευθεία βαίνει τελικά παράλληλα στον άξονα Χ).

Η δυαδική λογιστική παλινδρόμηση αποτελεί μια διωνυμική εξίσωση στην οποία η μεταβλητή απόκρισης Υ είναι το τυχαίο αποτέλεσμα εμφάνισης μιας από δύο δυνητικές εκβάσεις του τύπου επιτυχία ή αποτυχία όπως π.χ. είναι το αποτέλεσμα της ρίψης ενός

νομίσματος δύο διαφορετικών όψεων (κορώνα-γράμματα), η ρίψη ενός ζαριού όπου το αποτέλεσμα εμφάνισης του αριθμού 6 θεωρείται επιτυχία και των λοιπών αριθμών αποτυχία, η θετική ψήφος εκλογής ενός πολιτικού εκπροσώπου κτλ.
Η δίτιμη λογιστική παλινδρόμηση έχει τη μορφή:

$$f(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

όπου z είναι η μεταβλητή εισόδου και f(z) το αποτέλεσμα αυτής. Στα πλεονεκτήματα της εξίσωσης συγκαταλέγεται και το γεγονός ότι η μεταβλητή εισόδου λαμβάνει θετικές και αρνητικές τιμές ενώ το αποτέλεσμα αυτής f(z) περιορίζεται σε εύρος τιμών μεταξύ 0 και 1. Αναλυτικότερα, η μεταβλητή z εκπροσωπεί τη δράση μιας ομάδας ανεξάρτητων μεταβλητών ενώ η f(z) προσδιορίζει την πιθανότητα ενός συγκεκριμένου αποτελέσματος λόγω της δράσης της ομάδας αυτής. Η μεταβλητή z (λογιστική) εκφράζει επίσης το μέτρο της ολικής συνεισφοράς όλων των συμμετεχουσών ανεξάρτητων μεταβλητών στο μοντέλο και ορίζεται ως:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \ldots + \beta_\kappa X_\kappa$$

όπου β_0 είναι το ύψος της κλίσης της γραμμής παλινδρόμησης και ισούται με την τιμή z όταν οι τιμές όλων των ανεξάρτητων μεταβλητών ισούνται με 0, ενώ β_i είναι οι συντελεστές παλινδρόμησης καθένας των οποίων εκφράζει το μέγεθος συνεισφοράς της αντίστοιχης μεταβλητής. Θετική τιμή του συντελεστή δηλώνει ότι η επεξηγηματική μεταβλητή αυξάνει την πιθανότητα της επιτυχημένης έκβασης (να συμβεί δηλαδή το γεγονός), αρνητική τιμή σημαίνει ότι η μεταβλητή μειώνει την πιθανότητα αυτής της έκβασης. Υψηλή τιμή του συντελεστή σημαίνει ότι η ανεξάρτητη μεταβλητή επηρεάζει πολύ ισχυρά την πιθανότητα να συμβεί το γεγονός ή μη, ενώ χαμηλή τιμή δηλώνει μικρή επίδραση της ανεξάρτητης μεταβλητής στην πιθανότητα εμφάνισης της ανάλογης έκβασης.

Συνοψίζοντας, η λογιστική παλινδρόμηση χρησιμεύει στην περιγραφή της σχέσης που αναπτύσσεται μεταξύ μιας ή περισσότερων ανεξάρτητων μεταβλητών (π.χ. ηλικία, φύλο, τοξική συγκέντρωση ουσίας) και μιας δυαδικής μεταβλητής απόκρισης εκφρασμένης

ως πιθανότητα δυνάμενη να πάρει μία από δύο τιμές, όπως π.χ. θετική (1) αρνητική (0), παρόν ενδεχόμενο (1) απόν ενδεχόμενο (0), επιζών (1) θανών (0), αρεστός (1) δυσάρεστος (0).

Η φύση των ανεξάρτητων μεταβλητών εισόδου στην εξίσωση της πολλαπλής λογιστικής παλινδρόμησης μπορεί να είναι ποσοτική, τακτική ή ονομαστική (αδιαβάθμητη κατηγορική). Για παράδειγμα, η πιθανότητα ένα άτομο να υποστεί καρδιακό επεισόδιο σε συγκεκριμένο χρονικό διάστημα (εξαρτημένη μεταβλητή) μπορεί να προβλεφθεί από ένα πλήθος ανεξάρτητων μεταβλητών όπως είναι η ηλικία, το φύλο, ο δείκτης μάζας σώματος, η φυσική αγωγή, το ιστορικό του ασθενούς κτλ. Η ηλικία ενδέχεται να ενταχθεί στη εξίσωση είτε ως ποσοτική (με την πραγματική τιμή της) είτε ως τακτική: 15-25 ετών, 25-40, 40-60, >60. Η φυσική αγωγή ως ονομαστική μεταβλητή (άθληση ή μη), ο δείκτης μάζας σώματος (Body Mass Index - BMI) ως ποσοτική ή τακτική (30), και το ιστορικό ως ονομαστική (ύπαρξη προδιάθεσης ή μη). Επιστήμες όπως η ιατρική, οι κοινωνικές επιστήμες και το marketing καταφεύγουν συχνά στην εφαρμογή της πολυωνυμικής λογιστικής παλινδρόμησης.

Οι πιθανότητες που συγκλίνουν υπέρ της εμφάνισης ενός γεγονότος ή πρόθεσης εκφράζονται ως λόγος ζεύγους ακέραιων τιμών (odds) όπου ο αριθμητής προσδιορίζει την πιθανότητα που έχει το προσδοκώμενο γεγονός να συμβεί και ο παρονομαστής την πιθανότητα να μη συμβεί. Έτσι, αν p είναι η πιθανότητα να εμφανιστεί το γεγονός και 1-p η πιθανότητα να μη συμβεί τότε ο λόγος των πιθανοτήτων θα είναι p/(1-p). Για παράδειγμα, η πιθανότητα να ανασυρθεί μια κάρτα σπαθί από μια τράπουλα 52 φύλλων είναι 25% δηλαδή μία στις τέσσερις ή αριθμητικά 13/52=1:4 ή και ¼. Με ανάλογο τρόπο εκφράζεται και η πιθανότητα μη εμφάνισης μιας κάρτας σπαθί η οποία ισούται με 4:1, αντιστρέφοντας απλώς τους όρους του κλάσματος, (1-p)/p.

Η παραπάνω σχέση (logit) κάλλιστα μπορεί να ενσωματωθεί στο μοντέλο της παλινδρόμησης σε λογαριθμική μορφή ως εξής,

$$\text{Logit (p)} = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \ldots + \beta_\kappa X_\kappa$$

Οι συντελεστές της παλινδρόμησης υπολογίζονται με τη βοήθεια της εκτίμησης της **μέγιστης πιθανοφάνειας (Maximum Likelihood Estimate – MLE)**, ως:

$$L = \prod_{i=1}^{n} f(x_i \theta)$$

ή προτιμότερο με τη λογαριθμική έκδοση αυτής,

$$L = \sum_{i=1}^{n} \ln\ f(x_i \theta)$$

όπου θ είναι μια παράμετρος της μεταβλητής η οποία μπορεί να μεταβάλλεται ελεύθερα. Η προβλεπόμενη τιμή για κάθε παρατήρηση θα ισούται με

$$l = \frac{1}{n} \ln (L)$$

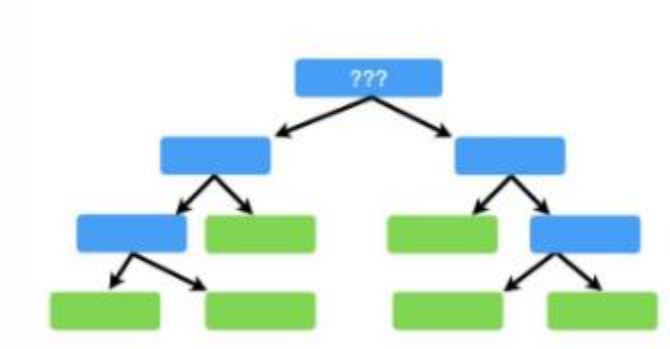
Η συνάρτηση της πιθανοφάνειας έκβασης ενός γεγονότος (likelihood) δείχνει πόσο κατάλληλα ένα παρατηρούμενο δείγμα περιγράφεται από κάποιες τιμές παραμέτρων π.χ. μέσος όρος, τυπική απόκλιση. Άρα, η μεγιστοποίηση της συνάρτησης της πιθανότητας έκβασης καθορίζει τις παραμέτρους εκείνες που είναι οι πλέον ικανές να παράγουν τα παρατηρούμενα στοιχεία. Από άποψη στατιστικής βαρύτητας, η MLE προτείνεται για εφαρμογές σε μεγάλα δείγματα καθόσον είναι ευέλικτη, προσαρμόζεται εύκολα στην παραγωγή πολλών διαφορετικού τύπου μοντέλων, το χειρισμό διαφορετικής φύσης στοιχείων και περιέχει ακριβέστερες μετρήσεις. Η αξιοπιστία των αποτελεσμάτων της λογιστικής παλινδρόμησης επηρεάζεται κατά πολύ από το δειγματοληπτικό μέγεθος της έρευνας. Ένας χρυσός κανόνας υπαγορεύει την αντιστοιχία του αριθμού των επιθυμητών εκβάσεων προς τον αριθμό των ανεξάρτητων μεταβλητών να προσδιορίζεται από τη σχέση 10:1. Εάν υπάρχουν ονομαστικές ανεξάρτητες μεταβλητές, όπως, για παράδειγμα, διχοτομικές, ο παραπάνω κανόνας θα ισχύει για το μέγεθος των παρατηρήσεων της ολιγοπληθέστερης κατηγορίας.

Για δίτιμες εξαρτημένες μεταβλητές, η άριστη άμεση πρόβλεψη της συμμετοχής μιας μεταβλητής ως μέλους σε ομάδα με τη μέθοδο της διακριτικής ανάλυσης επιβάλλει την

ύπαρξη της πολυμεταβλητής κανονικότητας των ανεξάρτητων μεταβλητών αφενός και την ισότητα των διακυμάνσεων-συνδιακυμάνσεων (ομοιογένεια) στις δύο ομάδες, προϋποθέσεις όχι απαραίτητες κατά τη δίτιμη λογιστική παλινδρόμηση. Συμπερασματικά, για την εφαρμογή του υποδείγματος της δίτιμης λογιστικής παλινδρόμησης είναι αναγκαίες πολύ λιγότερες προϋποθέσεις από αυτές που απαιτεί η διακριτική ανάλυση. Μάλιστα, ακόμη κι αν ικανοποιούνται όλες οι προϋποθέσεις για την εφαρμογή της διακριτικής ανάλυσης, η δίτιμη λογιστική παλινδρόμηση λειτουργεί εξαιρετικά καλά, με την απαραίτητη, βέβαια, προϋπόθεση ότι το μέγεθος των παρατηρήσεων είναι τουλάχιστον 10-20 ανά ανεξάρτητη μεταβλητή.

3.4 Δένδρα αποφάσεων

Τα Δέντρα Απόφασης (Decision Trees) ή Δέντρα Ταξινόμησης (Classification Trees) είναι πιθανόν ένα από τα πιο συχνά χρησιμοποιούμενα μοντέλα της Επιβλεπόμενης Μηχνανικής Μάθησης, τόσο για τα αξιόπιστα αποτελέσματα που προσφέρουν, όσο και για την εύκολη ερμηνεία τους. Τα Δέντρα Απόφασης χρησιμοποιούνται για το διαχωρισμό του συνόλου δεδομένων σε κλάσεις που ανήκουν στη μεταβλητή στόχο. Συνήθως, η συνάρτηση στόχος έχει δύο κλάσεις, 0 (αληθές) ή 1 (ψευδές). Ωστόσο και στη περίπτωση που η συνάρτηση στόχος έχει παραπάνω από δύο κλάσεις έχουν δημιουργηθεί αλγόριθμοι για την αντιμετπώπιση τέτοιων προβλημάτων [4]. Ακόμη, τα Δέντρα Απόφασης μπορούν να χρησιμοποιουθούν και στη περίπτωση που η συνάρτηση στόχος παίρνει αριθμητικές τιμές, τότε έχουμε τα λεγόμενα Δέντρα Παλινδρόμησης (Regression Trees) [5]. Επομένως, η φυσική ερμηνεία της μεταβλητής είναι αυτή που θα καθορίσει το τύπο του Δέντρου Απόφασης. Σε αυτή την ενότητα θα περιγράψουμε τη δομή των Δέντρων Αποφάσεων για κατηγοριοποίηση (classification) και γνωστούς αλγορίθμους αυτών, όπως ο Αλγόριθμος C4.5 και ID3 [6].



Σχήμα 7: Δέντρο απόφασης

Όπως δηλώνει και το όνομα των Δέντρων Αποφάσεων, το μοντέλο προς εκπαίδευση αναπαριστάται σε δενδρική μορφή με εσωτερικούς και τερματικούς κόμβους. Κάθε εσωτερικός κόμβος, συμπεριλαμβανομένης και της ρίζας, αναπαριστά ένα από τα χαρακτηριστικά του συνόλου εκπαίδευσης, ενώ οι τερματικοί κόμβοι ή αλλίως φύλλα αντιστοιχούν στις τιμές των κατηγοριών που έχουν οριστεί. Τα χαρακτηριστικά εισόδου μπορούν να είναι διακριτά ή συνεχή, ομοίως και τα χαρακτηριστικά της τιμής των φύλλων. Στην περίπτωση που η τιμή εξόδου είναι διακριτή τιμή τότε έχουμε Κατηγοριοποίηση, ενώ όταν η τιμή εξόδου είναι συνεχής συνάρτηση τότε έχουμε Παλινδρόμηση.

Ένα πολύ βασικό χαρακτηριστικό στα Δέντρα Απόφασης που πρέπει να αναφέρουμε, πρίν προχωρήσουμε στο τρόπο δημιουργίας τους, είναι η πολυπλοκότητα που εμφανίζουν καθώς έχει κρίσιμη επίδραση στην ακρίβεια του μοντέλου. Αναμενόμενο είναι να προτιμούμε ένα δέντρο μικρού μεγέθους για να διατηρούμε την φυσική του ερμηνεία. Προφανώς ένα πιο περίπλοκο δέντρο εξασφαλίζει τις περισσότερες φορές μια ελαφρώς καλύτερη ακρίβεια πρόβλεψης, όμως σε σύγκριση με την αύξηση της πολυπλοκότητας του είναι προτιμότερο. Ο πιο συνήθης τρόπος για τον υπολογισμό της πολυπλοκότητάς του είναι: το πλήθος των κόμβων, το πλήθος των φύλλων, το βάθος του δέντρου και το πλήθος των χαρακτηριστικών που χρησιμοποιήθηκαν για τη δημιουργία του. Επομένως, η πολυπλοκότητα ενός Δέντρου Απόφασης εξαρτάται από το πλήθος των χαρακτηριστικών διάσπασης n, το μέγεθος k του συνόλου εκπαίδευσης καθώς και το σχήμα του παραγόμενου Δέντρου Απόφασης. Έτσι, η πολυπλοκότητα χρόνου για τη δημιουργία του μοντέλου είναι $O(n \cdot k \cdot \log(k))$ καθώς και η πολυπλοκότητα χρόνου πρόβλεψης ενός συνόλου n

παραδειγμάτων εξαρτάται από το ύψος του δέντρου και είναι O(n · log(k)) υποθέτοντας τη χειρότερη περίπτωση για την πολυπλοκότητα ύψους O(log(k)). Ο αναγνώστης μπορεί να δεί το Αλγοριθμικό Σχήμα 1 για τη δημιουργία ενός Δέντρου Απόφασης.

Αλγόριθμος 1: Αλγόριθμος Δημιουργίας Δέντρου Απόφασης

B1. Αρχικοποίηση με ένα κόμβο που περιέχει όλες τις εγγραφές

B2. Διάσπαση του κόμβου με βάση κάποιου κριτήριο διαχωρισμού σε κάποιο από τα γνωρίσματα. Επιλέγεται το καλύτερο γνώρισμα διαχωρισμού

B3. Αναδρομική επανάληψη του B2

B4. Επανάληψη της διαδικασίας έως ότου ικανοποηθεί κάποιο κριτήριο τερματισμού

Σχήμα 8: Αλγοριθμικό σχήμα-Δημιουργία Δέντρου Απόφασης

Είναι φανερό από το παραπάνω ψευδοκώδικα ότι για το διαχωρισμό των γνωρισμάτων είναι απαραίτητο να ορίσουμε μια ποσότητα που θα μας παρέχει πληροφορία για το γνώρισμα που παρέχει πληροφορία ώστε να αντιστοιχηθεί με τον αντίστοιχο κόμβο. Τη πληροφορία αυτή για το κάθε γνώρισμα, μας τη δίνει το στατιστικό εργαλείο κέρδος εντροπίας (information gain), η τιμή του οποίου δείχνει κάτα πόσο αυτό το γνώρισμα είναι ικανό να διαχωρίσει το σύνολο εκπαίδευσης στις επιθυμητές κατηγορίες. Για τον αυστηρό ορισμό του κέρδου εντροπίας είναι απαραίτητο να ορίσουμε την έννοια της εντροπίας (entropy) [7].

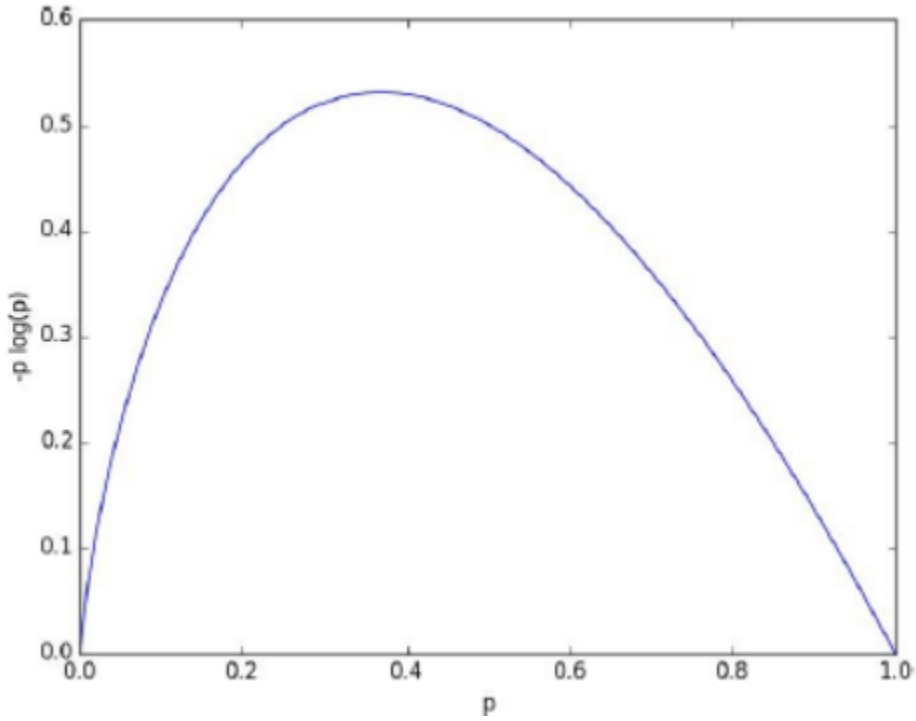
Η εντροπία είναι ενα βασικό εργαλείο για τη θεωρία της πληροφορίας και αποτελεί ένα μέτρο της αβεβαιότητας για ένα σύνολο δεδομένων. Δοθέντος ενός δίτιμου συνόλου δεδομένων S, δηλαδή που διακρίνεται σε δύο κατηγορίες, έστω A και B. Τότε η εντροπία είναι ίση με:

$$E(S) = -P(A)\ log_2(P(A)) - P(B)\ log_2(P(B))$$

όπου P(A), P(B) το ποσοστό των παρατηρήσεων που ανήκουν στη κατηγορία A και B αντίστοιχα. Γενικότερα,

$$E(S) = -\sum_{i=1}^{k} p_i \, log_2 \, (p_i)$$

όπου k είναι οι κλάσεις της συνάρτησης στόχου και p_i το ποσοστό των δειγμάτων από το σύνολο δεδομένων που ανήκουν στην i κατηγορία [48]. Ο αναγνώστης μπορεί να δεί τη γραφική παράσταση του −p log(p) στο ακόλουθο σχήμα.



Σχήμα 9: γράφημα του −p log(p)

Είναι χρήσιμο να τονίσουμε ότι χρησιμοποιούμε το λογάριθμο με βάση 2, γιατί ο υπολογιστής αναγνωρίζει μόνο τις τιμές 0 και 1. Ακόμη, ορίζουμε τη πράξη 0 log (0) = 0. Τέλος, όπως είναι φανερό και από το Σχήμα 8, η εντροπία παίρνει τιμές ανάμεσα στο 0 και 1, όσο μεγαλύτερη είναι η εντροπία τόσο πιο αβέβαιοι είμαστε για τις προβλέψεις μας για μια νέα παρατήρηση.

Με βάση την εντροπία λοιπόν, ορίζουμε ως κέρδος πληροφορίας για ένα γνώρισμα A τη παρακάτω ποσότητα,

$$\text{Gain (S, A)} = \text{E(S)} - \sum_{i\ E\ A} \frac{|\Sigma_i|}{|\Sigma|} E(\Sigma_i)$$

Όπου, $|\Sigma_i|$ είναι το πλήθος των δεδομένων του S για τα οποία η τιμή του γνωρίσματος A είναι ίση με i. Αντίστοιχα, $\frac{|\Sigma_i|}{|\Sigma|}$ είναι το ποσοστό των παρατηρήσων για τα οποία το γνώρισμα A έχει τιμή ίση με i. Σε κάθε βήμα επιλέγεται για κόμβος το γνώρισμα με το μεγαλύτερο κέρδος πληροφορίας. Ακόμη, ένα άλλο μέτρο της εντροπίας είναι το Gini Index, το οποίο ορίζεται ως:

$$\text{Gini} = 1 - \sum_{i=1}^{k} p_i^2$$

Η τιμή του Gini κυμαίνεται μεταξύ 0 και 0.5, αλλά με διαφορετικές ιδιότητες από αυτή της εντροπίας. Οποιαδήποτε στατιστικό εργαλείο από τα δύο παραπάνω μπορεί να χρησιμοποιηθεί για το καθορισμό των κόμβων [8].

Η παραπάνω διαδικασία, όπως περιγράψαμε και στον ψευδοκώδικα, επαναλαμβάνεται εώς ότου ικανοποιηθεί κάποιο κριτήριο τερματισμού. Τα πιο σύνηθες κριτήρια τερματισμού είναι τα παρακάτω:

a. Όλα τα ενδεχόμενα στο σύνολο των εκπαιδευτικών δεδομένων ανήκουν σε μία μόνο κατηγορία.
b. Το μέγιστο βάθος του δέντρου, το οποίο δίνεται από τον χρήστη να έχει επιτευχθεί.
c. Ο αριθμός των περιπτώσεων στο τερματικό κόμβο είναι μικρότερος από τον ελάχιστο αριθμό των υποθέσεων των μητρικών κόμβων.

3.4.1 Βασικοί αλγόριθμοι εκμάθησης δένδρων

Αλγόριθμος ID3

Ο αλγόριθμος (ID3 (Iterative Dichotomiser 3) είναι ένας αλγόριθμος που εφευρέθηκε από τον Ross Quinlan για τη δημιουργία ενός Δέντρου Αποφάσης [9]. Ο Αλγόριμος ID3 ιστορικά εμφανίστηκε νωρίτερα από τον Αλγόριθμο C4.5 και χρησιμοποιείται στους τομείς της Μηχανικής Μάθησης και της επεξεργασίας φυσικής γλώσσας. Ο Αλγόριθμος κατά την υλοποίηση του προσπαθεί να προσδιορίσει κάθε φορά εκείνη τη παράμετρο που του δίνει τη μεγαλύτερη πληροφορία. Συνεπώς κατά τη δημιουργία ενός κόμβου υπολογίζεται το κέρδος πληροφορίας για κάθε παράμετρο και επιλέγεται εκείνη η παράμετρος που παρέχει τη περισσότερη πληροφορία.

Ο Αλγόριθμος ID3 είναι αρκετά εύκολος στη κατανόηση του τρόπου δημιουργίας ενός Δέντρου Απόφασης, καθώς και εύκολα προγραμματίσιμος σε κάποια γλώσσα προγραμματισμού. Για αυτό το λόγο πολλές μέθοδοι βασίστηκαν στη λογική αυτού του αλγορίθμου αντιμετωπίζοντας όμως τα μειονεκτήματα του αλγορίθμου. Ένα βασικό μειονέκτημα του αλγορίθμου είναι οτι δεν μας προσφέρει το ελάχιστο δέντρο. Αυτό οφείλεται στην εύρεση του βέλτιστου τρόπου για το διαχωρισμό των τιμών κάθε παραμέτρου. Ακόμη, προβλήματα συναντόνται και στη περίπτωση που οι μεταβλητές είναι συνεχής, γιατί τότε πρέπει να διαχωρίσουμε το σύνολο χωρίς να χαθεί πληροφορία. Σε μια τέτοια περίπτωση θα πρέπει πρώτα να διακριτοποιούμε τα δεδομένα μας. Για καλύτερη κατανόηση, παραθέτουμε το Αλγοριθμικό Σχήμα 2 του Αλγορίθμου ID3:

Αλγόριθμος 2: Αλγόριθμος ID3

ID3(Παραδείγματα, Ιδιότητα-Στόχος, Ιδιότητες)

B1. Δημιούργησε ένα αρχικό κόμβο για το δέντρο

B2. Εάν όλα τα παραδείγματα είναι θετικά, Επέστρεψε την Ρίζα ως δέντρο με ένα κόμβο, με ετικέτα = +

B3. Εάν όλα τα παραδείγματα είναι αρνητικά, Επέστρεψε την Ρίζα ως δέντρο με ένα κόμβο, με ετικέτα = -

B4. Εάν ο αριθμός των προβλεπόμενων ιδιοτήτων είναι κενός, τότε Επέστρεψε την Ρίζα ως δέντρο με ένα κόμβο, με ετικέτα = την πιο κοινή τιμή της ιδιότητας-στόχου των παραδειγμάτων. Αλλιώς Ξεκίνα

B5. A = Η ιδιότητα που κατηγοριοποιεί καλύτερα τα παραδείγματα

B6. Ιδιότητα Δέντρου απόφασης για τη ρίζα = A

B7. Για κάθε πιθανή τιμή, v_i, του A,

B8. Πρόσθεσε ένα νέο κλάδο κάτω από τη Ρίζα, που να αντιστοιχεί στη δοκιμή A = v_i

B9. Θέσε Παραδείγματα(v_i), ως το υποσύνολο των παραδειγμάτων που έχουν την τιμή v_i για το A

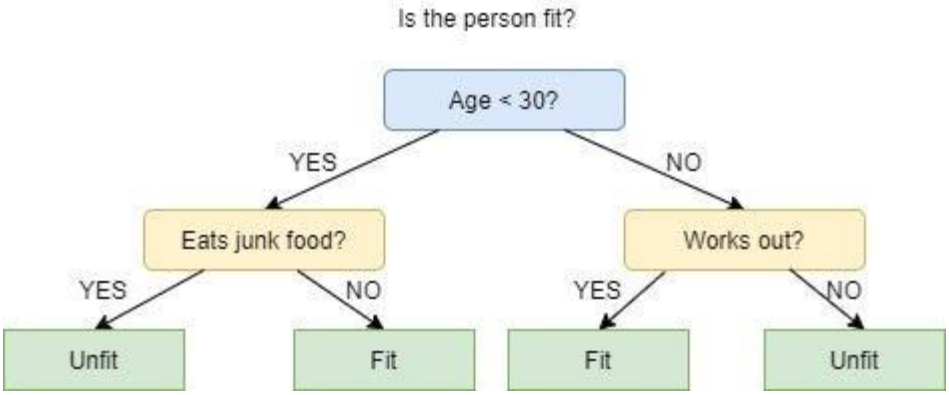
B10. Αν Παραδείγματα(v_i) είναι κενό, τότε κάτω από αυτό τον νέο κλάδο πρόσθεσε έναν κόμβο-φύλλο με ετικέτα = την πιο κοινή τιμή στόχο στα παραδείγματα. Αλλιώς κάτω από αυτό τον νέο κλάδο πρόσθεσε το υποδέντρο ID3(Παραδείγματα(v_i), Ιδιότητα-Στόχος, Ιδιότητες – A)

B11. Τέλος

B12. Επέστρεψε Ρίζα

Σχήμα 10: Αλγοριθμικό σχήμα- ID3

Παρακάτω παρουσιάζεται σχηματικά ένα παράδειγμα εφαρμογής του αλγορίθμου ID3 για την εξαγωγή του αν ένα άτομο έχει καλή φυσική κατάσταση, με βάση ορισμένες άλλες μεταβλητές όπως η ηλικία, η άθληση και οι διατροφικές συνήθειες.



Σχήμα 11: Παράδειγμα εφαρμογής αλγορίθμου ID3

Αλγόριθμος C4.5

Ο αλγόριθμος C4.5 αποτελεί επέκταση του Αλγορίθμου ID3 ο οποίος δημιουργήθηκε από τον Ross Quinlan για να επιφέρει λύσεις σε προβλήματα που εμφανίστηκαν από τον προγενέστερο αλγόριθμο [10]. Τα Δέντρα Απόφασης που παράγονται από τον αλγόριθμο C4.5 κατά κύριο λόγο χρησιμοποιούνται για ταξινόμηση και για αυτό το λόγο χαρακτηρίζεται και ως στατιστικός ταξινομητής. Είναι σημαντικό να αναφέρουμε ότι είναι ένας από τους πιο διάσημους αλγορίθμους για Δέντρα Αποφάσεων, κατακτόντας τη πρώτη θέση μεταξύ των 10 κορυφαίων αλγορίθμων στην Εξόρυξη Δεδομένων [11].

Ο Αλγόριθμος C4.5 έκανε μια σειρά βελτιώσεων του Αλγορίθμου ID3, οι κυριότερες είναι οι παρακάτω,

a. Χειρισμός τόσο συνεχών όσο και διακριτών χαρακτηριστικών. Για το χειρισμό συνεχών χαρακτηριστικών, δημιουργεί ένα κατώφλι και στη συνέχεια χωρίζει τη λίστα σε εκείνα των οποίων η τιμή χαρακτηριστικού είναι πάνω από το όριο και εκείνες που είναι μικρότερες ή ίσες με αυτήν.

b. Χειρίζεται δεδομένα εκπαίδευσης με ελλειπές τιμές. Οι τιμές χαρακτηριστικών που λείπουν απλώς δεν χρησιμοποιούνται στους υπολογισμούς κέρδους και εντροπίας.
c. Χειρισμός χαρακτηριστικών με διαφορετικό κόστος.
d. Κλάδεμα δέντρων μετά τη δημιουργία του δέντρου, προσπαθώντας να αφαιρέσει κλαδιά που δεν βοηθούν αντικαθιστώντας τα με κόμβους φύλλων.

Λαμβάνοντας υπόψη τις παραπάνω βελτιώσεις, ο Αλγόριθμος πετυχένει έναν πολύ καλό συνδυασμό ποσοστού σφάλματος και ταχύτητας. Υπάρχουν πολλές ακόμη βελτιώσεις του Αλγορίθμου C4.5, όπως για παράδειγμα ο Αλγόριθμος EC4.5, ο οποίο υποστηρίζεται ότι βρίσκει το ίδιο Δέντρο Απόφασης με αυτό του Αλγοριθμου C4.5 με επίδοση εώς και πέντε φορές καλύτερο [12].

Είναι φανερό όπως και κάθε μέθοδος να διαθέτει κάποια Πλεονεκτήματα και Μειονεκτήματα, έτσι και στη περίπτωση των Δέντρων Απόφασης, είναι χρήσιμο να λάβουμε υπόψιν μας τα παρακάτω Πλεονεκτήματα και Μειονεκτήματα [13].

Πλεονεκτήματα Δέντρων:

a. Χρησιμοποιούνται τόσο για κατηγορικές όσο και για αριθμητικές μεταβλητές
b. Είναι εύκολα ερμηνεύσιμα λόγω της δομής τους
c. Ο εντοπισμός των ευαίσθητων σημείων των διαφόρων ενεργειών είναι εύκολος, πράγμα που διευκολύνει τον χρήστη να αποφύγει τυχόν κεκτημένα σφάλματα
d. Μπορούν να χειριστούν δεδομένα με ελλιπείς τιμές
e. Δέν απαιτούν υποθέσεις για το σύνολο δεδομένων, όπως ανεξαρτησία δεδομένων ή γραμμικότητα

Μειονεκτήματα Δέντρων:

a. Συχνά συναντάται το φαινόμενο της υπερπροσαρμογής (overfitting), κυρίως όταν τα δεδομένα του συνόλου εκπαίδευσης είναι λίγα με συνέπεια την αποτυχία κατηγοριοποίησης νέων δεδομένων

b. Αρκετές φορές χρειάζεται το στάδιο του κλαδέματος (pruning) για μείωση του μεγέθους τους
c. Μερικές φορές ο υπολογισμός τους μπορεί να είναι πολύ πιο πολύπλοκος σε σύγκριση με άλλους αλγόριθμους
d. Είναι ιδιαίτερα ευαίσθητα σε αλλαγές του συνόλου εκπαίδευσης, δηλαδή μικρές μεταβολές του συνόλου εκπάιδευσης μπορούν να προκαλέσουν τελειώς διαφορετική δομή του δέντρου

3.4.2 Ensemble Τεχνικές

Οι τεχνικές ensemble αναφέρονται σε μεθόδους που συνδυάζουν πολλαπλά μοντέλα μηχανικής μάθησης για να δημιουργήσουν έναν ισχυρότερο μοντέλο. Η ιδέα πίσω από αυτές τις τεχνικές είναι ότι η συνέργεια πολλών μοντέλων μπορεί να οδηγήσει σε καλύτερη προβλεπτική ικανότητα από ό, τι ένα μεμονωμένο μοντέλο.

Ένα από τα πιο δημοφιλή σύνολα τεχνικών ensemble είναι αυτές που βασίζονται στα δέντρα αποφάσεων. Τα δέντρα αποφάσεων είναι μοντέλα μηχανικής μάθησης που αποφασίζουν για την κατηγορία ενός παραδείγματος με βάση μια σειρά από συνθήκες και κανόνες που ορίζονται από τον αλγόριθμο εκπαίδευσης. Οι ensemble τεχνικές που βασίζονται στα δέντρα αποφάσεων αξιοποιούν την ιδέα της "σοφίας των πλήθους", δηλαδή της συλλογικής σοφίας πολλών μοντέλων.

Υπάρχουν διάφορες εφαρμογές ensemble τεχνικών στα δέντρα αποφάσεων, συμπεριλαμβανομένων των παρακάτω:

<u>Bagging:</u>

Ο αλγόριθμος bagging (bootstrap aggregating) λαμβάνει τυχαία δείγματα από το σύνολο εκπαίδευσης και εκπαιδεύει ένα δέντρο αποφάσεων για κάθε δείγμα. Τα δέντρα συνδέονται σε ένα σύνολο (ensemble) και οι τελικές προβλέψεις λαμβάνονται μέσω ψηφοφορίας ή μέσου όρου των προβλέψεων των δέντρων. Ο αλγόριθμος bagging βελτιώνει τη σταθερότητα και την ακρίβεια των προβλέψεων.

Random Forests:

Τα random forests είναι μια παραλλαγή του bagging που χρησιμοποιεί τυχαία επιλεγμένα χαρακτηριστικά σε κάθε κόμβο των δέντρων αποφάσεων. Αυτό το χαρακτηριστικό τυχαιότητας οδηγεί σε πιο πολυμεταβλητές προβλέψεις.

Boosting:

Η τεχνική της ενίσχυσης (boosting) εστιάζει στην εκπαίδευση δέντρων αποφάσεων που επικεντρώνονται στα παραδείγματα που δυσκολεύουν το μοντέλο. Τα δέντρα εκπαιδεύονται σειριακά, με κάθε νέο δέντρο να προσπαθεί να διορθώσει τα λάθη του προηγούμενου δέντρου. Με αυτόν τον τρόπο, η τεχνική της ενίσχυσης βελτιώνει συνεχώς την απόδοση του συνόλου.

Stacking:

Η τεχνική του stacking συνδυάζει πολλαπλά μοντέλα δέντρων αποφάσεων με ένα μετα-μοντέλο που μαθαίνει να συνδυάζει τις προβλέψεις των μοντέλων βάσης. Το μετα-μοντέλο μπορεί να είναι ένας απλός ταξινομητής, όπως ένας λογιστικός παλμός, που παίρνει τις προβλέψεις των μοντέλων βάσης ως είσοδο και εκπαιδεύεται να παράγει την τελική πρόβλεψη. Το stacking επιτρέπει την εκμετάλλευση των δυνατοτήτων διαφορετικών μοντέλων για βελτιωμένη ακρίβεια και απόδοση.

Οι ensemble τεχνικές που βασίζονται στα δέντρα αποφάσεων έχουν ευρεία εφαρμογή σε πολλούς τομείς, συμπεριλαμβανομένης της ταξινόμησης, της παλινδρόμησης και της ανίχνευσης ανωμαλιών ενώ με την κατάλληλη χρήση ensemble τεχνικών, μπορεί να επιτευχθεί αυξημένη ακρίβεια και αξιοπιστία στην πρόβλεψη των μοντέλων μηχανικής μάθησης.

3.4.3 Bagging

Το bagging είναι μία από τις τεχνικές συνεργατικής μάθησης (ensemble learning) που χρησιμοποιείται στη μηχανική μάθηση. Σκοπός του bagging είναι να βελτιώσει την απόδοση ενός μοντέλου πρόβλεψης δημιουργώντας πολλά υπο-μοντέλα που λειτουργούν ανεξάρτητα μεταξύ τους και συνδυάζοντας τις προβλέψεις τους.

Η μέθοδος Bootstrap Aggregating - Bagging είναι μια συλλογική μέθοδος, που βασίζεται σε μεθόδους επαναληπτικής δειγματοληψίας με αντικατάσταση από ένα σύνολο δεδομένων με χρήση διαφορετικού μοντέλου για το κάθε υποσύνολο δεδομένων [14]. Η δειγματοληψία ακολουθεί την ομοιόμορφη κατανομή πιθανότητας και κάθε δείγμα έχει ίσο μέγεθος δεδομένων. Να σημειώσουμε ότι είναι δυνατόν κάποια δεδομένα να εμφανισθούν περισσότερο από μία φορά, μιας και η δειγματοληψία γίνεται με επανατοποθέτηση [15]. Η μέθοδος Bagging προσπαθεί να μειώσει τη διακύμανση και το σφάλμα γενίκευσης, δημιουργώντας Τ διαφορετικούς «μαθητές» και εκπαιδεύοντάς τους σε ένα υποσύνολο του αρχικού συνόλου εκπαίδευσης. Έτσι, η τελική απόφαση προκύπτει από τον κανόνα της πλειοψηφίας ανάμεσα σε όλους τους σχηματιζόμενους εκτιμητές για κάθε σύνολο [16].

Έστω ότι το πλήθος των κατηγοριοποιητών, που προβλέπουν τη κατηγορία c_j για το Χ, δίνεται από τη σχέση:

$$u_j(X) = |\{L_i(X) = \ c_j \ | i = 1, \dots, T\}|$$

Ο συλλογικός κατηγοριοποιητής, συμβολίζεται με L_{Total} και προβλέπει την κατηγορία ενός σημείου δοκιμής Χ μέσω πλειοψηφικής ψηφοφορίας μεταξύ των k κατηγοριών:

$$L_{Total}(X) = argmax_{\ cj}\{u_j(X) \big| \ j = 1, \dots, k\}$$

Για καλύτερη κατανόηση παραθέτουμε τον ψευδοκώδικα της μεθόδου στο ακόλουθο αλγοριθμικό σχήμα:

B1. Καθορισμός των $T = \{L_1, \ldots, L_n\}$ διαφορετικών μαθητών

B2. Για $i = 1, \ldots, T$ επανέλαυε

B21. Δημιούργησε ενα υποσύνολο D_i ίσου μεγέθους με δειγματοληπτική μέθοδο επανατοποθέτησης από το γενικό σύνολο δεδομένων D

B22. Εκπαίδευσε τον αλγόριθμο L_i στο υποσύνολο D_i και θέσε Ac_i την ακρίβεια του αλγορίθμου

B3. Θέσε $L_{Total}(\underline{X_{test}}) = arg\max_{c_j} \sum_{i=1}^{T} I(c_j = Ac_i(\underline{X_{test}})$ ή
$L_{Total}(\underline{X_{test}}) = arg\max_{c_j} \frac{1}{T} \sum_{i=1}^{T} Ac_i(c_j \mid \underline{X_{test}})$

Σχήμα 12: Αλγοριθμικό σχήμα-Bagging

3.4.4 Boosting

Η μέθοδος Boosting είναι ακόμη μια συνεργατική μέθοδος που εκπαιδεύει τους βασικούς κατηγοριοποιητές με χρήση διαφορετικών δειγμάτων [17]. Ωστόσο, η βασική ιδέα της μεθόδου βρίσκεται στη προσεκτική επιλογή των δειγμάτων με τέτοιο τρόπο, ώστε να ενισχύεται η απόδοση της τελική μεθόδου για τα στιγμιότυπα που είναι δύσκολο να κατηγοριοποιηθούν [18]. Ξεκινάει με ένα κατηγοριοποιητή L_1, όπου εκπαιδεύεται σε ένα αρχικό δείγμα D_1 του συνόλου D και προσιορίζει το ποσοστό σφαλμάτων της εκπαίδευσής του. Στη συνέχεια, για την κατασκευή του D_2, επιλέγουμε τα εσφαλμένα κατηγοριοποιημένα στιγμιότυπα που έχουν υψηλότερη πιθανότητα και αφού εκπαιδευτεί ο κατηγοριοποιητής L_2, προσδιορίζεται το ποσοστό σφαλμάτων της εκπαίδευσής του. Έπειτα, κατά τη κατασκευή του D_3, μεγαλύτερη πιθανότητα να επιλεγούν έχουν εκείνα τα στιγμιότυπα που είναι δύσκολο να κατηγοριοποιηθούν από τον κατηγοριοποιητή L_1και L_2. Επαναλαμβάνεται αυτή η διαδικασία k φορές. Τέλος, ο συνεργατικός κατηγοριοποιητής προκύπτει μέσω σταθμισμένης ψηφοφορίας σε σχέση με την έξοδο των k κατηγοριοποιητών $L_1, \ldots, L_k$ [19]. Έστω ότι το πλήθος των κατηγοριοποιητών που προβλέπουν την κατηγορία c_j για το X δίνεται από τη σχέση:

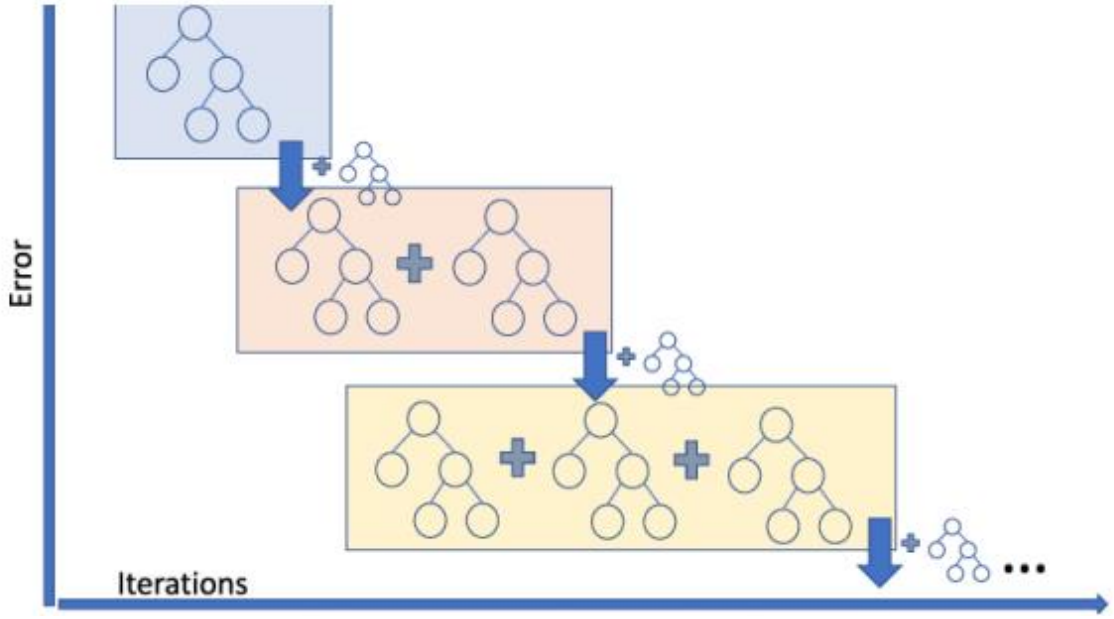
$$u_j(X) = |\{L_i(X) = c_j | i = 1, \dots, T\}|$$

Ο συνεργατικός κατηγοριοποιητής (L_{Total}) προβλέπει την κατηγορία ενός σημείου δοκιμής Χ μέσω της παρακάτω σχέσης:

$$L_{Total}(X) = argmax_{cj}\{w_j u_j(X) \big| j = 1, \dots, k\}$$

Σε αντίθεση με τη μέθοδο Bagging, που χρησιμοποιεί ανεξάρτητα τυχαία διανύσματα από το σύνολο D, η μέθοδος Boosting στηρίζεται σε σταθμισμένα δείγματα για την κατασκευή διαφορετικών συνόλων εκπαίδευσης, με το τρέχον δείγμα να εξαρτάται από το προηγούμενο. Το όφελος από τη μέθοδο Boosting μεγιστοποιείται όταν οι βασικοί κατηγοριοποιητές είναι ασθενείς, δηλαδή όταν έχουν ένα ποσοστό σφαλμάτων που είναι ελαφρώς μικρότερο από το ποσοστό ενός τυχαίου κατηγοριοποιητή. Η ιδέα πίσω από αυτή τη μέθοδο είναι η εξής: παρότι ο κατηγοριοποιητής L_1 είναι πιθανόν να μην έχει καλή απόδοση για όλα τα στιγμιότυπα εκπαίδευσής του, ο κατηγοριοποιητής L_2 ίσως βοηθήσει στη κατηγοριοποίηση αυτών των περιπτώσεων, που ο L_2 δεν μπόρεσε να ταξινομήσει σωστά. Αντίστοιχα, ο κατηγοριοποιητής L3 ίσως βοηθήσει στη κατηγοριοποίηση στιγμιοτύπων, όπου οι L_1, L_2 δεν μπόρεσαν να ανταπεξέλθουν ικανοποιητικά [20]. Λόγω του ότι μπορούν με πολλούς διαφορετικούς τρόπους να τεθούν τα βάρη των στιγμιοτύπων για τη δειγματοληψία, καθώς και για το τρόπο που συνδυάζονται οι μαθητές $L_1, L_2, \dots, L_k$. Μία από τις πιο δημοφιλείς παραλλαγές της μεθόδου Boosting είναι ο Adaptive Boosting - AdaBoosting, που οποίου την περιγραφή μπορεί να μελετήσει αναλυτικά ο αναγνώστης στο βιβλίο [21].

Στο ακόλουθο σχήμα παρουσιάζεται μια οπτική αναπαράσταση της μεθόδου.



Σχήμα 13: Boosting [22]

3.4.5 Random Forests:

Η μέθοδος Random Forests ανήκει, επίσης, στη κατηγορία συνεργατικών μεθόδων για προβλήματα Ταξινόμησης και Παλινδρόμησης, είτε για δύο ή περισσότερες κλάσεις [23]. Η βασική ιδέα της μεθόδου είναι να διαχωρίσουμε το σύνολο δεδομένων σε υποσύνολα και να κατασκευάσουμε Δέντρα Απόφασης για κάθε υποσύνολο. Για κάθε νέο δεδομένο, εξετάζεται σε όλα τα Δέντρα Απόφασης που έχουν δημιουργηθεί και η τελική απόφαση προκύπτει με βάση τον κανόνα της πλειοψηφίας. Η επιτυχία της μεθόδου Random Forests βασίζεται σε δύο διαδικασίες τυχαιοποίησης:

1. Εκπαίδευση κάθε δέντρου σε ένα διαφορετικό δείγμα περιπτώσεων
2. Σε κάθε βήμα δεν εξετάζονται όλες οι p μεταβλητές για το διαχωρισμό αλλά συνήθως, $\sqrt{p}$ μεταβλητές με τυχαία επιλογή και ύστερα, επιλέγεται ο καλύτερος διαχωρισμός μεταξύ αυτών

Να σημειώσουμε για τη δεύτερη διαδικασία τυχαιοποίησης ότι δημιουργεί διαφορετικά Δέντρα Απόφασης επιτρέποντας σε πολύ συσχετισμένες μεταβλητές να παίζουν σχεδόν ισοδύναμους ρόλους (διαφορετικά, η ελαφρώς πιο προγνωστική μεταβλητή

θα επιλέγεται πάντα «φαινόμενο Υπερπροσαρμογής»). Έτσι, αυτή η διαδικασία μειώνει το σφάλμα πρόβλεψης [24].

Η επιλογή μεταβλητών μπορεί να χρησιμοποιηθεί για να εκτιμήσουμε τη σχετική σημασία της j επεξηγηματικής μεταβλητής. Για να συμβεί αυτό, θυμηθείτε ότι ένα Random Forest ελέγχει κάθε παρατήρηση σε όλα τα Δέντρα Απόφασης, για τα οποία η παρατήρηση είναι εκτός συνόλου εκπαίδευσης και μετρά τον αριθμό των ψήφων για τη σωστή κλάση που λαμβάνει κάθε παρατήρηση. Αυτή η ψηφοφορία μέτρησης μπορεί να συγκριθεί με τον αντίστοιχο αριθμό ψήφων μετά από τυχαία μετατόπιση των τιμών της μεταβλητής j μεταξύ των δειγμάτων. Αν μετατρέψουμε τις τιμές του X_j και τις δοκιμάσουμε στα Δέντρα Απόφασης μετρώντας πάλι τον αριθμό των σωστών ψήφων, θα δούμε πως η διαφορά του αριθμού των σωστών ψήφων σύμφωνα με τις δύο διαδικασίες είναι η πρωταρχική σημασία για τη μεταβλητή j. Συνήθως, αυτά είναι τυποποιημένα, επομένως, το άθροισμα είναι ένα [25].

Να σημειώσουμε ότι η μέθοδος Random Forests για μεγάλο αριθμό Δέντρων Αποφάσεων δεν εμφανίζει το φαινόμενο της Υπερπροσαρμογής, παρόλο που το φαινόμενο αυτό εμφανίζεται έντονα στα Δέντρα Απόφασης. Αυτό δεν σημαίνει βέβαια πως και κάθε Δέντρο Απόφασης είναι ένας αξιόπιστος ταξινομητής. Ακόμη, ειναι σημαντικό να αναφερθεί ότι το σφάλμα γενίκευσης της μεθόδου Random Forests περιορίζεται από έναν όρο, που εκφράζει τη συσχέτιση μεταξύ των Δέντρων Απόφασης και μια άλλη ποσότητα που εκφράζει τη δυναμική των ταξινομητών. Η δυναμική ενός συνόλου ταξινομητών αναφέρεται στη μέση απόδοση των ταξινομητών. Όσο πιο μεγάλο είναι το περιθώριο, τόσο πιο πιθανό είναι ο ταξινομητής να προβλέψει σωστά την κατηγορία του χαρακτηριστικού Χ. Δηλαδή, όσο αυξάνεται η συσχέτιση μεταξύ των δένδρων ή αντίστοιχα μειώνεται η δυναμική του δάσους, τότε και το όριο του σφάλματος γενίκευσης τείνει να αυξηθεί. Οι διαδικασίες τυχαιοποίησης βοηθούν στη μείωση του βαθμού συσχέτισης μεταξύ των Δέντρων Απόφασης, και κατ' επέκταση στη βελτίωση του σφάλματος γενίκευσης ενός συνδυαστικού μοντέλου. Επομένως, με αυτόν τον τρόπο η μέθοδος δεν παρουσιάζει το φαινόμενο του overfitting κατά τη δημιουργία περισσότερων Δέντρων Αποφάσεων, αλλά αντίθετα το σφάλμα γενίκευσης περιορίζεται [26]. Για καλύτερη κατανόηση παραθέτουμε τον ψευδοκώδικα της μεθόδου [27] στο ακόλουθο σχήμα:

B1. Επέλεξε IDT: το Δέντρο Απόφασης, T: αριθμός επαναλήψεων, S: σύνολο δεδομένων, n: μέγεθος υποσυνόλου και N: αριθμός χαρακτηριστικών που χρησιμοποιούνται σε κάθε κόμβο

B2. Για κάθε $i = 1, \ldots, T$

B21. Επέλεξε υποσύνολο S_i μεγέθους n με επανατοποθέτηση

B22. Δημιούργησε ταξινομητή M_i χρησιμοποιώντας το $IDT(N)$ στο υποσύνολο S_i

B3. Εξαγωγή κάθε ταξινομητή $M_i \forall i = 1, \ldots, T$ και εξαγωγή τελικής απόφασης μέσω της μεθόδου πλειοψηφίας

Σχήμα 14: Αλγοριθμικό σχήμα-Random Forest

3.4.6 Light Gradient Boosting Machine

Ο αλγόριθμος Light Gradient Boosting Machine (LightGBM), αποτελεί επέκταση του αλγορίθμου Gradient Boosting που αναλύθηκε προηγουμένως και είναι ένας από τους πιο δημοφιλείς αλγορίθμους της Boosting οικογένειας στην περιοχή της μηχανικής μάθησης.

Ο LGBM είναι ένας αλγόριθμος με βάση τα δέντρα αποφάσεων και είναι γνωστός για την υψηλή του απόδοση και ταχύτητα εκτέλεσης. Αυτό το επιτυγχάνει μέσω της χρήσης τεχνικών όπως η (leaf-wise) ανάπτυξη του δέντρου και η εκμετάλλευση της παράλληλης επεξεργασίας.

Οι βασικές αρχές λειτουργίας του LGBM είναι οι εξής:

<u>Gradient Boosting:</u>

Ο LGBM χρησιμοποιεί την τεχνική του Gradient Boosting για την εκπαίδευση του μοντέλου. Σε κάθε επανάληψη, δημιουργεί ένα νέο δέντρο απόφασης που διορθώνει τα σφάλματα του προηγούμενου δέντρου, μειώνοντας σταδιακά το σφάλμα του μοντέλου.

Αρχικοποίηση:

OLGBM αρχικοποιεί το μοντέλο με ένα απλό δέντρο απόφασης και στη συνέχεια προσθέτει σταδιακά νέα δέντρα για τη βελτίωση των προβλέψεων.

Leaf-wise ανάπτυξη:

Αντί να χρησιμοποιεί την παραδοσιακή επίπεδη (level-wise) ανάπτυξη του δέντρου, ο LGBM υιοθετεί τη leaf-wise ανάπτυξη του δέντρου, κατα την οποία επιλέγει το φύλλο με τη μεγαλύτερη ελαχιστοποίηση στο σφάλμα σε κάθε επανάληψη. Αυτό οδηγεί σε μεγαλύτερη ακρίβεια και ταχύτερη εκπαίδευση του μοντέλου.

Παράλληλη επεξεργασία:

Ο LGBM εκμεταλλεύεται τη δυνατότητα παράλληλης επεξεργασίας για να επιταχύνει την εκπαίδευση του μοντέλου. Μπορεί να εκτελεστεί σε πολυνηματικό (multithread) περιβάλλον και να χρησιμοποιήσει πόρους υπολογιστικής παράλληλης επεξεργασίας για να επιτύχει υψηλή απόδοση.

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. 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.

3.5 Similarities

Όταν θέλουμε να παράξουμε προβλέψεις σχετικά με την πιθανή έκβαση ενός αποτελέσματος και έχουμε στη διάθεση μας ένα σύνολο δεδομένων για τα οποία γνωρίζουμε τις παρελθοντικές καταστάσεις και ετικέτες που τα συνοδεύουν, ένας άλλος τρόπος να το πραγματοποιήσουμε, είναι μέσω αλγοριθμικών τεχνικών που χρησιμοποιούν ορισμένους ευριστικούς μηχανισμούς.

Πιο συγκεκριμένα διακρίνουμε αλγορίθμους content-based filtering, οι οποίοι αναλύουν το περιεχόμενο που σχετίζεται με τα διάφορα προϊόντα-αντικείμενα ή αλλιώς μεταβλητές εισόδου και στους οποίους οι χρήστες και τα αντικείμενα αντιπροσωπεύονται από ένα σύνολο χαρακτηριστικών, και τους αλγορίθμους collaborative-filtering οι οποίοι δε χρειάζονται πληροφορία για το περιεχόμενο των items αλλά βασίζονται στην υπόθεση ότι ένας χρήστης τείνει να ενδιαφέρεται για προϊόντα-υπηρεσίες που προτιμούν παρόμοιοι με αυτόν χρήστες.

Στα πλαίσια της παρούσας διπλωματικής εργασίας κατα την οποία διαθέτουμε ένα χρηματοοικονομικό σύνολο δεδομένων απαρτιζόμενο από πληθώρα πελατών και της πληροφορίας χρήσης ή όχι του καθενός ενός προϊόντος ή υπηρεσίας μιας τράπεζας, έχουμε τη δυνατότητα να προβούμε σε πρόβλεψη μιας μελλοντικής κατάστασης μέσω της τεχνικής του συνεργατικού φιλτραρίσαμτος ή αλλιώς collaborative filtering και συγκεκριμένα του user based collaborative filtering που παρουσιάζεται στη συνέχεια.

3.5.1 Content based filtering

Αυτόι οι αλγόριθμοι προτείνουν items (προϊόντα-υπηρεσίες) με βάση το τι άρεσε στον χρήστη στο παρελθόν. Βασική ιδέα είναι πως ένας χρήστης είναι πιθανό να έχει την ίδια γνώμη για παρόμοια αντικείμενα. Ο τρόπος μέσω του οποίου επιτυγχάνονται τα πραπάνω, είναι η εκπαίδευση του συστήματος σχετικά με τις προτιμήσεις του χρήστη μέσω user feedback (Explicit ή Implicit) μέσω του οποίου χτίζεται ένα user profile. Το profile

περιλαμβάνει πληροφορία για τα items of interest του χρήστη, (π.χ. συγκεκριμένες ταινίες, βιβλία, CDs κτλ.) και βρίσκει similar items τα οποία και προτείνει στον χρήστη. Ουσιαστικά, το input είναι τα items of interest που βρίσκονται στο user profile και το output οι προβλέψεις δηλαδή τα προϊόντα που θα αγοράσει ή όχι ο χρήστης. Η όλη διαδικασία μπορεί να θεωρηθεί σαν μια αναζήτηση εγγράφων στις μηχανές αναζήτησης. Το user profile με τα items of interest είναι το query, ενώ τα items που περιέχονται στο site είναι το document base από το οποίο ψάχνουμε να βρούμε τα similar items που θα γίνουν recommend (πρόβλεψη χρήσης-αγοράς).

Πως αναγνωρίζεται όμως ένα item σαν similar item?

Α) Τα items περιγράφονται με βάση:

- Τα χαρακτηριστικά τους (π.χ. μια ταινία έχει α) είδος, β) σκηνοθέτη, γ) ηθοποιούς κτλ)
- Κάποια άλλα free tags που τα περιγράφουν
- Το κείμενο που περιγράφει το item

Β) Τα items που έχουν παρόμοια χαρακτηριστικά ή tags ή περιγράφονται με παρόμοιο κείμενο, θεωρούνται similar

- Δεν έχουν όλες οι λέξεις (χαρακτηριστικά, tags ή keywords του κειμένου) που περιγράφουν ένα item την ίδια βαρύτητα
- Πιο σημαντικές θεωρούνται οι λέξεις που εμφανίζονται πιο συχνά από τον μέσο όρο και που είναι πιο περιγραφικές για κάποια κλάση αντικειμένων (Π.χ. για περιγραφές restaurant, λέξεις όπως “noodle”, “shrimp”, “basil”, “exotic”, “salmon” είναι σημαντικές)

Για να συμπεράνει το σύστημα ποια items είναι similar χρησιμοποιούνται τεχνικές ανάλυσης περιεχομένου και machine learning (Π.χ Bayesian Classifiers, Tf–idf weight (term frequency–inverse document frequency)

Στα πλεονεκτήματα των συστημάτων content-based filtering εντάσσονται τα ακόλουθα:

- Μόνο τα rankings του χρήστη για τον οποίον προορίζεται το recommendation αρκούν
- Μπορούν να δώσουν recommendations σε χρήστες με προτιμήσεις που δεν είναι δημοφιλείς
- Μπορούν να προτείνουν νέα items καθώς και items που δεν είναι δημοφιλή
- Μπορούν να εξηγούν για ποιον λόγο πρότειναν κάποιο item, παρουσιάζοντας μια λίστα με τα χαρακτηριστικά του item που οδήγησαν στο recommendation

Από την άλλη, στα μειονεκτήματα ανήκουν:

- "User cold-start" problem (Πρέπει να μάθουν ποια χαρακτηριστικά του περιεχόμενου των items είναι σημαντικά για τον χρήστη και αυτό απαιτεί χρόνο)
- Το σύστημα εξακολουθεί να προτείνει items με βάση τις παλιές προτιμήσεις ακόμα και αν οι προτιμήσεις κάποιου χρήστη αλλάξουν.
- Δεν γίνεται εκμετάλλευση κριτικών για items που προέρχονται από άλλους χρήστες

3.5.2 User-based collaborative filtering

Βασική ιδέα των αλγορίθμων αυτής της κατηγορίας, είναι πως οι χρήστες που συμφώνησαν στο παρελθόν (σχετικά με προϊόντα-υπηρεσίες) τείνουν να συμφωνήσουν πάλι στο μέλλον. Οι αλγόριθμοι αυτοί προκειμένου να προβλέψουν το opinion ενός χρήστη (ενεργός χρήστης) για ένα item (για το οποίο δεν έχει φυσικά ακόμα εκφράσει άποψη) χρησιμοποιούν τα opinions των similar users. Το similarity ανάμεσα σε χρήστες βασίζεται στο κατά πόσο τα opinions που είχαν στο παρελθόν για άλλα items ήταν παρόμοια. Στηρίζονται συνεπώς στην έννοια της γειτονιάς του ενεργού χρήστη, στην οποία ανήκουν χρήστες με παρόμοιες βαθμολογίες και προτιμήσεις για έναν μεγάλο αριθμό items που έχουν καταναλωθεί από τον ενεργό χρήστη.

Για το σκοπό αυτό, ανακύπτει η ανάγκη υπολογισμού μιας ομοιότητας μεταξύ των χρηστών-πελατών που θα αποτελεί και τον ευριστικό μηχανισμό βάσει του οποίου θα πραγματοποιούνται και οι προβλέψεις.

Ορισμένες τεχνικές υπολογισμού του similarity μεταξύ των δειγμάτων είναι:

- Ευκλείδια απόσταση (Euclidean distance)- Euclidean similarity
- Pearson correlation
- Cosine similarity

Έχοντας φτιάξει ουσιαστικά έναν πίνακα συχετίσεων μεταξύ των χρηστών αξιοποιώντας τους μαθηματικούς τύπους των παραπάνω τεχνικών, είμαστε πλέον σε θέση να προβλέψουμε την πιθανότητα αγοράς-χρήσης μιας υπηρεσίας-προϊόντος μέσω της χρήσης των τιμών ομοιότητας μεταξύ των εξεταζόζεμων χρηστών καθώς και μιας aggregated συνάρτησης η οποία υποδυκνύει και την ζητούμενη πιθανότητα. Αν για παράδειγμα υποθέσουμε ότι έχουμε 5 χρήστες (δείγματα εισόδου) και με βάση τις κατηγορικές μεταβλητές που τους συνοδεύουν οδηγούμαστε στον ακόλουθο πίνακα ομοιότητας για τον χρήστη1,

	User 1	User 2	User 3	User 4	User 5
User 1	1	0,67	0,41	0,72	0,3

Πίνακας 2: Πίνακας ομοιότητας

μπορούμε να πάρουμε στη συνέχεια π.χ. τους 2 κοντινότερους σε αυτόν χρήστες και να προβλέψουμε πως η πιθανότητα να αγοράσει-βαθμολογήσει ένα προιόν δίνεται από μια συνάρτηση (aggregated function) σαν την ακόλουθη:

$$predictedRating_{U1} = \frac{sim(u1,u2) \cdot rating_{U2} + sim(u1,u4) \cdot rating_{U4}}{sim(u1,u2) + sim(u1,u4)}$$

Σχήμα 15: Πιθανή συνάρτηση για πρόβλεψη με την τεχνική collaborative filtering

Η όλη διαδικασία μοιάζει πολύ με τον γνωστό αλγόριθμο K-nearest-neighbour, όπου στην προκειμένη περίπτωση λάβαμε K=2. Ο τρόπος που θα εφαρμοστεί αυτή τη τεχνική στα πλαίσια της παρούσας εργασίας αναλύεται σε επόμενο κεφάλαιο.

Κλείνοντας, αξίζει εδώ να σημειώσουμε πως στα βασικά μειονεκτήματα αυτής της μεθόδου εντάσσονται τα παρακάτω:

- User Cold-Start problem (Οι νέοι χρήστες δεν έχουν κάνει πολλές αγορές, δυσκολεύοντας τον ορθό υπολογισμό των similarities)
- Σποραδικότητα (sparsity) των αγορών (Αν το σύνολο των items είναι μεγάλο, οι χρήστες μπορεί να έχουν μόνο λίγα items)
- Scalability (Με εκατομμύρια χρήστες και items οι υπολογισμοί μπορεί να γίνουν πολύ αργοί)

Κεφάλαιο 4: Πειραματική μεθοδολογία

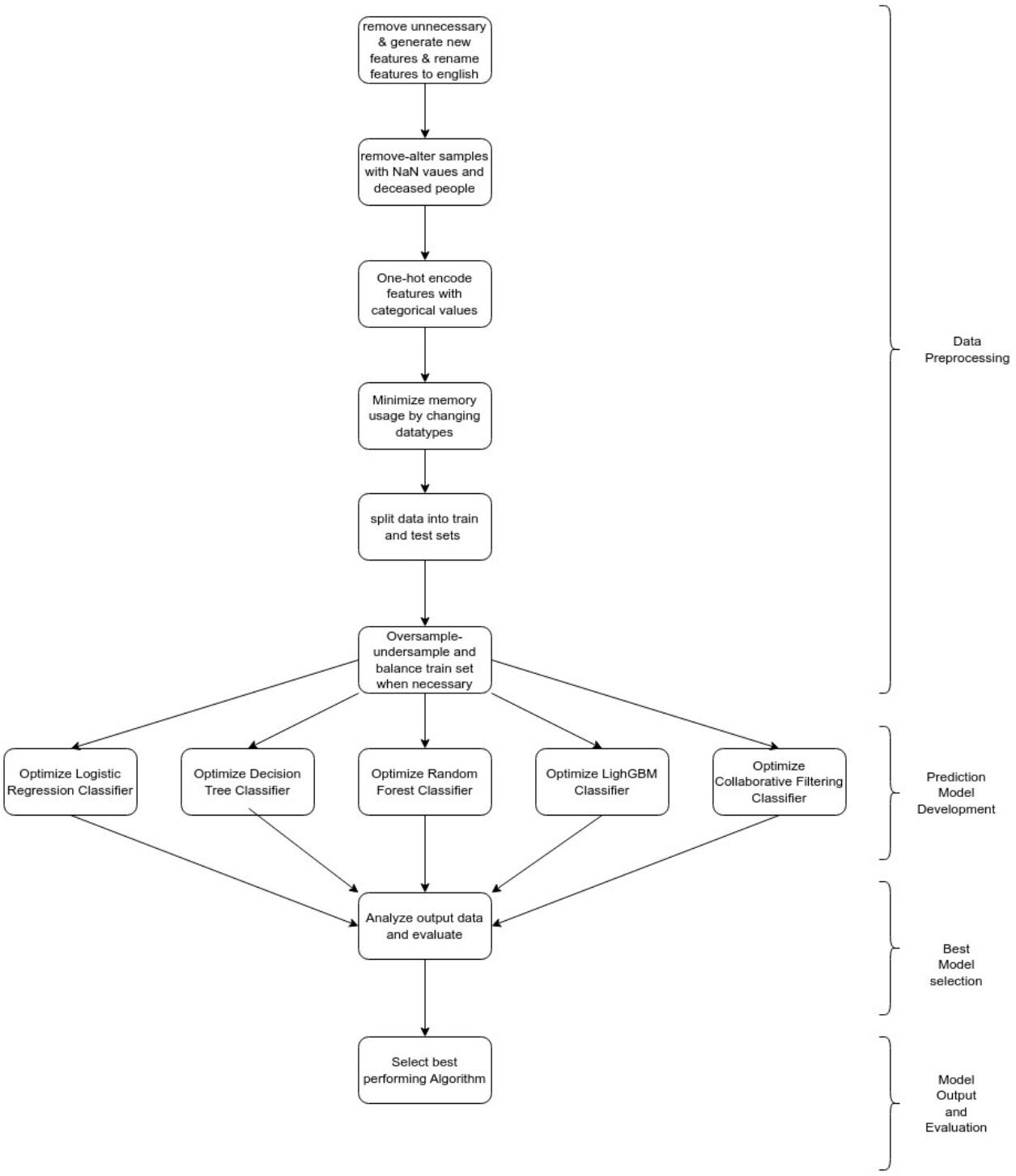
4.1 Γενική περιγραφή πειραματικού πλαισίου παραγωγής προβλέψεων

Η βασική ιδέα του προτεινόμενου πλαισίου είναι η αυτοματοποίηση της διαδικασίας παραγωγής προβλέψεων από σύνολα δεδομένων που αφορούν τη χρήση-απόκτηση προϊόντων ή υπηρεσιών από πελάτες της Ισπανικού πολυεθνικού ομίλου επιχειρήσεων Santander, και συγκεκριμένα από την τράπεζα Santander Bank. Το πλαίσιο αυτό, λειτουργεί ως ένα black box που δέχεται ως είσοδο ένα σύνολο δεδομένων με 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 την απόκτηση-χρήση ή μη προιόντων και υπηρεσιών, των πελατών μιας τράπεζας, ο ορίζοντας πρόβλεψης είναι ο επόμενος μήνας, και τα δεδομένα εισόδου είναι είτε του προηγούμενου μήνα είτε όλα τα δεδομένα που έχουν συγκεντρωθεί σε βάθος μηνών. Οι προβλέψεις αυτές αποτελούν ένα καλό και σημαντικό πρώτο βήμα για την ανάλυση της χρήσης των προϊόντων και υπηρεσιών της τράπεζας, με σκοπό την δημιουργία ενός ολοκληρωμένου συστήματος προτάσεων. 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 στην τράπεζα σημαντικές πληροφορίες για την κατάσταση χρήσης του συνόλου δεδομένων που χρησιμοποιείται προς μελέτη.

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/forehandy/Product_ Recommendation και περιλαμβάνει τα εξής διακριτά βήματα, τα οποία θα αναλυθούν στη συνέχεια:

- Ανάγνωση και 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 δοκιμή των τιμών των υπερπαραμέτρων τους, για την παραγωγή προβλέψεων.
- Εξαγωγή αποτελεσμάτων και αξιολόγηση μοντέλων. Στο προτελευταίο στάδιο αξιολογείται το μοντέλο που προέκυψε με χρήση μετρικών. Σε αυτό το βήμα δοκιμάζονται οι συνδυασμοί των υπερπαραμέτρων που συνθέτουν κάθε μοντέλο, παίρνοντας το μέσο όρο των κατατάξεων τους σύμφωνα με ορισμένες μετρικές και ανακυρήσσεται ως βέλτιστο το μοντέλο με τα καλύτερα αποτελέσματα.
- Επιλογή βέλτιστου μοντέλου από το σύνολο της πειραματικής διαδικασίας

Παρακάτω παρουσιάζεται η σχηματική αναπάρασταση των βημάτων της προτεινόμενης μεθοδολογίας:



Σχήμα 16: Σχηματική αναπαράσταση πειραματικής 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 από το μέσο όρο ή με πρόβλεψη βάσει των υπόλοιπων χαρακτηριστικών του συγκεκριμένου παραδείγματος ή ακόμα και να εξαιρεθεί τελείως το συγκεκριμένο δείγμα αν πρόκειται για δείγμα εκπαίδευσης. Σε άλλες περιπτώσεις είναι καλύτερο να παραμείνουν ως ελλειπούσες, ίσως με την προσθήκη νέου χαρακτηριστικού και χρήση one-hot encoding. Επίσης, πολλές φορές είναι χρήσιμη η δημιουργία νέων χαρακτηριστικών από τα υπάρχοντα, τα οποία δίνουν πιο σημαντικές και ουσιαστικές πληροφορίες. Ένα παράδειγμα τέτοιου χαρακτηριστικού είναι ο τύπος ενός πελάτη όταν δίνεται από μικρού εύρους συνεχείς αριθμητικές τιμές. Σε αυτή την περίπτωση είναι προτιμότερο να παραχθούν στήλες αντίστοιχου πλήθους με τις δυνατές τιμές του αρχικού χαρακτηριστικού, οι οποίες θα λαμβάνουν την τιμή 1 και 0 ανάλογα με το σε ποιά κατηγορία ανήκει ο κάθε πελάτης. 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 Μετονομασία, αφαίρεση και δημιουργία χαρακτηριστικών

Το αρχικό βήμα της προεπεξεργασίας είναι η κατανόηση των δεδομένων. Προκειμένου να επιτευχθεί αυτό, μετονομάζουμε τα χαρακτηριστικά όλων των δειγμάτων σύμφωνα με μια περιγραφική και ουσιώδη ονομασία που θα είναι κατανοητή τόσο σε εμάς

όσο και σε κάθε χρήστη του συγκεκριμένου συνόλου δεδομένων. Ύστερα, ανάλογα με το ποιά εκ των χαρακτηριστικών προσδίδουν αξιόλογη πληροφορία που θα αποδειχθεί χρήσιμη για τις προβλέψεις της μεθόδου, κρίνουμε αν για λόγους πληρότητας ή επίδοσης και εκπαίδευσης των μοντέλων, είναι προτιμότερο να αφαιρεθούν ή τροποποιηθούν ή και να παραμείνουν. Τελικά αφού έχουμε ολοκληρώσει τα δύο προηγούμενα βήματα και έχουμε κατανοήσει σε βάθος την πληροφορία που προσφέρεται από τα δείγματα του συνόλου δεδομένων μας, μπορούμε να προβούμε σε δημιουργία νέων χαρακτηριστικών, τα οποία κρίνουμε ότι θα συμβάλλουν θετικά στην παραγωγή προβλέψεων.

4.2.2 Αφαίρεση-τροποποίηση κενών τιμών και «αδιάφορων» δειγμάτων

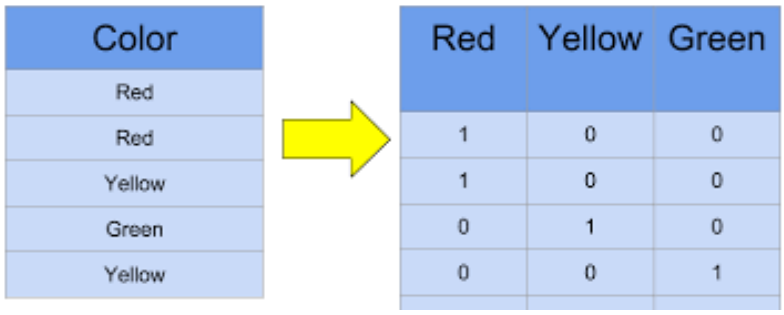
Το δέυτερο στάδιο της προεπεξεργασίας, είναι η τροποποίηση ή και αφαίρεση όλων εκείνων των δειγμάτων για τα οποία κάποιο εκ των χρακτηριστικών παρουσιάζει κενή τιμή. Συγκεκριμένα, ανάλογα με το αν το εν λόγω feature του υπό εξέταση δέιγματος δύναται να συμπληρωθέι μέσω κάποιου γενικού κανόνα όπως Μ.Ο., μηδενική τιμή (0) κ.λ.π., κρίνουμε την αφαίρεση ή την τροποποίηση του. Αξίζει εδώ να σημειωθεί πως είθισται να λαμβάνουμε γενικές αποφάσεις σε σχέση με τον τρόπο που θα χειριστούμε τέτοιου είδους «ελλείψεις» μιας και δεν είναι συνοχικά σωστό να κρίνουμε κάθε δείγμα κατά περίπτωση. Αυτό σημάινει πως αν για παράδειγμα παρατηρήσουμε πως το περιεχόμενο της στήλης (feature) income είναι κενό (Null) για ένα δείγμα τότε θα πρέπει αν επιλέξουμε να το μετατρέψουμε σε 0 για το συγκεκριμένο δείγμα, να πράξουμε αναλόγως και για όλα τα υπόλοιπα δείγματα που παρουσιάζουν την ίδια συμπεριφορά.

Επιπλέον των παραπάνω, αφαιρούμε απο το σύνολο των δεδομένων μας κάθε εγγραφή της οποίας η μελέτη και ανάλυση δεν αξίζει να μελετηθεί ή δεν συμβάλλει στην απόκτηση γνώσης για την εξαγωγή ενός συμπεράσματος. Για παράδειγμα, στην εν λόγω εφαρμογή, κρίνουμε αναγκαίο να αφαιρεθούν όλοι οι πελάτες που έχουν πάυσει να είναι εν ζωή και συνεπώς η μελλοντική κατάσταση τους σχετικά με την απόκτηση-αγορά ή χρήση

ενός πρόϊόντος ή υπηρεσίας της τράπεζας είναι τόσο τετριμμένη όσο και άνευ σημασίας για τη μελέτη μας.

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 17: Παράδειγμα one-hot-encoding [28]

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 Ελαχιστοποίηση χρήσης 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 αλλά και για να τηρούνται όλα τα απαραίτητα κριτήρια συμβατότητας αναφορικά με τη μορφή των δεδομένων και το υπό εκπαίδευση μοντέλο. Για παράδειγμα για τον logistic regression classifier είναι απαράιτητο όλες οι τιμές των χαρακτηριστικών να είναι numeric.

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.

4.2.5 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 ακολουθεί τη λογική ότι δύο διαδοχικοί μήνες αποτελούν τα train και test sets αντίστοιχα κατά αύξουσα χρονολογική σειρά. Έτσι αν για παράδειγμα ως train set ορίσουμε το σύνολο των δειγμάτων κατά το μήνα Ιανουάριο τότε ως test set ορίζεται ο μήνας Φεβρουάριος. Αξίζει εδώ να σημειωθεί πως εξετάζονται μόνο τα δείγματα-πελάτες για τα οποίους έχουμε παρουσία και τους δύο διαδοχικούς μήνες της μελέτης-πρόβλεψης.

4.2.6. Ισορρόπηση train set

Επόμενη προεργασία που απαιτείται πριν την εκπαίδευση των μοντέλων πρόβλεψης, είναι η ισορρόπηση (balance) 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 πελάτες που διαθέτουν το υπό εξέταση προϊόν ή υπηρεσία έναντι των υπολοίπων που δεν το έχουν, 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 δυνατή επίδοση, 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 έδειξε η μέθοδος 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, στην εν λόγω εργασία,

καταλήξαμε πως το τυχαίο under sampling είναι προτιμότερο λόγω του πολύ μεγάλου όγκου δεδομένων και των πολύ μεγάλων χρόνων εκπαίδευσης των μοντέλων.

4.3 Ανάπτυξη μοντέλων πρόβλεψης

Έχοντας ολοκληρώσει το σύνολο της πειραματικής διαδικασίας, αναδυκνείεται πως οι ταξινομητές με την καλύτερη επίδοση σε προβλήματα αναφορικά με την πρόβλεψη αγοράς ενός προϊόντος ή υπηρεσίας της τράπεζας είναι οι: Decision Tree, Random Forest, και LightGBM. Πρόκειται για μοντέλα τα οποία με εκκίνηση αυτό του Decision Tree, παρουσιάζουν κλιμακόυμενη πολυπλοκότητα, με το Random Forest και το LightGBM να ακολουθούν αντίστοιχα. Και τα τρία αυτά μοντέλα βασίζονται στη θεωρία των δέντρων απόφασης, γεγονός που τα καθιστά κατάλληλα για την πρόβλεψη της συγκεκριμένης χρήσης. 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 μεγιστοποιούν τη συνάρτηση που έχουμε θέσει ως στόχο. Στην περίπτωσή μας, χρησιμοποιούμε τη μετρικές AUC, Recall, Precision και Accuracy, οι οποίες αναλύθηκαν προηγουμένως στο κεφάλαιο 2.5.

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 αρκετά 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 ο αλγόριθμος του grid search ούτως ώστε να έχουμε όσο το δυνατόν πιο ακριβή αποτελέσματα.

Για το συμπερασμό του συνδυασμού των βέλτιστων υπερπαραμέτρων για κάθε μοντέλο, δημιουργήθηκε μία κατάταξη της τιμής του μέσου όρου (Μ.Ο.) των κατατάξεων των μετρικών AUC, precision και recall που παράγει ο κάθε ένας συνδυασμός παραμέτρων από αυτούς που ακολοθούν στους επόμενους πίνακες. Η ενέργεια αυτή πραγματοποιηθήκε τόσο σε επίπεδο αλγορίθμου και ύστερα προϊόντος, όσο και σε επίπεδο προϊόντος και ύστερα αλγορίθμου(μοντέλου), μιας και στη συγκεκριμένη εργασία το κάθε πείραμα έτρεχε αυτόνομα για κάθε προϊόν ή υπηρεσία.

Παρακάτω αναγράφονται οι χώροι αναζήτησης, δηλαδή το σύνολο των τιμών που λαμβάνει η κάθε υπερπαράμετρος για το κάθε μοντέλο που αναπτύχθηκε. Λαμβάνοντας υπόψιν το σύνολο όλων των εξαγόμενων συνδυασμών ανά μοντέλο δύναται και η ευκαιρία λεπτομερούς μελέτης της επίδρασης κάθε υπερπαραμέτρου.

Log. Regression Classifier	
Υπερπαράμετρος	Χώρος αναζήτησης
max_iter	{50, 100, 500, 1000}

Πίνακας 3: Χώρος αναζήτησης υπερπαραμέτρων Logistic Regression Classifier

Decision Tree Classifier	
Υπερπαράμετρος	Χώρος αναζήτησης
criterion	{‘gini’, ‘entropy’}
max_depth	{None, 4, 8, 16}
max_features	{0.6, 0.75, 0.9, 1.0}
min_samples_leaf	{1, 2, 4, 10, 20}
min_samples_split	{2, 4, 10, 20, 40}

Πίνακας 4: Χώρος αναζήτησης υπερπαραμέτρων Decision Tree Classifier

Random Forest Classifier	
Υπερπαράμετρος	Χώρος αναζήτησης
criterion	{‘gini’, ‘entropy’}
max_features	{0.6, 0.75, 0.9, 1.0}
n_estimators	{100, 250, 500, 1000}

Πίνακας 5: Χώρος αναζήτησης υπερπαραμέτρων Random Forest Classifier

LightGBM Classifier	
Υπερπαράμετρος	Χώρος αναζήτησης
colsample_bytree	{0.8, 0.9, 1.0}
learning_rate	{0.005, 0.01, 0.05, 0.1, 0.3}
max_depth	{-1, 4, 8, 16}
n_estimators	{100, 250, 500, 1000}

Πίνακας 6: Χώρος αναζήτησης υπερπαραμέτρων LightGBM Classifier

Επιπλέον, σημειώνουμε πως για την εξαγωγή συμπερασμάτων σχετικά με τις μεθόδους του Collaborative filtering δεν επεκταθήκαμε πέραν της παραγωγής προβλέψεων σύμφωνα με τον κανόνα της πλειοψηφίας, κατά τον οποίο για την πρόβλεψη αγοράς ενός

προϊόντος ή υπηρεσίας από τον υπό εξέταση χρήστη, προβλέψαμε αγορά ή μη σύμφωνα με το αν η πλειοψηφία των 100 πιο όμοιων με αυτόν χρήστες έχουν το ίδιο προϊόν ή όχι.

4.4 Εξαγωγή αποτελεσμάτων και αξιολόγηση μοντέλων

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, accuracy, και AUC, με τη μεγαλύτερη βαρύτητα να δίνεται στο AUC καθώς είναι και αυτό που δείχνει την πληρέστερη εικόνα για το συγκεκριμένο πεδίο εφαρμογής.

Σημειώνουμε εδώ πως η αξιολόγηση συνίσταται σε δύο διαφορετικά σενάρια πρόβλεψης. Το πρώτο και καθολικό είναι αυτό κατα το οποίο κάθε μοντέλο με τις υπερπαραμέτρους του καλείται να προβλέψει για κάθε προϊόν που εξετάζουμε το άν και κατα πόσο οι πελάτες θα συνεχίσουν να έχουν, θα συνεχίσουν να μην έχουν, θα πάψουν να χρησιμοποιούν ή θα αποκτήσουν ένα προϊόν ή υπηρεσία. Επιχειρώντας να εκφράσουμε αυτό το transition με όρους πρόβλεψης έχουμε τις ακόλουθες τέσσερεις περιπτώσεις 1→1, 0→0, 1→0 και 0→1 αντίστοιχα. Το δεύτερο σενάριο είναι αυτό όπου αξιολογούμε τα μοντέλα μόνο σε επίπεδο απόκτησης προϊόντος ή υπηρεσίας από έναν πελάτη ο οποίος δεν είχε το συγκεκριμένο προϊόν ή υπηρεσία. Περιοριζόμαστε λοιπόν μόνο στις μεταβάσεις από 0 σε 1 (0→1). Το δεύτερο αυτό σενάριο είναι πιο κοντά και στη λογική ενός recommendation system το οποίο για να αξιολογηθεί με γνώμονα οικονομικού κέρδους για την τράπεζα, αρκεί κυρίως να εξετάσουμε ποιούς πελάτες πρέπει να στοχεύσουμε ώστε να προτείνουμε το υπό πρόβλεψη προϊόν.

4.5 Επιλογή βέλτιστων μοντέλων

Ο τρόπος σύμφωνα με τον οποίο καταλήγουμε στο ποιά μοντέλα είναι τα βέλτιστα είναι να εξετάσουμε τις τιμές που παρουσιάζουν για κάθε προϊόν και αλγόριθμο συγκεντρωτικά και στα ίδια σύνολα δεδομένων. Πιο συγκεκριμένα, προκειμένου να αποτιμήσουμε και να κατατάξουμε τα υπό εξέταση μοντέλα, αφού δοκιμάσαμε κάθε πιθανό συνδυασμό υπερπαραμέτρων κατατάξαμε κατά φθίνουσα σειρά κάθε συνδυασμό υπερπαραμέτρων κάθε μοντέλου, για κάθε μια μετρική εκ των AUC, Precision και Recall. Στη συνέχεια θεωρήσαμε ως συνολικό Ranking κάθε μοντέλου την τιμή που προκύπτει ως ο μέσος όρος των 3 αυτών rankings.

Για παράδειγμα αν εξετάσουμε το μοντέλο Decision Tree η διαδικασία έχει ώς εξής:

1) Δεδομένων των τιμών των υπερπαραμέτρων όπως παρουσιάστηκαν στον πίνακα 3, διαπιστώνουμε πως ο χώρος αναζήτησης διαθέτει:

   (#criterion) x (#max_depth) x (#max_features) x (#min_samples_leaf) x (#min_samples_split)

   πιθανούς συνδυασμούς προς εξέταση, όπου με **#hyperparameter** συμβολίσαμε το πλήθος τιμών που λαμβάνει μία υπερπαράμετρος. Αναλυτικά το προκύπτον νόυμερο σύμφωνα με τον πίνακα είναι **2 x 4 x 4 x 5 x 5 = 800** συνδυασμοί.

2) Κάθε ένας από τους προκύπτοντες συνδυασμούς αποκτά ένα rank ως προς κάθε μία απο τις μετρικές AUC, Recall, Precision σύμφωνα με την τιμή της εκάστοτε μετρικής (AUC Rank 1 έχει ο συνδυασμός με τη μεγαλύτερη τιμή AUC, AUC Rank 2 ο συνδυασμός με τη δέυτερη μεγαλύτερη τιμή κ.ο.κ.. Αντίστοιχα και για τις άλλες 2 μετρικές.)

3) Δημιουργούμε νέο rank που το ονομάζουμε AUC-PRECISION-RECALL Rank ως:

$$\textbf{AUC-PRECISION-RECALL Rank} = \frac{\boldsymbol{AUC\ RANK + PRECISION\ RANK + RECALL\ RANK}}{\mathbf{3}}$$

4) Κατατάσσουμε τους συνδυασμούς κατά φθίνουσα σειρά αυτής της μετρικής
5) Ο συνδυασμός με το μικρότερο **AUC-PRECISION-RECALL Rank** είναι και ο πιό αποδοτικός για την πειραματικής μας διενέργεια.

Σημειώνεται εδώ πως αυτή η διαδικασία πραγματοποιήθηκε για κάθε ένα εκ των τεσσάρων προϊόντων που αποτέλεσαν την πειραματική μας μελέτη και για κάθε ένα από τα μοντέλα που αναπτύχθηκαν. Ύστερα, επιλέγοντας τον βέλτιστο συνδυασμό υπερπαραμέτρων καταγράψαμε τις τιμές των AUC, Precision, Recall και Accuracy και διαπιστώσαμε ποιά απο τα μοντέλα συμπεριφέρονται αποδοτικότερα στην πρόβλεψη αγοράς προϊόντων απο την τράπεζα. Παράλληλα παρήξαμε συμπεράσματα και συγκεντρωτικά για την μεμονομένη μεταβολή κάθε μιας μεμονωμένα υπερπαραμέτρου για κάθε προϊόν και μοντέλο.

Κεφάλαιο 5: Πειραματική διαδικασία και αποτελέσματα

5.1 Περιγραφή πειραματικού συνόλου δεδομένων

Για τη μελέτη της συμπεριφοράς των καταναλωτών και τις προθέσεις τους σχετικά με την πρόθεση χρήσης ή όχι προϊόντων και υπηρεσιών θα μπορούσαμε να αντλήσουμε δεδομένα από πολλές διαφορετκές πηγές καθώς ή πρόβλεψη τόσο της ζήτησης όσο και η επιθυμία να προοικονομηθούν οι ενέργειες και ανάγκες των καταναλωτών είναι ζήτημα ύψιστης σημασίας για κάθε επιχείρηση ή οργανισμό του οποίου σκοπός είναι η απόκτηση κέρδους. Στα πλαίσια της συγκεκριμένης πειραματικής μελέτης και διαδικασίας τα δεδομένα επιλέχθηκαν από τον χρηματοοικονομικό τομέα.

Συγκεκριμένα χρησιμοποιήθηκαν δεδομένα του ομίλου επιχειρήσεων Santander Group ο οποίος αποτελεί μια Ισπανική πολυεθνική εταιρεία οικονομικών υπηρεσίων. Όλα τα δεδομένα προήλθαν από τον τραπεζικό τομέα του ομίλου και την Santander Bank, που περιλαμβάνει τα δεδομένα των χρηστών για τις ποικίλες υπηρεσίες που προσφέρει η ομώνυμη τράπεζα. Τα δεδομένα αυτά δημοσιεύθηκαν από την εταιρεία με τη μορφή διαγωνισμού στην πλατφόρμα Kaggle της Google, στον οποίο οι συμμετέχοντες καλούνταν να προβλέψουν ποιά θα είναι η συμπεριφορά των υπό εξέταση πελατών εντός ενός μελλοντικού ορίζοντα πρόβλεψης ενός μήνα, με σκοπό την δημιουργία ενός συστήματος προτάσεων το οποίο θα μπορεί με μεγάλη ακρίβεια να προβλέπει ποιά προϊόντα είναι πιθανότερο να αγοράσει κάθε πελάτης. Η εφαρμογή αυτή μπορέι να χρησιμοποιηθεί με πολλούς τρόπους από την Santander Bank και να της επιφέρει μεγάλα κέρδη σε πολλά επίπεδα.

Ο διαγωνισμός μπορεί να βρεθεί στον ακόλουθο σύνδεσμο:

https://www.kaggle.com/competitions/santander-product-recommendation/overview

Το σύνολο δεδομένων αποτελείται από τα δεδομένα που θα χρησιμοποιηθούν τόσο για την εκπαίδευση όσο και την αξιολόγηση των μοντέλων που θα αναπτυχθούν. Τα δεδομένα αποτελούνται στο σύνολο τους από 13.647.310 δείγματα τα οποία αντιπροσωπεύουν την κατάσταση των πελατών σε βάθος 18 μηνών. Από τα 13.647.310 δείγματα-γραμμές, για την πειραματική μας διαδικασία αφαιρέσαμε όλα αυτά για τα οποία κάποια εκ των χαρακτηριστικών συμπεριελάμβαναν προβληματικές τιμές και δεν ήταν παρόντα και στους 18 μήνες ανάλυσης, ενώ αφαιρέσαμε και τους αποθανόντες πελάτες για τους οποίους μια πρόβλεψη είναι άνευ νοήματος, καταλήγοντας τελικώς στο σύνολο των 6.137.217 δειγμάτων. Για τη μελέτη της πρόβλεψης της συμπεριφοράς των καταναλωτών, περιοριστήκαμε σε δέιγματα που αφορούσαν μόνο τον μήνα Ιανουάριο του έτους 2015 και είχαν εξαχθεί την 28η ημέρα. Αυτό το σύνολο αποτέλεσε το train set ενώ το test set συνίσταται από τους ίδιους ακριβώς πελάτες και την κατάσταση τους κατά το μήνα Φεβρουάριο και συγκεκριμένα την ημερομηνία 28/02/2015. Η ανάλυση μας και ο πειραματικός σχεδιασμός συντελέστηκε για κάθε μοντέλο μόνο για 4 προϊόντα απο σύνολο 24 για λόγους ταχύτητας και ευκολίας. Τα προϊόντα που επιλέχθηκαν φροντίσαμε να ανήκουν κάθε ένα σε μία κατηγορία απο τις κύριες κατηγοριες στις οποίες διαχωρίζονται τα προϊόντα και οι υπηρεσίες τον τραπεζών, όλα είχαν ικανοποιητικό και αρκετά ισορροπημένο πλήθος χρηστών και μη, και είναι τα ακόλουθα:

- **Credit card (πιστωτική κάρτα)**
- **Current account (τρεχούμενος λογαριασμός)**
- **Funds (κεφάλαια)**
- **Long-term deposits (μακροπρόθεσμες καταθέσεις)**

Σημειώνουμε πως οι περισσότεροι πελάτες προέρχονται από την χώρα στην οποία εδράζει η συγκεκριμένη τράπεζα η οποία είναι η Ισπανία (ES) και πως η κατανομή των πελατών στους διάφορους νομούς και πόλεις της Ισπανίας είναι ομοιόμορφη όπως προέκυψε απο την σχετική στατιστική ανάλυση στο dataset.

Στην επόμενη σελίδα, παρουσίαζονται αναλυτικά ορισμένα από τα χαρακτηριστικά του συνόλου δεδομένων που χρησιμοποιήσαμε. Για το πλήρες σύνολο των χαρακτηριστικών με την περιγραφή τους, προτρέπεται ο αναγνώστης να οδηγηθεί στο σύνδεσμο του διαγωνισμό στην ενότητα data (δεδομένα) όπως δόθηκε παραπάνω.

Feature (Χαρακτηριστικό)	Περιγραφή
'fetch_date',	Ημερομηνία εξαγωγής δεδομένων, ο πίνακας χωρίζεται σε τμήματα βάσει αυτού
'cust_code',	μοναδικός κωδικός πελάτη (ID)
'emp_index',	δείκτης εργαζομένου
'country',	χώρα
'sex',	φύλο
'age',	ηλικία
spouse_index',	δείκτης για το αν οπελάτης είναι σύζυγος υπαλλήλου
'deceased',	δείκτης για το αν ο πελάτης βρίσκεται εν ζωή
prov_code',	διεύθυνση πελάτη
'prov_name',	όνομα επαρχιακής διεύθυνσης
'cust_date',	ημερομηνία που έγινε πελάτης
'new_cust',	δέικτης αν είναι νέος πελάτης
'joining_channel',	κανάλι μέσω του οποίου έγινε πελάτης της τράπεζας
'income',	εισόδημα
.	.
.	.
.	.

Πίνακας 7: Ορισμένα εκ των χαρακτηριστικών του συνόλου δεδομένων (Τίτλος-Περιγραφή)

5.2 Προεπεξεργασία συνόλου δεδομένων

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. Στο σύνολο δεδομένων που επιλέχθηκε έγινε η κατάλληλη τροποποίηση για την εξαγωγή εγκυρότερων αποτελεσμάτων στα ακόλουθα σημεία:

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 “emp_index”, “cust_type”, “ cust_rel” και “segmentation”. Αναφορικά με το “segmentation” πριν πραγματοποιηθεί one-hot-encoding, ελέγξαμε αν υπάρχει correlation (συσχέτιση) μεταξύ αυτού και του χαρακτηριστικού “income”, μιας και αν υπήρχε συσχέτιση δε θα έπρεπε να προβούμε σε τροποποίηση. Σε αυτή την περίπτωση θα το κρατούσαμε με αριθμητικές τιμές κατά αύξοντες αριθμούς όπως αρχικά εμφανιζόταν στο σύνολο δεδομένων ενώ σε αντίθετη περίπτωση όχι. Ο λόγος πίσω απο αυτή την ενέργεια είναι ότι δε θέλουμε να μας επηρεάζει τις προβλέψεις μια στήλη (χαρακτηριστικό) για την οποία η διαφοροποίηση των τιμών τις σε 1, 2 ή 3 δεν έχει διαβαθμιστική αξία. Για όσους δεν προέβησαν σε αναλυτική μελέτη του συνόλου δεδομένων όπως δόθηκε στον εξωτερικό σύνδεσμο του διαγωνισμού, αναφέρουμε πως το εν λόγω χαρακτηριστικό εκφράζει μια κοινωνική βαθμίδα από VIP πελάτες έως απόφοιτους κολλεγίων κ.λ.π. συνεπώς μια συσχέτιση με το εισόδημα ήταν αρκετά πιθανή. Το χαρακτηριστικό “income” τροποποιήθηκε έτσι ώστε όλες οι κενές τιμές να τεθούν ίσες με το μηδέν (0). Επιπλέον, το “country” μετατράπηκε, αρχικά, από string σε μορφή integer για να μπορεί να διαχειριστεί με μεγαλύτερη ευκολία καθώς η αρχική μορφή μπορεί να είναι εύκολα αντιληπτή από τον άνθρωπο αλλά όχι από ορισμένα μοντέλα μηχανικής μάθησης. Αντίστοιχη δουλειά έγινε και για τα χαρακτηριστικά: “sex”, “indrel”, “residence_index”, “foreigner_index’, και “segmentation”, όλα εκ των οποίων υπέστησαν αλλαγές. Επιπρόσθετα ορισμένα χαρακτηριστικά κατα την στατιστική ανάλυση κρίθηκαν άσκοπο να συμμετεχουν στην πειραματική διαδικασία οπότε και αφαιρέθηκαν. Σε αυτήν την κατηγορία ανήκουν τα:

“spouse_index”, “fetch_date” και “cust_code”. Τα 2 τελευταία ωστόσο τονίζουμε πως αρχικά υπέστησαν επεξεργασία σε datetime και integer αντίστοιχα, στη συνέχεια χρησιμοποιήθηκαν στο balance του train set και έπειτα αφαιρέθηκαν καθαρά και μόνο για την εκπαιδευση των μοντέλων. Προβλέφθηκε να ξαναπροστεθούν στη συνέχεια για την εξαγωγή αποτελσμάτων.

5.3 Ισορρόπηση συνόλου δεδομένων

Όπως εξηγήθηκε και στο κεφάλαιο 4 η ανάγκη για ισορρόπηση των δεδομένων του συνόλου δεδομένων εκπαίδευσης είναι επιτακτική προκειμένου να εξαχθούν όσο το δυνατόν πιο αξιόπιστες προβλέψεις από το κάθε μοντέλο. Για το σκοπό αυτό εφαρμόστηκε η τεχνική under sampling συνυπολογίζοντας τον ιδιαίτρα μεγάλο όγκο δεδομένων. Για την παραγωγή προβλέψεων αρκεστήκαμε σε σύνολο τεσσάρων πρϊόντων ενώ πραγματοποιήσαμε 2 διαφορετικά πειράματα ανά πρϊόν και αλγόριθμο, ($1^{ο}$ σενάριο-overall: 0→0, 0→1, 1→0, 1→1 και $2^{ο}$ σενάριο 0→1) όπως προαναφέρθηκε. Ανάλογα με το αν η κυρίαρχη κλάση ήταν πελάτες που διαθέτουν το προϊόν ή όχι, περιορίσαμε τον όγκο δειγμάτων πλειοψηφίας διατηρώντας πάντα σύνολο πελατών που βρίσκεται και στους δύο μήνες (train & test).

5.4 Δείκτες αξιολόγησης

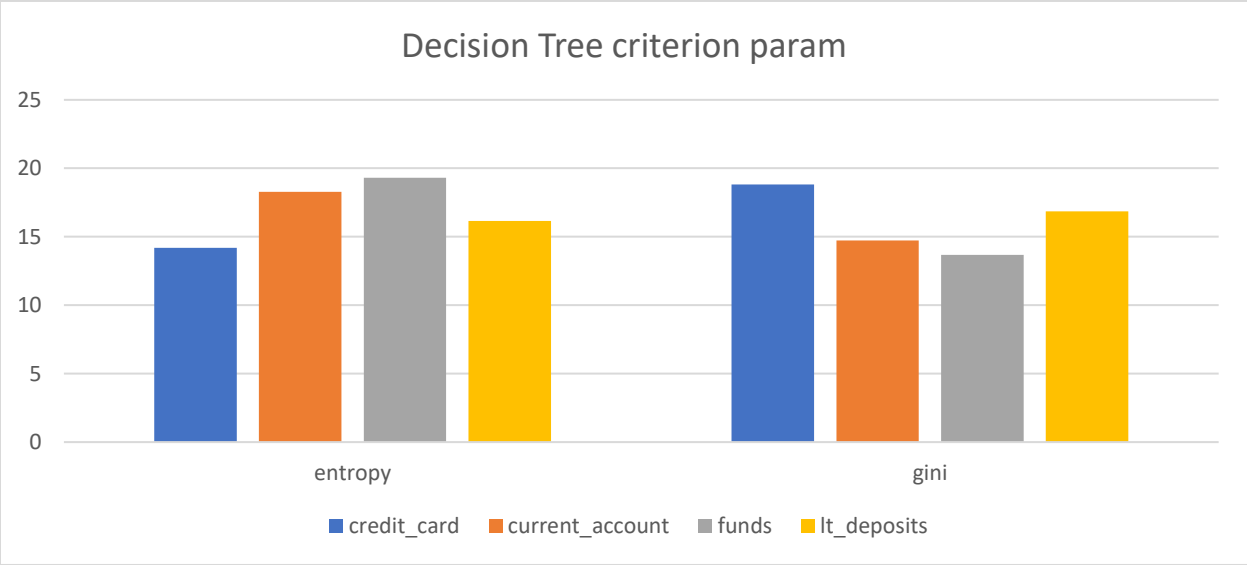
Όπως εξηγήθηκε και κατά την ενότητα 4.5, η ανάλυση των αποτελεσμάτων προκειμένου να βρούμε ποιές υπερπαράμετροι ορίζουν τα βέλτιστα αποτελέσματα για κάθε μοντέλο, έγινε με χρήση του:

$$\textbf{AUC-PRECISION-RECALL Rank} = \frac{\boldsymbol{AUC\ RANK + PRECISION\ RANK + RECALL\ RANK}}{\mathbf{3}}$$

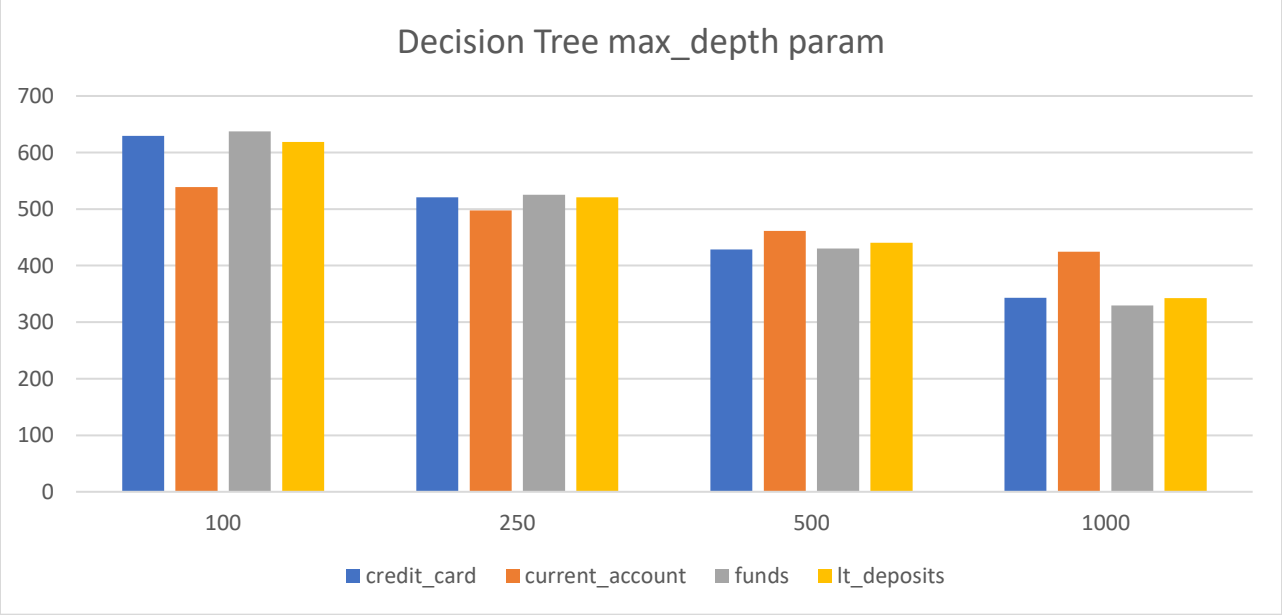
Έπειτα από εκτενή μελέτη και στατιστική ανάλυση των αποτελεσμάτων που παρήχθησαν από κάθε μοντέλο και για κάθε προϊόν, στους ακόλουθους πίνακες φαίνεται ο τρόπος με τον οποίο η μεταβολή κάθε υπερπαραμέτρου επηρεάζει την παραπάνω μετρική για το πρώτο σενάριο των καθολικών προβλέψεων. (Σημείωση: ***Average Rank= AUC-PRECISION-RECALL Rank***)

DECISION TREE

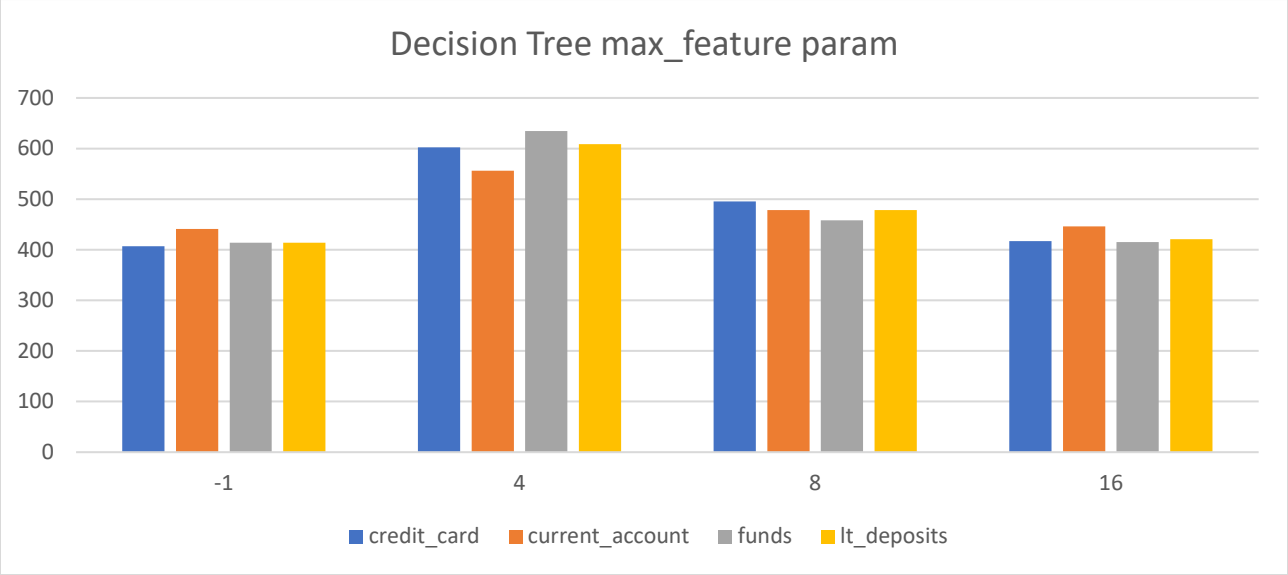
(y-axis → Average Rank, x-axis → hyperparameter value)



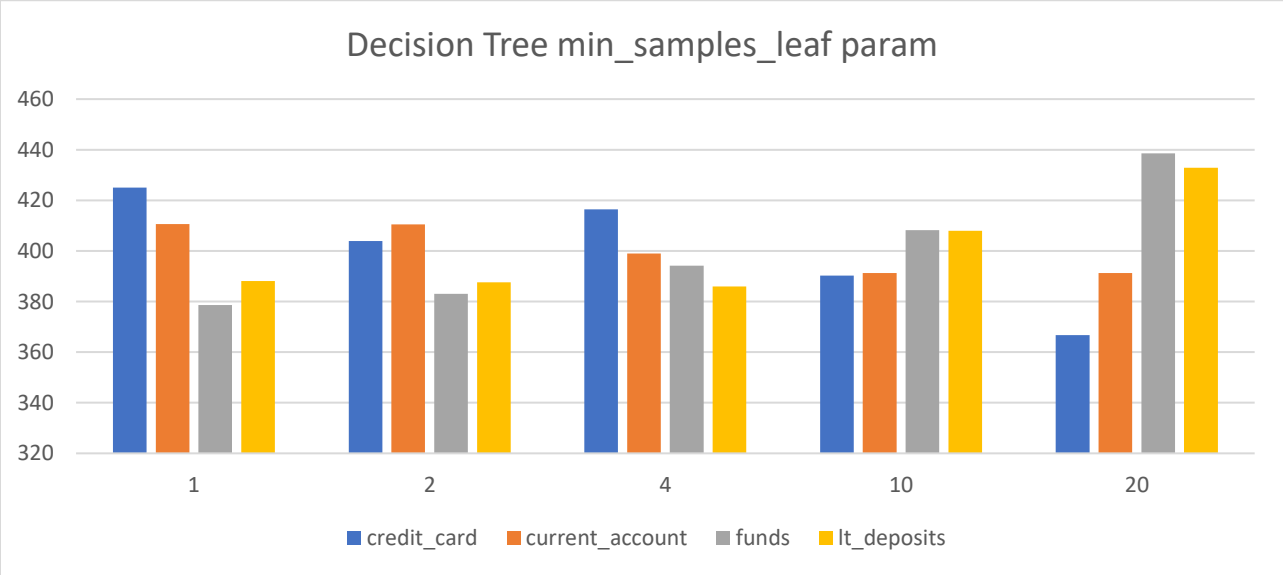
Σχήμα 18:μεταβολή της μετρικής Average Rank για το μοντέλο Decision Tree, ανά προϊόν, για την υπερπαράμετρο criterion



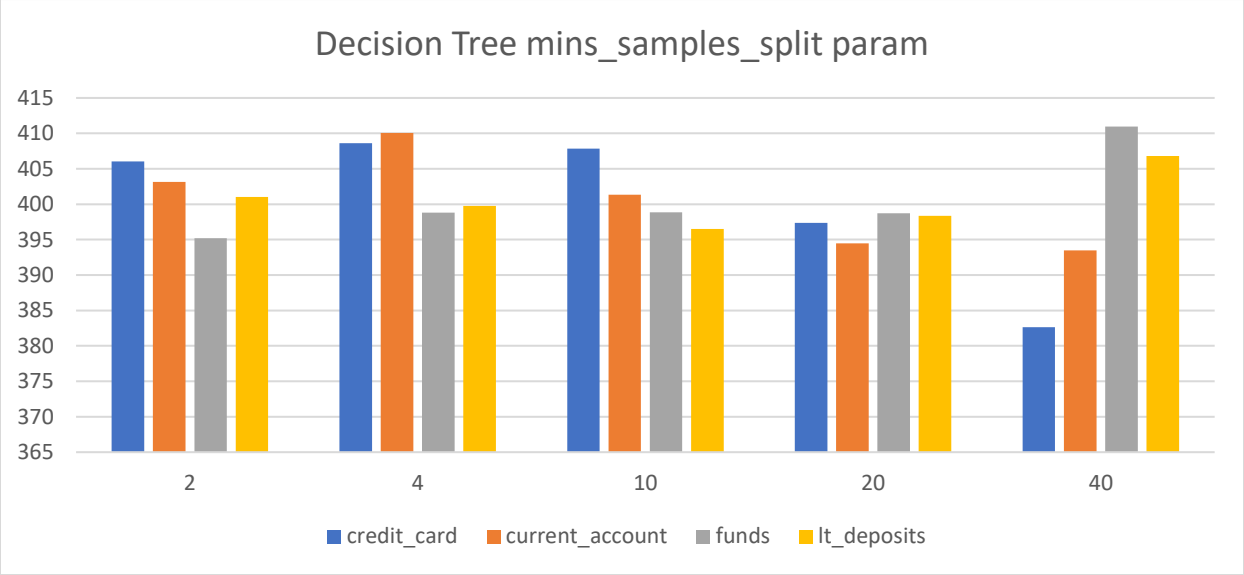
Σχήμα 19:μεταβολή της μετρικής Average Rank για το μοντέλο Decision Tree, ανά προϊόν, για την υπερπαράμετρο max_depth



Σχήμα 20:μεταβολή της μετρικής Average Rank για το μοντέλο Decision Tree, ανά προϊόν, για την υπερπαράμετρο max_feature



Σχήμα 21:μεταβολή της μετρικής Average Rank για το μοντέλο Decision Tree, ανά προϊόν, για την υπερπαράμετρο min_samples_leaf

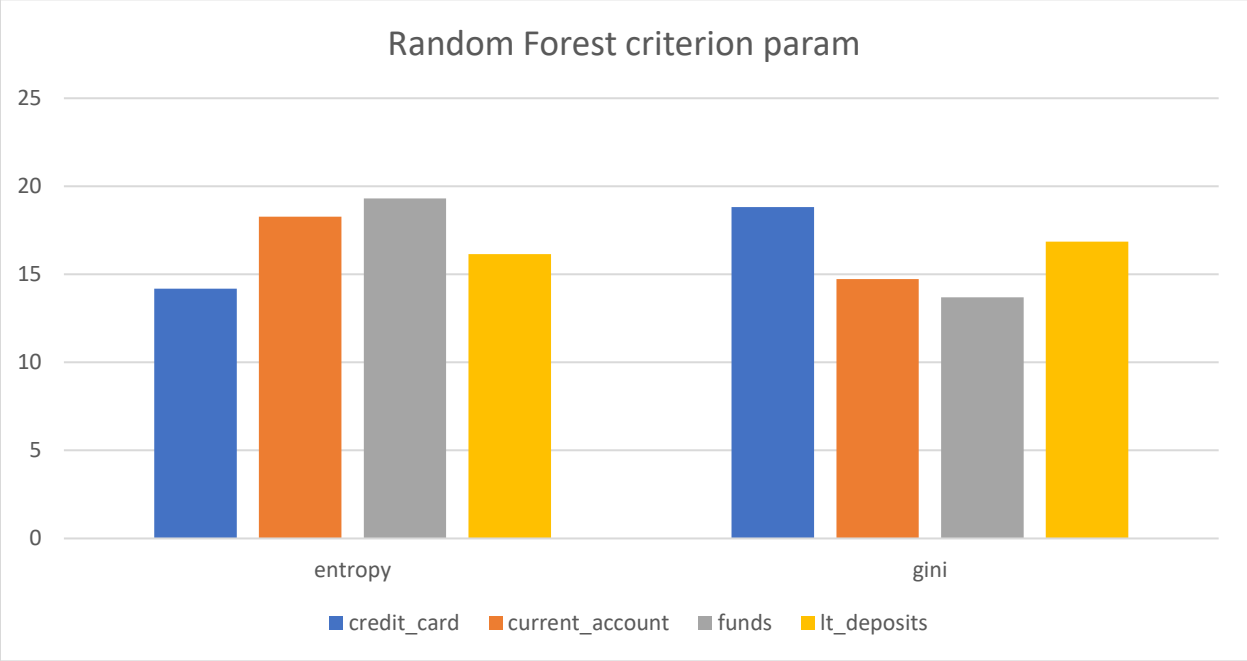


Σχήμα 22:μεταβολή της μετρικής Average Rank για το μοντέλο Decision Tree, ανά προϊόν, για την υπερπαράμετρο min_samples_split

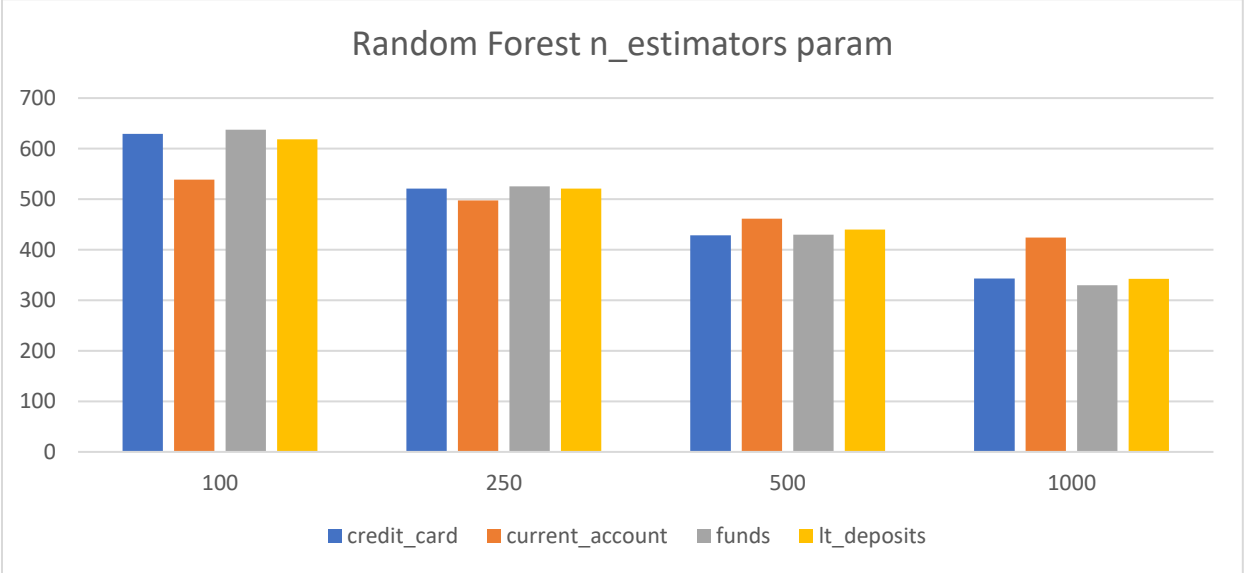
Μέσω των σχημάτων 18 έως 22, γίνεται εμφανές πως ανάλογα με την μεταβολή κάθε υπερπαραμέτρου οδηγούμαστε σε καλύτερα ή χειρότερα αποτελέσματα για το εν λόγω μοντέλο. Υψηλότερη τιμή στον άξονα y σηματοδοτεί μείωση της επίδοσης καθώς υπενθυμίζουμε πως η μετρική μας υποδηλώνει την κατάταξη της παραμετροποίησης του μοντέλου. Καταλήγουμε με βεβαιότητα λοιπόν στις τιμές των υπερπαραμέτρων που πρέπει να θεσπίσουμε προκειμένου να λάβουμε τις εγκυρότερες προβλέψεις. Καθ' ομοίωση με τα παραπάνω λαμβάνουμε τα αντίστοιχα bar charts και για τα άλλα μοντέλα που ακολουθούν.

RANDOM FOREST

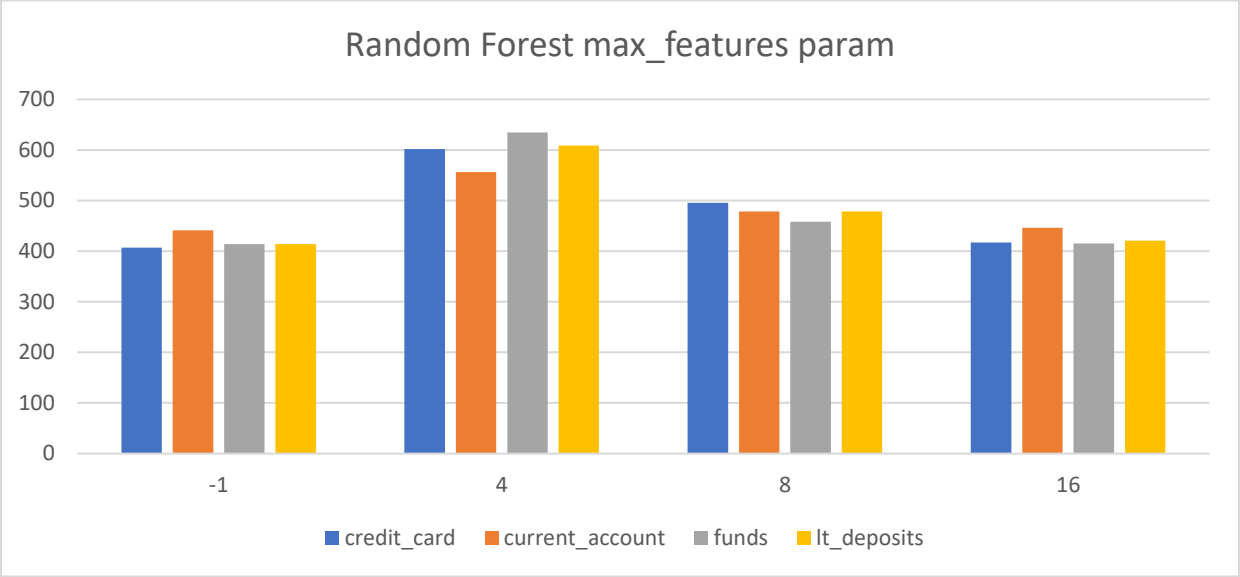
(y-axis → Average Rank, x-axis → hyperparameter value)



Σχήμα 23: μεταβολή της μετρικής Average Rank για το μοντέλο Random Forest, ανά προϊόν, για την υπερπαράμετρο criterion



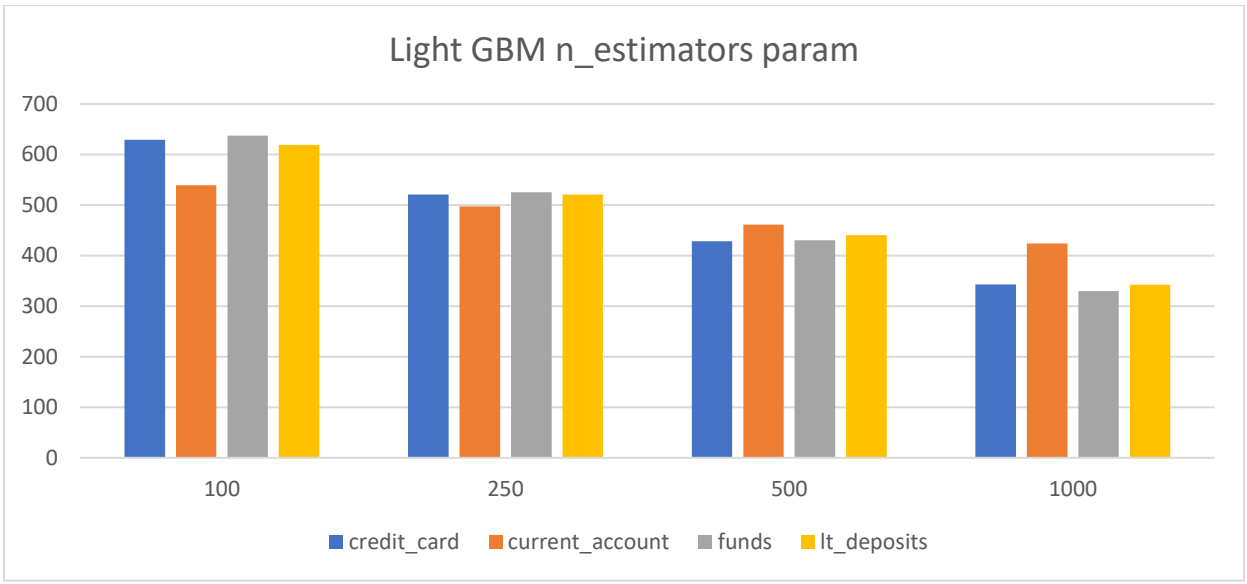
Σχήμα 24: μεταβολή της μετρικής Average Rank για το μοντέλο Random Forest, ανά προϊόν, για την υπερπαράμετρο n_estimators



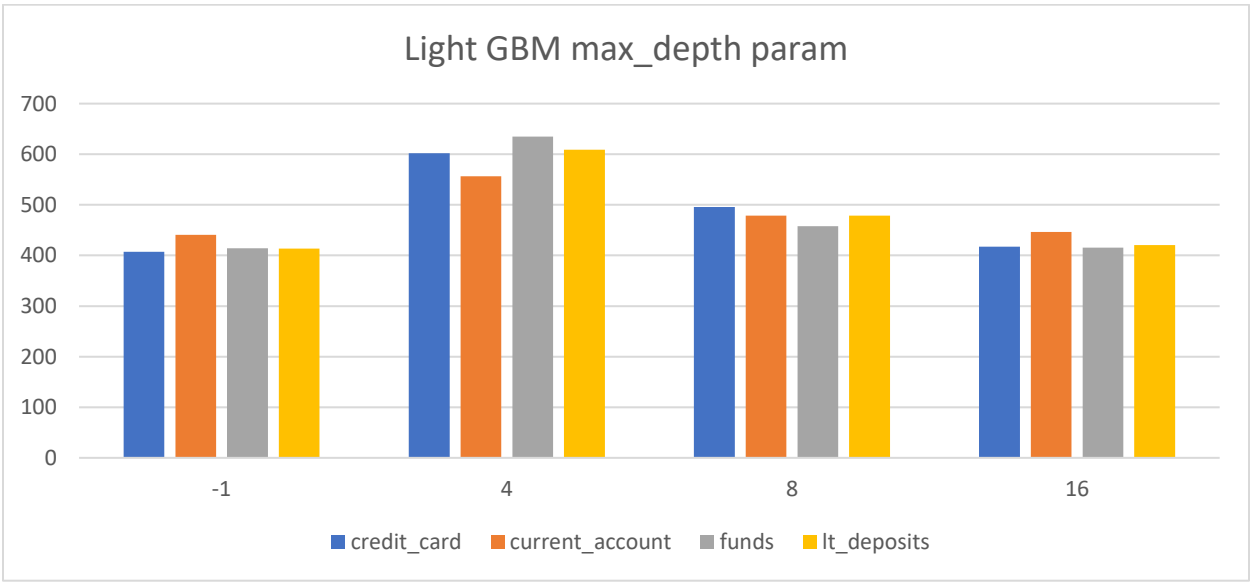
Σχήμα 25: μεταβολή της μετρικής Average Rank για το μοντέλο Random Forest, ανά προϊόν, για την υπερπαράμετρο max_features

LIGHT GBM

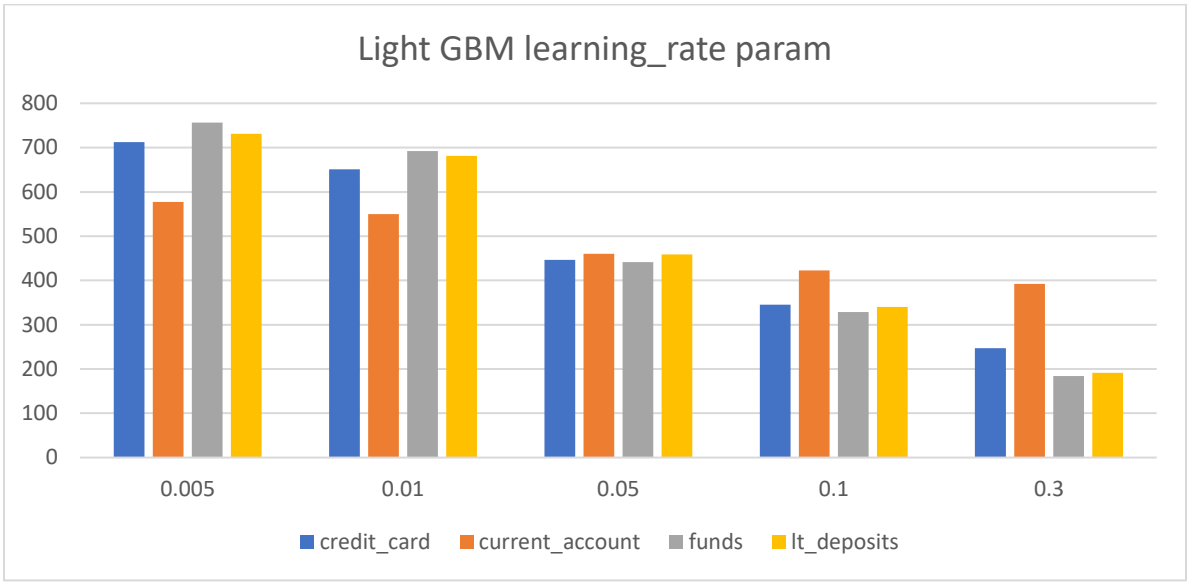
(y-axis → Average Rank, x-axis → hyperparameter)



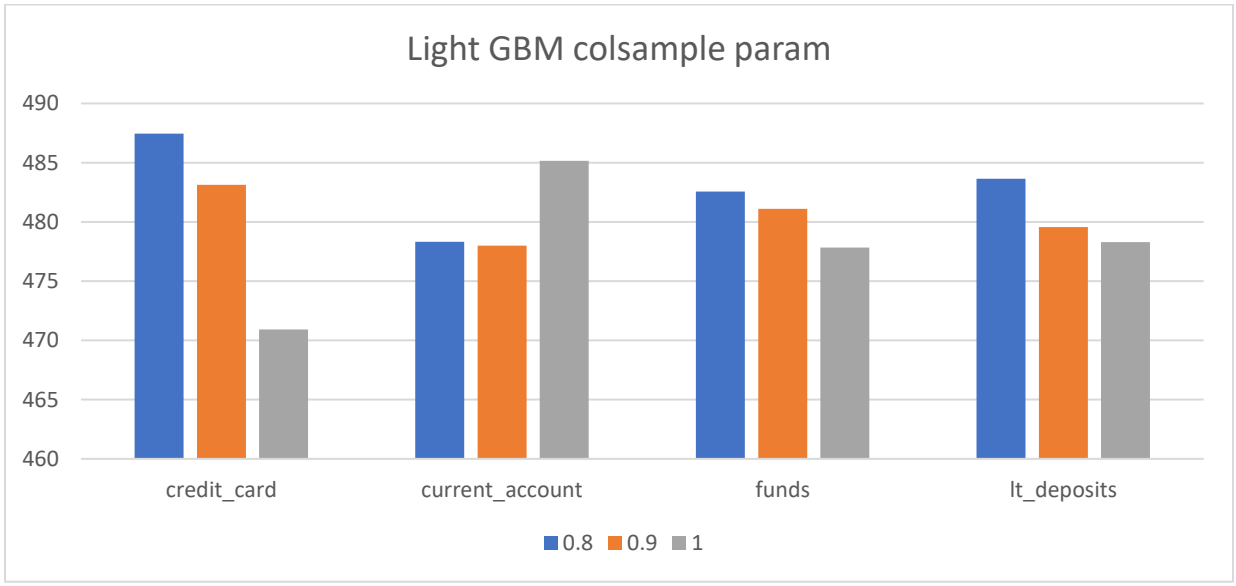
Σχήμα 26: μεταβολή της μετρικής Average Rank για το μοντέλο Light GBM, ανά προϊόν, για την υπερπαράμετρο n_estimators



Σχήμα 27: μεταβολή της μετρικής Average Rank για το μοντέλο Light GBM, ανά προϊόν, για την υπερπαράμετρο max_depth



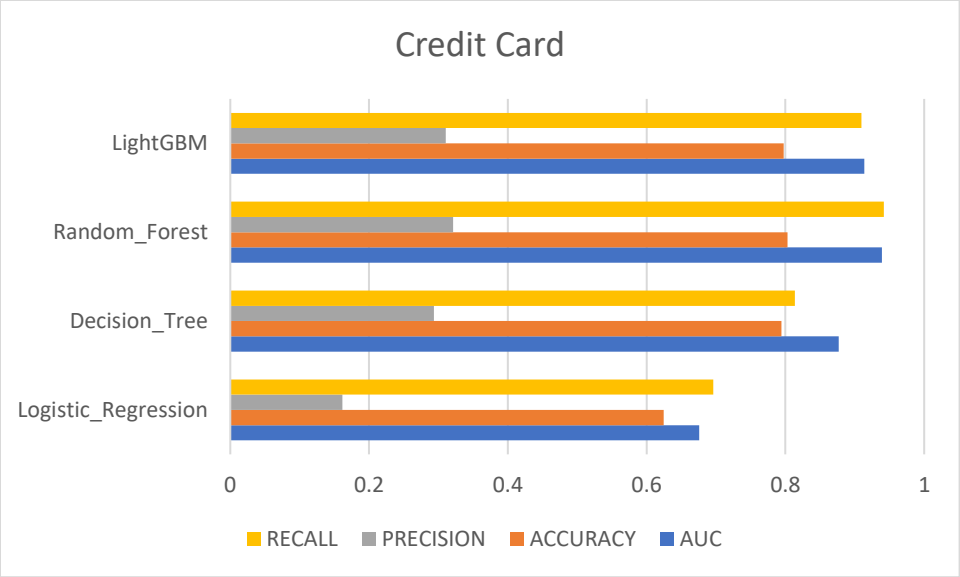
Σχήμα 28: μεταβολή της μετρικής Average Rank για το μοντέλο Light GBM, ανά προϊόν, για την υπερπαράμετρο learning_rate



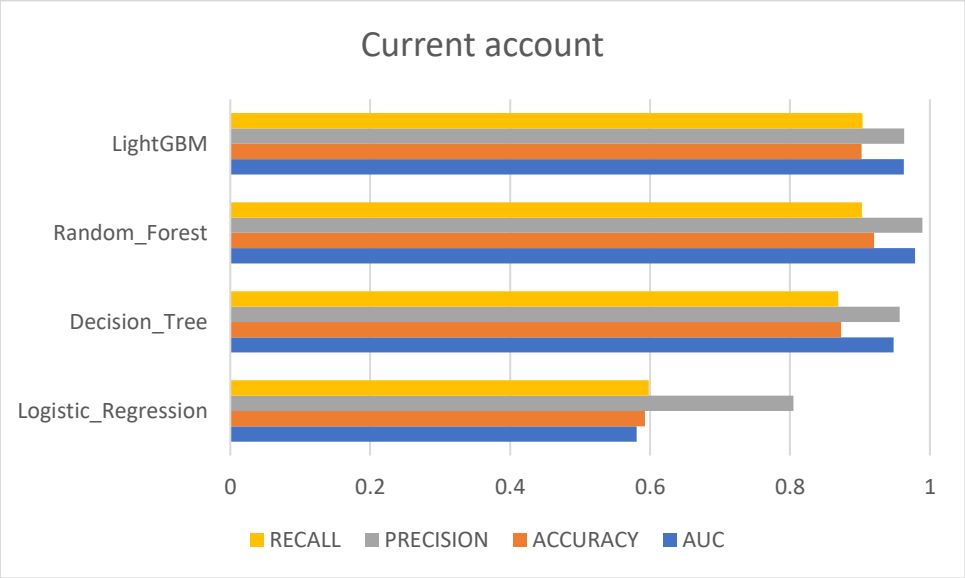
Σχήμα 29: μεταβολή της μετρικής Average Rank για το μοντέλο Light GBM, ανά προϊόν, για την υπερπαράμετρο colsample

Το επόμενο πείραμα περιλαμβάνει τον έλεγχο διαφόρων ταξινομητών για την εύρεση αυτών που αρμόζουν καλύτερα στο συγκεκριμένο πρόβλημα, δηλαδή την πρόβλεψη της συμπεριφοράς των πελατών της τράπεζας σχετικά με την πρόθεση τους να αγοράσουν προϊόντα ή υπηρεσίες. Μελετήθηκαν και πάλι τα ίδια 4 προίοντα και εξετάστηκαν οι ταξινομητές Logistic Regression, Decision Tree, Random Forest και LightGBM. Το πείραμα αυτό διεξήχθη με τα συμπεράσματα των προηγούμενων πειραμάτων, δηλαδή χρησιμοποιήθηκε Random under sampling σε όλο το σύνολο δεδομένων. Τα αποτελέσματα παρουσιάζονται παρακάτω, στα σχήματα 24-27 και έχουν εξαχθεί λαμβάνοντας κάθε φορά τη βέλτιστη παραμετροποίηση κάθε μοντέλου για το υπό εξέταση προϊόν. Από την παρακάτω γραφική παράσταση φαίνεται ότι ο ταξινομητής Logistic Regression δεν έχει καλή επίδοση. Πιο συγκεκριμένα, υπάρχει σημαντική διαφορά μεταξύ αυτού και των υπόλοιπων ταξινομητών στο σύνολο των μετρικών. 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 ταξινομητής Random Forest. Η προκύπτουσα κατάταξη κατά φθίνουσα σειρά επιδόσεων όπως συμπεράινεται απο τα ακόλουθα σχήματα είναι:

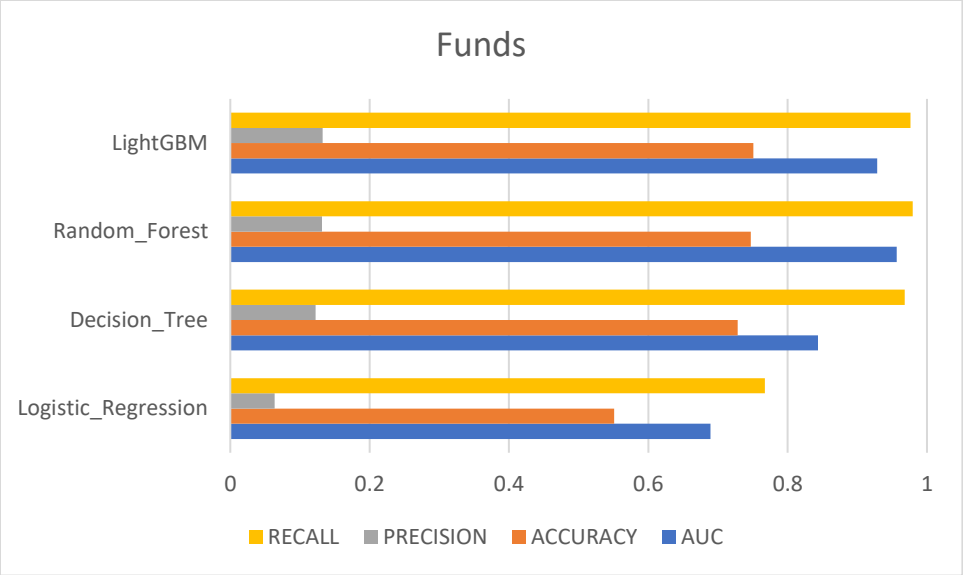
1) *Random forest*
2) *Light GBM*
3) *Decision Tree*
4) *Logistic Regression (Benchmark model)*



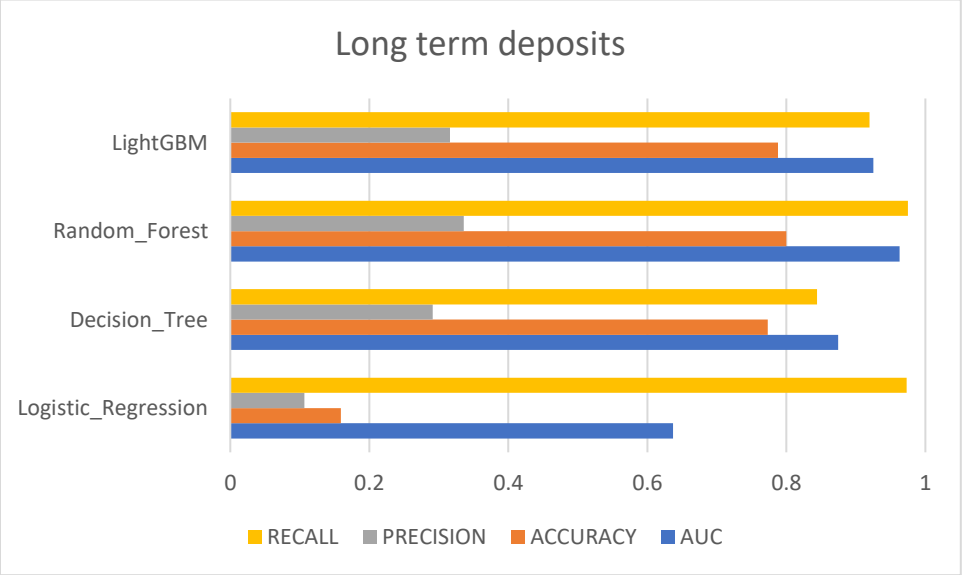
Σχήμα 30: μετρικές RECALL, PRECISION, ACCURACY και AUC για τα μοντέλα Light GBM, Random Forest, Decision Tree, Logistic Regression για το προϊόν credit_card



Σχήμα 31: μετρικές RECALL, PRECISION, ACCURACY και AUC για τα μοντέλα Light GBM, Random Forest, Decision Tree, Logistic Regression για το προϊόν current_account

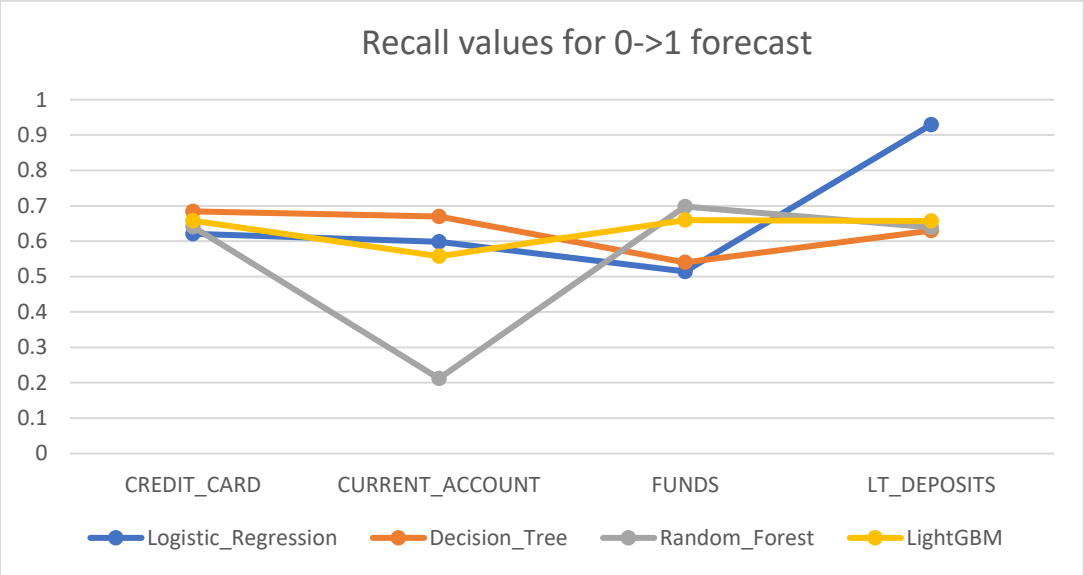


Σχήμα 32: μετρικές RECALL, PRECISION, ACCURACY και AUC για τα μοντέλα Light GBM, Random Forest, Decision Tree, Logistic Regression για το προϊόν funds



Σχήμα 33: μετρικές RECALL, PRECISION, ACCURACY και AUC για τα μοντέλα Light GBM, Random Forest, Decision Tree, Logistic Regression για το προϊόν long-term deposits

Μη αναμενόμενη συμπεριφορά ωστόσο φαίνεται να παρουσιάζεται στην περίπτωση διεξαγωγής του πειράματος του 2ου σεναρίου (0→1 transition), κατά το οποίο υπενθυμίζουμε ότι εξετάζουμε την επίδοση των μοντέλων στην πρόβλεψη αγοράς πρϊόντων από πελάτες που δεν διέθεταν τα εν λόγω 4 προϊόντα. Συγκεκριμένα για το σκοπό αυτό της μέτρησης κρίθηκε σκόπιμο η αποτίμηση της επίδοσης να γίνει με τη μετρική Recall η οποία αναφέρεται στο ποσοστό των σωστών θετικών προβλέψεων σε σχέση με τον συνολικό αριθμό των πραγματικών θετικών περιπτώσεων δηλαδή το άθροισμα όσων πελατών ορθώς προβλέφθηκε ότι θα αγοράσουν ένα προϊον και εκείνων που λανθασμένα προβλέφθηκε ότι δε θα αγοράσουν το προϊόν.



Σχήμα 34: τιμή της μετρικής RECALL για τα μοντέλα Light GBM, Random Forest, Decision Tree, Logistic Regression για τα προϊόντα long-term deposits, credit_card, current_account, funds κατά την πρόβλεψη απόκτησης (0→1) κάθε ενός από τα αντίστοιχα προϊόντα.

5.4 Feature Importance (σημασία χαρακτηριστικών)

Μια πρόσθετη αξιοσημείωτη γραφική παράσταση, είναι αυτή που αναδεικνύει τη σημασία των χαρακτηριστικών που χρησιμοποιήθηκαν για την εκπαίδευση και συγκεκριμένα του μοντέλου που παράγει ο ταξινομητής Random Forest.

Το feature importance βασισμένο στο permutation importance είναι μια τεχνική που χρησιμοποιείται για να αξιολογήσει τη σημαντικότητα των χαρακτηριστικών σε ένα μοντέλο ταξινόμησης με τον αλγόριθμο Random Forest.

Η βασική ιδέα πίσω από το permutation importance είναι ότι αν ανακατέψουμε τις τιμές ενός συγκεκριμένου χαρακτηριστικού στο σύνολο δεδομένων μας, οποιαδήποτε σημασία είχε αυτό το χαρακτηριστικό για το μοντέλο θα χαθεί. Συνεπώς, μπορούμε να υπολογίσουμε πόσο σημαντικό είναι το χαρακτηριστικό μετρώντας την απόκλιση της απόδοσης του μοντέλου μετά το ανακάτεμα (αναδιάταξη) των τιμών του χαρακτηριστικού.

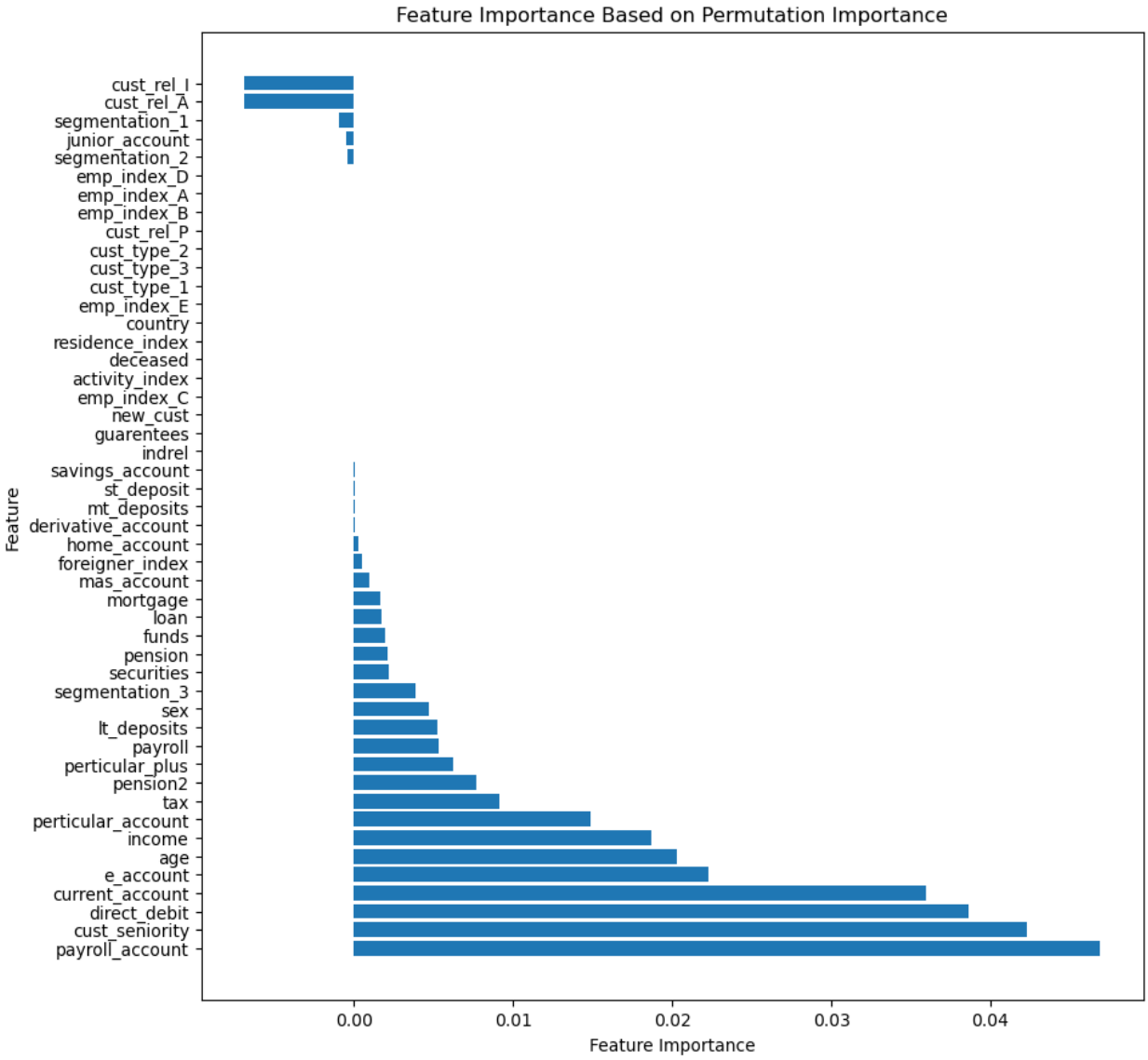
Για να υπολογίσουμε το permutation importance, ακολουθούμε τα εξής βήματα:

- Εκπαιδεύουμε ένα Random Forest μοντέλο χρησιμοποιώντας το σύνολο δεδομένων μας.
- Αξιολογούμε την απόδοση του μοντέλου μας στο σύνολο και κρατάμε το αρχικό σκορ.
- Για κάθε χαρακτηριστικό, αναδιατάσσουμε τις τιμές του στο σύνολο δεδομένων ελέγχου.
- Ξανα-αξιολογούμε την απόδοση του μοντέλου μετά την αναδιάταξη των τιμών του χαρακτηριστικού.
- Υπολογίζουμε τη διαφορά των σκορ πριν και μετά την αναδιάταξη για κάθε χαρακτηριστικό. Αυτή η διαφορά αντιπροσωπεύει το permutation importance του χαρακτηριστικού.
- Ταξινομούμε τα χαρακτηριστικά βάσει της σημαντικότητάς τους, με βάση τη διαφορά των σκορ πριν και μετά την αναδιάταξη.

Με τον τρόπο αυτό, μπορούμε να πάρουμε μια εκτίμηση της σημασίας κάθε χαρακτηριστικού στην πρόβλεψη του μοντέλου. Τα χαρακτηριστικά που έχουν μεγαλύτερη διαφορά σκορ πριν και μετά την αναδιάταξη θεωρούνται πιο σημαντικά για το μοντέλο.

Για παράδειγμα, αν υποθέσουμε ότι είχαμε ένα χαρακτηριστικό "Ηλικία" με τιμές [25, 30, 35, 40, 45]. Κατά την εφαρμογή της τεχνικής του permutation importance, θα τροποποιήσουμε τις τιμές του χαρακτηριστικού με τυχαίο τρόπο. Παραδείγματος χάριν, η νέα ανακατευμένη σειρά μπορεί να είναι [35, 45, 30, 40, 25].

Από τη δική μας μέτρηση για την καθολική πρόβλεψη του πρώτου σεναρίου (0→0, 0→1, 1→0, 1→1) για το προϊόν «credit_card», βλέπουμε ότι τα σημαντικότερα χαρακτηριστικά είναι ορισμένα άλλα προϊόντα που χρησιμοποιήθηκαν στη διαδικασία εκπαίδευσης και παραγωγής προβλέψεων από το μοντέλο. Συγκεκριμένα, πρϊόντα όπως ο λογαριασμός μισθοδοσίας, η άμεση χρέωση και ο τρεχούμνεος λογαριασμός, αποτελούν χαρακτηριστικά που η κατοχή τους αναδυκνείει την πιθανότητα απόκτησης του προϊόντος της πιστωτικής κάρτας ενώ επιπλέον φαίνεται πως ορισμένες κατηγορικές μεταβλητές όπως το εισόδημα, η ηλικία, το φύλο και “αρχαιότητα” ενός πελάτη επηρέαζουν και αυτά την εξέλιξη της πρόβλεψης γεγονός αρκετά αναμενόμενο.



Σχήμα 35: Αξία των χαρακτηριστικών για την εκπαίδευση του Random Forest στην πειραματική διαδικασία πρόβλεψης του προϊόντος credit_card

Κεφάλαιο 6: Συμπεράσματα και Προεκτάσεις

6.1 Συμπεράσματα

Η εφαρμογή της μηχανικής μάθησης καθώς και άλλων στατιστικών τεχνικών σε προβλήματα ταξινόμησης είναι ένας κλάδος ιδιαίτερα διαδεδομένος τα τελευταία χρόνια. Το γεγονός αυτό οφείλεται κατά κύριο λόγο στη μεγάλη επιτυχία που έχουν οι εν λόγω τεχνικές στην επίλυση των προβλημάτων τέτοιου είδους, καθώς και στο πολυπληθές πεδίο εφαρμογών τους. Έχουν αναπτυχθεί ιδιάιτερα προηγμένοι αλγοριθμοι μηχανικής μάθησης για την αντιμετώπιση προβλημάτων ταξινόμησης όπως Random Forests, Gradient Boosting Machines και τα Νευρωνικά Δίκτυα, επιδεικνύοντας παράλληλα εξαιρετικές επιδόσεις και ακρίβεια ενώ τα πολλά εύχρηστα εργαλεία και βιβλιοθήκες, όπως η scikit-learn και η TensorFlow, που διευκολύνουν την εφαρμογή της μηχανικής μάθησης σε προβλήματα τέτοιου τύπου συντελούν προσθέτοντας μεγάλη ευχέρεια, άνεση και ταχύτητα. Αυτά τα εργαλεία παρέχουν ένα ευρύ φάσμα αλγορίθμων και λειτουργιών που είναι εύκολο να χρησιμοποιηθούν από επαγγελματίες και ερευνητές. Είναι, λοιπόν, φυσικό η προσπάθεια εφαρμογής της μηχανικής μάθησης και των νευρωνικών δικτύων στον κλάδο της πρόβλεψης της συμπεριφοράς των καταναλωτών, με σκοπό την αξιοποίηση της ακρίβειας και της ταχύτητας των αποτελεσμάτων που ο άνθρωπος δε μπορεί να συναγωνιστεί να είναι στο απόγειο της.

Με μια πρώτη ματιά στα αποτελέσματα της πειραματικής διαδικασίας γίνεται εύκολα σαφές ότι οι τεχνικές αυτές μπορούν να αποτελέσουν ιδιαίτερα αποδοτικό μέσο παραγωγής προβλέψεων για τον καθορισμό της αγοράς προϊόντων και υπηρεσιών από πελάτες μιας εταιρείας, με βασική προϋπόθεση την κατάλληλη προεπεξεργασία των δεδομένων και τη ρύθμιση συγκεκριμένων παραμέτρων. Τα αποτελέσματα της παρούσας διπλωματικής εργασίας επιβεβαιώνουν την παραπάνω θέση και παρουσιάζουν ενδιαφέρον ως προς την πρόβλεψη της αγοράς προϊόντων μιας τράπεζας. Συγκεκριμένα:

- Έγινε σαφής η ανωτερότητα των μοντέλων που βασίζονται στα δέντρα αποφάσεων στα προβλήματα ταξινόμησης. Πιο συγκεκριμένα, οι ταξινομητές Decision Tree, Random Forest, και LightGBM που χρησιμοποιήθηκαν κατά την πειραματική διαδικασία είχαν καλύτερη επίδοση σε όλες τις μετρικές που εξετάστηκαν σε σχέση με τον ταξινομητή Logistic Regression και τις υπόλοιπες απλούστερες τεχνικές συνεργατικού φιλτραρίσματος. 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 ο τύπος δεδομένων του πελάτη που σχετίζεται πιο άμεσα με την αγορά νέων προϊόντων. Πιο συγκεκριμένα, παρατηρήθηκε ότι ορισμένα δημογραφικά δεδομένα και τα δεδομένα χρήσης άλλων προϊόντων και υπηρεσιών ήταν σημαντικά για την εκπαίδευση των μοντέλων πρόβλεψης και, κατά συνέπεια, την ταξινόμηση του πελάτη ως αγοραστή ή όχι. Αντιθέτως τα one-hot-encoded χαρακτηριστικά δείχνουν να μην επηρεάζουν την πρόβλεψη ενώ και άλλα κατηγορικά features όπως η «χώρα» και το αν είναι νέος πελάτης ή όχι δεν επηρεάζουν καθόλου. Συμπερασματικά η εικόνα που λαμβάνουμε είναι ανάμεικτη καθώς διαφαίνεται ότι το σύνολο σημαντικών χαρακτηριστικών για το υπό εξέταση προϊόν μπορεί να διαφέρει κατά περίπτωση.

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 παραγωγής προτάσεων για υφιστάμενους πελάτες, κατοχυρώνοντας την ικανοποίηση τους από την εταιρεία, ενώ προσελκύονται και νέοι πελάτες στους οποίους μπορούν να γίνου προτάσεις ιδιαίτερα ακριβείς ως προς τις ανάγκες και προτιμήσεις τους, μέσω της σωστής διαχείρισης των αποτελεσμάτων της μεθοδολογίας, με την προσφορά σχετικών προϊόντων για τα οποία τα μοντέλα προέβλεψαν ότι έχουν τη μεγαλύτερη πιθανότητα να αποκτήσουν. Επιπλέον

καθίσταται δυνατή η παροχή εκπτώσεων ή ο καθορισμός άλλων ενεργειών που πιθανώς να συντελέσουν σε μία αμφίδρομα κερδοφόρα κατάσταση. Η πρόβλεψη της συμπεριφοράς των πελατών είναι αδιαμφισβήτητα σημαντική για μια εταιρεία ή έναν οργανισμό, ιδιαίτερα αν πρόκειται για εταιρεία που απασχολεί προσωπικό το οποίο προσφέρει τις υπό μελέτη υπηρεσίες, ή αν πρόκειται για παραγωγική βιομηχανία που επιθυμεί να καθορίσει τους όγκους παραγωγής της και συνεπώς το κέρδος της. Αυτό ισχύει καθώς είναι γεγονός ότι η απόκτηση νέων πελατών αποτελεί σημαντική επένδυση όπως και η διατήρηση και ικανοποίηση των ήδη υπαρχόντων. Με την αδυναία διατήρησης ενός πελάτη από την εταιρεία λόγω πλημμελούς ικανοποίησης του, η επένδυση που συνδέεται με την αρχική απόκτησή του χάνεται μαζί με τον πελάτη. Το ίδιο ακριβώς ισχύει και για την ανάπτυξη υπηρεσιών ή προϊόντων που απευθύνονται σε κοινό που δεν το ενδιαφέρει.

- Δίνει τη δυνατότητα να κατανοηθεί η συμπεριφορά των πελατών σε αντιστοιχία της προτίμησης των προϊόντων και υπηρεσιών. Μπορούν οι αρμόδιοι να αναλύσουν τη σημασία των χαρακτηριστικών των προϊόντων και υπηρεσιών που προσφέρουν καθώς και το κοινό που τα προτιμά, και να εντοπίσουν αυτά που είναι άμεσα συνδεδεμένα με τις τάσεις της αγοράς. Με τον τρόπο αυτό, ανακαλύπτουν τις προτιμήσεις των χρηστών και τις αδυναμίες των προϊόντων ή υπηρεσιών 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.

- Συμβάλλει στην προσαρμοστικότητα στις ανάγκες των πελατών βάσει της πρόβλεψης της συμπεριφοράς τους, και εν συνεχεία δημιουργείται αύξηση στην απόδοση των πωλήσεων, την ικανοποίηση των πελατών και την επικύρωση των πωλήσεων.

- Μπορεί να χρησιμοποιηθεί για την αναγνώριση ανεπιθύμητων συναλλαγών ή απάτης και γενικότερα ενεργειών. Με την ανάλυση των δεδομένων συναλλαγών και την αναγνώριση μοτίβων που αποκαλύπτουν ανορθόδοξη συμπεριφορά, οι εταιρείες δύνανται άμεσα να αναλάβουν δράση για τον περιορισμό του κινδύνου και την προστασία των πελατών τους συμβάλλοντας στην αύξηση της εμπιστοσύνης, στη μείωση της τυχόν απώλειας τους και φυσικά στην αύξηση της κερδοφορίας της ίδιας της επιχείρησης.

- Ένα εξίσου σημαντικό συμπέρασμα είναι ότι η εφαρμογή των μεθόδων που αναλύθηκαν, όχι μόνο εξυπηρετεί το αγοραστικό κοινό, αλλά δημιουργεί και καθορίζει νέες τάσεις στην αγορά, διαμορφώνοντας δυναμικές που ελέγχονται και καθοδηγούνται από την επιχείρηση που επιλέγει να τις εφαρμόσει.

6.2 Προεκτάσεις

Παρά την δυναμική της πειραματικής διαδικασίας και θεωρητικής μελέτης που διενεργήθηκε, λόγω των περιορισμών χρόνου και υπολογιστικών πόρων που περιλαμβάνονται στη διεξαγωγή μιας διπλωματικής εργασίας, υπάρχει πληθώρα βελτιώσεων που θα μπορούσαν να εφαρμοστούν με σκοπό την περαιτέρω αύξηση της απόδοσης των μοντέλων. Ορισμένες από αυτές είναι:

- Εκτενέστερη αναζήτηση των τιμών των υπερπαραμέτρων που χρησιμοποιούνται από τα μοντέλα. Με τον τρόπο αυτό, σε συνδυασμό με την αύξηση των επαναλήψεων και των συνδυασμών για τους οποίους εκτελείται η αναζήτηση, τα μοντέλα πρόβλεψης θα μπορούν να επιτύχουν βελτιωμένη απόδοση και, συνεπώς, καλύτερες τελικές προβλέψεις.

- Εξέταση περισσότερων μοντέλων πρόβλεψης για τον επακριβή προσδιορισμό αυτών που λειτουργούν καλύτερα για το συγκεκριμένο πρόβλημα ταξινόμησης. Για παράδειγμα, θα μπορούσαν να χρησιμοποιηθούν ταξινομητές όπως απλό Gradient Boosting, 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 περισσότερους τομείς και σύνολα δεδομένων (όχι απαραίτητα οικονομικούς). Με τον τρόπο αυτό, θα μπορέσουμε να αυτοματοποιήσουμε και να προσθέσουμε νέα στάδια στην προεπεξεργασία των δεδομένων των υπό εξέταση αλγορίθμων, προκειμένου να εφαρμοστούν και να καλύψουν ποικίλες ανάγκες του ανθρώπου και του περιβάλλοντός του.

- Προέκταση μεθόδων παραγωγής προβλέψεων σε τεχνικές πέραν της Μηχανικής μάθησης. Σημειώνεται εδώ ότι έγινε προσπάθεια διεξαγωγής πειραμάτων βασισμένων σε collaborative filtering χρησιμοποιώντας ευριστικές συναρτήσεις όπως η Ευκλείδια απόσταση και η Συνημιτονειδής ομοιότητα οι οποίες ωστόσο δεν έδειξαν να λειτουργούν αποτελεσματικά για αυτό και η ανάλυση τους παραλήφθηκε. Είμαστε αρκετά αισιόδοξοι ωστόσο ότι με περισσότερο διαθέσιμο χρόνο, πιθανώς να οδηγούμασταν σε αξιοσημείωτα αποτελέσματα με τη χρήση και περισσότερων ευριστικών συναρτήσεων, καθώς οι εν λόγω τεχνικές λαμβάνουν ιδιαίτερης εφαρμογής σε αντίστοιχα υφιστάμενα έργα.

- Θα προβαίναμε σε εξέταση μεγαλύτερου συνόλου προϊόντων και όχι μόνο τεσσάρων ενδεικτικών των κυριότερων κατηγοριών των προϊόντων της τράπεζας για το συγκεκριμένο dataset με σκοπό την πληρέστερη αποκόμιση γνώσης.

- Εξέταση μεγαλύτερου χρονικού διαστήματος και πιθανώς της συνολικής χρονοσειράς για την πρόβλεψη της συμπεριφοράς των καταναλωτών και τη μελέτη προτιμήσεων τους σε βάθος χρόνου καθώς η επιλογή μικρότερου διαστήματος, ενέχει στατιστικά μεγαλύτερη πιθανότητα αναξιόπιστης πρόβλεψης.

Βιβλιογραφία

[1] [https://www.javatpoint.com/supervised-machine-learning , www.javatpoint.com]

[2] [https://towardsdatascience.com/k-means-data-clustering-bce3335d2203 , https://towardsdatascience.com/]

[3] [https://en.wikipedia.org/wiki/Sigmoid_function]

[4] (J. Ross Quinlan. Induction of decision trees. Machine learning, 1(1):81–106, 1986.)

[5] (Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. Classification and regression trees. CRC press, 1984.)

[6] [Sotiris Kotsiantis. Supervised machine learning: A review of classification techniques. Informatica (Ljubljana), 31, 10 2007.]

[7] [Claude E Shannon. A mathematical theory of communication. Bell system technical journal, 27(3):379–423, 1948.]

[8] [Thomas M Cover and Joy A Thomas. Elements of information theory. John Wiley & Sons, 2012.]

[9] [J. Ross Quinlan. Induction of decision trees. Machine learning, 1(1):81–106, 1986.]

[10] [J Ross Quinlan. C4. 5: programs for machine learning. Elsevier, 2014.]

[11] [Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. Knowledge and information systems, 14(1):1–37, 2008.]

[12] [Salvatore Ruggieri. Efficient c4. 5 [classification algorithm]. IEEE transactions on knowledge and data engineering, 14(2):438–444, 2002.]

[13] [Vijay Kotu and Bala Deshpande. Data Science: Concepts and Practice. Morgan Kaufmann, 2018.]

[14] [Leo Breiman. Bagging predictors. Machine learning, 24(2):123–140, 1996]

[15] [Αικατερίνη Καρανικόλα. Κατηγοριοποίηση ομιλητων με χρήση αλγορίθμων μηχανικης μάθησης. PhD thesis, 2017]

[16] [Σταματης Κάρλος. Αναπτ υξη πρωτοτυπων μερικως επιβλεπομενων αλγοριθμων μηχανικης μαθησης ´. PhD thesis, 2020., ref: Mohammed J Zaki and Wagner Meira. Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press, 2014.]

[17] [Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In European conference on computational learning theory, pages 23–37. Springer, 1995]

[18] [Mohammed J Zaki and Wagner Meira. Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press, 2014.]

[19] [Mohammed J Zaki and Wagner Meira. Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press, 2014.]

[20] [Mohammed J Zaki and Wagner Meira. Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press, 2014.]

[21] [Mohammed J Zaki and Wagner Meira. Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press, 2014.]

[22] [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)]

[23] [Leo Breiman. Random forests. Machine learning, 45(1):5–32, 2001].

[24] [P Bickel, P Diggle, S Fienberg, U Gather, I Olkin, and S Zeger. springer series in statistics, 2009.]

[25] [P Bickel, P Diggle, S Fienberg, U Gather, I Olkin, and S Zeger. springer series in statistics, 2009.]

[26] [P Bickel, P Diggle, S Fienberg, U Gather, I Olkin, and S Zeger. springer series in statistics, 2009.]

[27] [Omer Sagi and Lior Rokach. Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4):e1249, 2018.]

[28] [https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding]