

ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



Πτυχιακή Εργασία

Ανάπτυξη εφαρμογής για την πρόβλεψη
αγοράς τραπεζικών προϊόντων μετά από
τηλεφωνικό marketing

Αλέξιος Κανταρτζής

Επιβλέπων: Θεμιστοκλής Έξαρχος

9 Ιουλίου 2024

Περιεχόμενα

1	Εισαγωγή	4
2	Βιβλιογραφική ανασκόπηση	7
3	Περιγραφή Δεδομένων	13
3.1	Επιλογή Δεδομένων	13
3.2	Περιγραφή Μεταβλητών	14
3.3	Τύποι δεδομένων	15
3.3.1	Ονομαστικά δεδομένα	16
3.3.2	Τακτικά δεδομένα	16
3.3.3	Δυαδικά δεδομένα	17
3.3.4	Διακριτά δεδομένα	17
3.3.5	Συνεχή δεδομένα	18
4	Μεθοδολογία	20
4.1	Στατιστική ανάλυση	20
4.1.1	Κατηγορηματικές μεταβλητές	21
4.1.2	Αριθμητικές μεταβλητές	23
4.2	Ανάλυση και επεξεργασία μεταβλητών	23
4.3	Προετοιμασία δεδομένων	37
4.3.1	Αφαίρεση μεταβλητών	37
4.3.2	Δυαδικές μεταβλητές	38
4.3.3	One-hot encoding	38
4.3.4	Binning	38
4.3.5	Label encoding	38
4.3.6	Μεταβλητές day και month	39
4.3.7	Target mean encoding	39
4.3.8	Κανονικοποίηση	39
4.3.9	Διόρθωση ανισορροπίας	40
4.3.10	Επιλογή αλγόριθμων μηχανικής μάθησης	40
4.3.11	K-fold cross validation	40
4.3.12	Παραμετροποίηση του μοντέλου	41
5	Αποτελέσματα	42

5.1	Μετρικές εκτίμησης	42
5.2	Συμπεράσματα αποτελεσμάτων	43
6	Υλοποίηση του συστήματος	45
6.1	Ανάπτυξη διεπαφής προγραμματισμού	45
6.2	Ανάπτυξη εφαρμογής διαχείρισης πελατολόγιου	46
7	Συμπεράσματα	51
8	Κώδικας	54

1 Εισαγωγή

Μία από τις πιο διαδεδομένες μορφές μάρκετινγκ που εφαρμόζεται από οντότητες οι οποίες στοχεύουν στην διαφήμιση και πώληση των προϊόντων ή υπηρεσιών που προσφέρουν, είναι το τηλεφωνικό μάρκετινγκ. Το τηλεφωνικό μάρκετινγκ αποτελεί μία άμεση μορφή μάρκετινγκ και πραγματοποιείται μέσω ειδικευμένων υπαλλήλων οι οποίοι καλούν απευθείας, μέσω τηλεφώνου, τον υποψήφιο πελάτη με σκοπό την διαφήμιση του προϊόντος/υπηρεσίας και κατ' επέκταση την πιθανή αγορά ή συνδρομή από αυτόν τον πελάτη. Μία από τις οντότητες οι οποίες εφαρμόζουν αυτές τις μεθόδους μάρκετινγκ για να διαφημίσουν τις υπηρεσίες που προσφέρουν είναι οι τράπεζες. Σύμφωνα με τους Page Luding (2003), οι περισσότεροι πελάτες κρατούν αρνητική στάση απέναντι στις στρατηγικές του άμεσου μάρκετινγκ όπως το τηλεφωνικό μάρκετινγκ, η οποία μάλιστα πολλές φορές τους δημιουργεί μία αρνητική προδιάθεση ως προς την αγορά του προϊόντος ή υπηρεσίας που τους διαφημίζεται. Για την αποφυγή προβλημάτων όπως αυτό, πολλές τράπεζες αναπτύσσουν μεθόδους για πιο στοχευμένο μάρκετινγκ κάνοντας ανάλυση και μελέτες πάνω σε δεδομένα που σχετίζονται με τις επιδόσεις και τάσεις που έχουν καταγράψει από παρελθοντικές εκστρατείες μάρκετινγκ. Ταυτόχρονα, με το πέρασμα του χρόνου, αλλά και με την εξέλιξη των μεθόδων συλλογής πληροφορίας, ο όγκος των δεδομένων αυτών διαρκώς αυξάνεται. Λόγω αυτού του προβλήματος, έχει δημιουργηθεί μία μεγάλη ανάγκη για συστήματα υποστήριξης αποφάσεων τα οποία θα μπορούν να διαχειριστούν αυτό τον μεγάλο όγκο δεδομένων και να αυτοματοποιήσουν πολλές από τις διαδικασίες που σχετίζονται με τις μεθόδους στοχευμένου μάρκετινγκ. Με την εξέλιξη των πόρων υπολογισμού και με την εφαρμογή μεθόδων μηχανικής μάθησης, η υλοποίηση τέτοιων συστημάτων έχει γίνει πιο προσβάσιμη.

Λόγω της αρνητικής στάσης των πελατών προς το τηλεφωνικό μάρκετινγκ, όπως αναφέραμε πριν, αλλά και λόγω της φύσης των δεδομένων που συλλέγονται από τις εκστρατείες μάρκετινγκ, η ανάπτυξη ενός ποιοτικού συστήματος υποστήριξης αποφάσεων μετατρέπεται σε μία δύσκολη διαδικασία. Η υλοποίηση αυτών των συστημάτων συνήθως γίνεται χρησιμοποιώντας μοντέλα πρόβλεψης τα οποία τα μπορούμε να τα αποκτήσουμε μετά την εκπαίδευση αλγόριθμων μηχανικής μάθησης. Για να γίνει αυτή η εκπαίδευση απαιτείται ένα σετ δεδομένων το οποίο σχετίζεται με το πρόβλημα που θέλουμε να λύσουμε. Η ποιότητα αυτού του σετ δεδομένων αποτελεί ίσως το σημαντικότερο δομικό κομμάτι αυτής της διαδικασίας μιας και από αυτή θα εξαρτηθεί ένα μεγάλο μέρος της ακρίβειας του συστήματος που θέλουμε να αναπτύξουμε. Στην περίπτωση του τηλεφωνικού μάρκετινγκ, τα δεδομένα που συλλέγονται περιέχουν μεταβλητές οι οποίες περιγράφουν πληροφορίες σχετικά με τον πελάτη, την τηλε-

φωνική κλήση ή ακόμη και με την εκστρατία μάρκετινγκ γενικότερα. Για την απόφαση ενός πελάτη σχετικά με την αγορά ή μη αγορά ενός προϊόντος ή υπηρεσίας που του διαφημίζεται, παίζουν πολλές φορές σημαντικό ρόλο δεδομένα που σχετίζονται με κοινωνικές ή ακόμη και ψυχολογικές τάσεις οι οποίες είναι πολύ δύσκολο να μετρηθούν ή ακόμη να καταγραφούν σε ένα σετ δεδομένων. Αυτή η ιδιοτροπία των δεδομένων αυτών προσθέτουν στην δυσκολία της ανάπτυξης συστημάτων υποστήριξης αποφάσεων.

Ένα μεγάλο κομμάτι της βιβλιογραφίας της οποίας θα εξετάσουμε σχετίζεται με την ανάπτυξη συστημάτων υποστήριξης αποφάσεων για την πρόβλεψη πιθανών συνδρομητών σε τραπεζικούς καταθετικούς λογαριασμούς κατόπιν τηλεφωνικού μάρκετινγκ. Η πλειοψηφία από τις έρευνες αυτές χρησιμοποιούν ως σετ δεδομένων, για την ανάλυση και υλοποίηση αυτών των συστημάτων, το Bank Marketing Dataset από το UC Irvine Machine Learning Repository. Σε αυτό εφαρμόζουν μεθόδους περιγραφικής ανάλυσης των δεδομένων, τεχνικές επιλογής μεταβλητών, τεχνικές προετοιμασίας δεδομένων και τεχνικές καταπολέμησης της ανισορροπίας του σετ δεδομένων. Το τελικό στάδιο αυτών των ερευνών συνήθως σχετίζεται με την εκπαίδευση και σύγκριση αλγόριθμων μηχανικής μάθησης με κριτήριο σύγκρισης την ακρίβεια πρόβλεψης την οποία προσφέρουν.

Ένα σημαντικό κομμάτι της ανάπτυξης συστημάτων υποστήριξης αποφάσεων είναι η ίδια η υλοποίηση και ενσωμάτωση τους σε κάποιο πλαίσιο εφαρμογής υπολογιστή ή ηλεκτρονικής υπηρεσίας. Μία σύγχρονη τάση που επικρατεί στον τομέα των ηλεκτρονικών υπηρεσιών είναι η χρήση διεπαφών προγραμματισμού για διαδικτυακές εφαρμογές οι οποίες έχουν ως στόχο τον καταμερισμό των διαδικτυακών εφαρμογών σε μικρές ανεξάρτητες υπηρεσίες. Με αυτή τη μέθοδο η εκάστοτε οντότητα η οποία έχει στην κατοχή της μία ηλεκτρονική εφαρμογή ή βρίσκεται στο στάδιο της υλοποίησης μίας ηλεκτρονικής εφαρμογής, μπορεί να προσθέσει λειτουργίες τις οποίες επιθυμεί σε αυτή χρησιμοποιώντας ηλεκτρονικές μικροϋπηρεσίες που έχουν αναπτυχθεί από τρίτες οντότητες, γλυτώνοντας με αυτό τον τρόπο ένα μέρος του κόστους ανάπτυξης και συντήρησης. Αυτή την τάση την παρατηρούμε ακόμη και στον τομέα της τεχνητής νοημοσύνης όπου πολλές οντότητες βλέποντας την επιτυχία κάποιων μοντέλων επεξεργασίας φυσικής γλώσσας όπως τα GPT μοντέλα, ενσωματώνουν αυτές τις τεχνολογίες, στις ηλεκτρονικές υπηρεσίες όπου κατέχουν, με τη χρήση διεπαφών προγραμματισμού (APIs) όπως το ChatGPT API. Αυτή η δομή ανάπτυξης μπορεί φυσικά να εφαρμοστεί και σε μοντέλα πρόβλεψης ως προς το τραπεζικό τηλεφωνικό μάρκετινγκ. Αυτό μπορεί να επιτευχθεί ενσωματώνοντας το μοντέλο πρόβλεψης που αναπτύσσεται σε μία διεπαφή προγραμματισμού που στόχο έχει την εξυπηρέτηση μίας ηλεκτρονικής διαδικτυακής μικροϋπηρεσίας. Σημαντικό

κομμάτι της ανάπτυξης μιας τέτοιας διεπαφής είναι η ελαχιστοποίηση της πολυπλοκότητας των λειτουργιών της δημιουργώντας μία ξεκάθαρη δομή ως προς την εισαγωγή και εξαγωγή των δεδομένων τα οποία ακολουθούν ένα ορισμένο pipeline. Με αυτό τον τρόπο η διεπαφή αποκτά μία ευελιξία ως προς την ενσωμάτωσή της σε τρίτες εφαρμογές ή ηλεκτρονικές υπηρεσίες.

Σκοπός αυτής της εργασίας είναι η υλοποίηση ενός συστήματος υποστήριξης αποφάσεων το οποίο θα αυτοματοποιεί τη διαδικασία της πρόβλεψης αποτελεσμάτων κατόπιν μίας εκστρατείας τηλεφωνικού μάρκετινγκ που έχει στόχο την συνδρομή υποψήφιων πελατών σε τραπεζικό καταθετικό λογαριασμό. Η υλοποίηση αυτού του συστήματος θα πραγματοποιηθεί χρησιμοποιώντας ένα σετ δεδομένων το οποίο σχετίζεται με το τραπεζικό μάρκετινγκ. Θα εφαρμόσουμε τεχνικές εξόρυξης και προετοιμασίας δεδομένων στις μεταβλητές οι οποίες περιέχονται στο σετ δεδομένων. Θα πραγματοποιηθεί η εκπαίδευση και η σύγκριση κάποιων αλγορίθμων μηχανικής μάθησης απ' τους οποίους θα γίνει η επιλογή του βέλτιστου ως προς την ικανότητα προβλέψεων σε αυτά τα δεδομένα, ο οποίος θα χρησιμοποιηθεί για την ανάπτυξη του μοντέλου πρόβλεψης και κατ' επέκταση της υλοποίησης του συστήματος υποστήριξης αποφάσεων. Η ανάπτυξη αυτού του συστήματος θα γίνει στα πλαίσια μίας διεπαφής προγραμματισμού η οποία θα φέρει την ικανότητα ενσωμάτωσης σε τρίτα λογισμικά διαχείρισης υποψήφιων πελατών τράπεζας. Η κύρια λειτουργία της διεπαφής θα είναι η αποδοχή δεδομένων υποψήφιων πελατών μέσω HTTP πρωτοκόλλου, η επεξεργασία αυτών των δεδομένων, και σε τελικό στάδιο η διαδικασία πρόβλεψης στα δεδομένα αυτά και η αποστολή των αποτελεσμάτων στον αρχικό πομπό. Ταυτόχρονα, θα εξετάσουμε τεχνικές αυτόματης βελτίωσης του συστήματος με την εισαγωγή καινούργιων δεδομένων. Μαζί με το σύστημα υποστήριξης αποφάσεων, θα γίνει η ανάπτυξη μίας εφαρμογής διαχείρισης πελατών, η οποία θα περιέχει γραφικό περιβάλλον και πρόσθετες λειτουργίες οι οποίες σχετίζονται με την διαχείριση πελατολόγιου. Σε αυτή την εφαρμογή θα γίνει η ενσωμάτωση του συστήματος υποστήριξης αποφάσεων για να πραγματοποιήσουμε μία ρεαλιστική προσομοίωση διαχείρισης και πρόβλεψης πελατών με τη χρήση τέτοιων συστημάτων.

2 Βιβλιογραφική ανασκόπηση

Η ανάπτυξη συστημάτων υποστήριξης αποφάσεων που σχετίζονται με τις στοχευμένες μεθόδους μάρκετινγκ σε τραπεζικές υπηρεσίες και προϊόντα αλλά και η ανάλυση αυτών των συστημάτων και των δεδομένων που χρησιμοποιούνται για την υλοποίηση αυτών των συστημάτων έχει δημιουργήσει έναν ολόκληρο τομέα επιστημονικής έρευνας. Ο τομέας αυτός προσπαθεί βρει λύσεις σε προβλήματα και δυσκολίες που αντιμετωπίζουμε κατά την διάρκεια της ανάπτυξης αυτών των συστημάτων υποστήριξης αποφάσεων. Σε αυτό το κεφάλαιο θα γίνει βιβλιογραφική ανασκόπηση σε μερικά από αυτά τα επιστημονικά άρθρα που σχετίζονται με την ανάπτυξη τέτοιων συστημάτων αλλά και την ανάλυση δεδομένων με σκοπό τη βελτίωση του τραπεζικού μάρκετινγκ.

Το πρώτο σκέλος της ανάπτυξης ενός τέτοιου συστήματος είναι η συλλογή δεδομένων που σχετίζονται με την επιρροή εκστρατιών τηλεφωνικού μάρκετινγκ πάνω σε υποψήφιους πελάτες. Το πιο διαδεδομένο σετ δεδομένων που χρησιμοποιείται για την υλοποίηση αυτών των συστημάτων είναι το Bank Marketing Dataset από το UC Irvine Machine Learning Repository. Σύμφωνα με τους Moro, Laureano, Cortez .ά. (2011), η δημιουργία αυτού του σετ δεδομένων έγινε από μία Πορτογαλική τράπεζα, με το δικό της κέντρο επικοινωνίας, συλλέγοντας δεδομένα που σχετίζονταν με τις τηλεφωνικές κλήσεις που πραγματοποιήθηκαν, αλλά και με προσωπικά δεδομένα περιγραφής του πελάτη στον οποίο γινόταν η τηλεφωνική κλήση. Για το τελικό αποτέλεσμα της κάθε τηλεφωνικής κλήσης που πραγματοποιήθηκε, έγινε ο ορισμός μίας δυαδικής μεταβλητής η οποία περιέχει την επιτυχία ή την αποτυχία της συνδρομής σε τραπεζικό καταθετικό λογαριασμό από την πλευρά του πελάτη. Στη συγκεκριμένη έρευνα πραγματοποιήθηκε η υλοποίηση μοντέλων πρόβλεψης, κάνοντας εκπαίδευση των αλγορίθμων Naive Bayes (NB), Decision Trees (DT), και Support Vector Machines (SVM). Για την εξέταση των επιδόσεων αυτών των αλγορίθμων χρησιμοποιήθηκαν οι μετρικές AUC (Area Under the ROC Curve) και ALIFT (Area Under the LIFT Curve) στις οποίες τις καλύτερες επιδόσεις τις είχε ο SVM (0.938, 0.887). Σύμφωνα με τους ερευνητές, η μεταβλητή με τη μεγαλύτερη σημασία στην πρόβλεψη ήταν η διάρκεια της τηλεφωνικής κλήσης (Call Duration) και στη συνέχεια ο μήνας της κλήσης (Month of Contact), και ο αριθμός των προηγούμενων τηλεφωνικών κλήσεων που έχουν πραγματοποιηθεί στον πελάτη (Nr. of Previous Contacts).

Σύμφωνα με τους Govindarajan (2015) μία από τις εξελίξεις στον τομέα της μηχανικής μάθησης είναι τα συνδυαστικά μοντέλα πρόβλεψης (ensemble methods) τα οποία απαρτίζονται από πολλά μοντέλα μέτρων επιδόσεων πρόβλεψης τα οποία συνδυάζονται για την υλοποίηση μο-

ντέλων πρόβλεψης υψηλών επιδόσεων. Στη συγκεκριμένη έρευνα προτείνεται η υλοποίηση συνδυαστικών μοντέλων πρόβλεψης με τη χρήση των αλγόριθμων radial basis neural networks (RBF NN) και support vector machines (SVM) και εφαρμόζοντάς αυτά τα μοντέλα στο Bank Marketing Dataset. Κατά την εξέταση των επιδόσεων των μοντέλων έγινε σύγκριση των συνδυαστικών μοντέλων πρόβλεψης με τις συμβατικές μορφές αυτών των αλγόριθμων. Η μετρική σύγκρισης όπου χρησιμοποιήθηκε κατά την εξέταση των μοντέλων είναι η μετρική ακρίβειας (accuracy). Στα αποτελέσματα της έρευνας παρατηρείτε ότι οι προτεινόμενοι συνδυαστικοί αλγόριθμοι φέρουν μεγαλύτερα ποσά ακρίβειας από τα συμβατικά μοντέλα.

Στην επόμενη έρευνα που εξετάσαμε, οι Patwary .ά. (2021) προτείνουν την υλοποίηση συνδυαστικών μοντέλων χρησιμοποιώντας τους αλγόριθμους naive bayes, support vector machines (SVM) και neural networks (NN). Οι μέθοδοι υλοποίησης συνδυαστικών μοντέλων πρόβλεψης που προτείνει η έρευνα είναι η μέθοδος bagging κατά την οποία το σετ δεδομένων χωρίζεται σε υποσύνολα για να γίνει παράλληλη εκπαίδευση πολλαπλών μοντέλων πρόβλεψης (estimators), η μέθοδος boosting κατά την οποία γίνεται σειριακή εκπαίδευση και παραμετροποίηση πολλαπλών μοντέλων πρόβλεψης, και η μέθοδος stacking όπου συνδυάζονται διαφορετικοί αλγόριθμοι μηχανικής μάθησης για την δημιουργία προβλέψεων. Η υλοποίηση αυτών των μοντέλων πρόβλεψης έγινε πάνω στο Bank Marketing Dataset. Οι μέθοδοι bagging και boosting εφαρμόστηκαν στον κάθε αλγόριθμο ξεχωριστά, ενώ για τη μέθοδο stacking χρησιμοποιήθηκαν οι αλγόριθμοι SVM και neural networks για την υλοποίηση του μοντέλου πρόβλεψης. Οι μετρικές όπου χρησιμοποιήθηκαν για την εξέταση των μοντέλων είναι οι accuracy, sensitivity (precision), specificity (recall), και error-rate. Για τον αλγόριθμο νευρωνικού δικτύου η μέθοδος bagging είχε τα καλύτερα αποτελέσματα και στις 4 μετρικές, ενώ για τους αλγόριθμους SVM και naive bayes η μέθοδος boosting είχε τα καλύτερα αποτελέσματα και στις 4 μετρικές. Στα αποτελέσματα της μεθόδου stacking, η μοναδική μετρική η οποία ήταν καλύτερη από τα συμβατικά μοντέλα ήταν το specificity. Τα καλύτερα αποτελέσματα όλης της έρευνας και στις 4 μετρικές τις έφερε η μέθοδος bagging η οποία εφαρμόστηκε πάνω στον αλγόριθμο neural networks.

Στην επόμενη έρευνα που εξετάσαμε, ο Hossam (2020) εξετάζει τις επιδόσεις περισσότερων συνδυαστικών αλγόριθμων ensemble models που εφαρμόστηκαν πάνω στο Bank Marketing Dataset. Οι δύο βασικές μέθοδοι υλοποίησης συνδυαστικών μοντέλων πρόβλεψης που χρησιμοποιήθηκαν είναι οι bagging και boosting. Οι αλγόριθμοι που χρησιμοποιήθηκαν για την εκπαίδευση είναι οι AdaBoost και LogitBoost για τις μεθόδους boosting, και οι Ρανδομ Φορεστ και Δεξισιον Τρεες (με υλοποίηση μέσω bagging) για τις μεθόδους bagging. Για την

εξέταση των μοντέλων πρόβλεψης χρησιμοποιήθηκαν οι μετρικές accuracy, f1-score, και AUC. Τα αποτελέσματα των αλγόριθμων που εξετάστηκαν δεν δείχνουν κάποια σημαντική διαφορά μεταξύ τους.

Σύμφωνα με τους Tang Zhu (2024), τα συνδυαστικά μοντέλα πρόβλεψης (ensemble models) φέρουν καλύτερες ικανότητες πρόγνωσης και πρόβλεψης, από τα συμβατικά μοντέλα, στο τραπεζικό μάρκετινγκ. Στην έρευνά τους προτείνουν την διαδοχική χρήση δύο συνδυαστικών αλγόριθμων για την βελτίωση των προβλέψεων. Ο πρώτος αλγόριθμος που επιλέχθηκε είναι ο random forest ο οποίος ακολουθεί τη μέθοδο εκπαίδευσης συγκέντρωσης με αντικατάσταση (bagging), ενώ ο δεύτερος είναι ο xgboost ο οποίος ακολουθεί τη μέθοδο εκπαίδευσης ενίσχυσης (boosting). Για την διαδικασία υλοποίησης και διεξαγωγής του μοντέλου, ο αλγόριθμος random forest χρησιμοποιείται για την επιλογή μεταβλητών και για τη δημιουργία των πρώτων προβλέψεων, στη συνέχεια αυτές οι προβλέψεις καταχωρούνται ως δεδομένα για την εκπαίδευση του αλγόριθμου xgboost ο οποίος ύστερα από αυτή τη διαδικασία δημιουργεί τις τελικές προβλέψεις. Κατά την εξέταση των αποτελεσμάτων, το προτεινόμενο μοντέλο έφερε καλύτερες επιδόσεις στις μετρικές accuracy, precision, recall και f1 συγκριτικά με τους υπόλοιπους αλγόριθμους μηχανικής μάθησης.

Η προετοιμασία δεδομένων και συγκεκριμένα των κατηγορηματικών μεταβλητών αποτελεί ένα σημαντικό κομμάτι για την επιτυχημένη εκπαίδευση αλγόριθμων μηχανικής μάθησης. Σύμφωνα με τους Dahouda Joe (2021), η συχνότητα και πυκνότητα κατηγορηματικών δεδομένων σε κάποια σετ δεδομένων δημιουργεί πολλαπλές διαστάσεις στα δεδομένα, πράγμα που μπορεί να οδηγήσει σε προβλήματα. Στην έρευνά τους προτείνουν μία δικιά τους εναλλακτική για την μέθοδο προετοιμασίας των κατηγορηματικών μεταβλητών. Η μέθοδος αυτή εφαρμόζεται μαζί με άλλες τεχνικές προετοιμασίας δεδομένων σε διάφορα σετ δεδομένων, εκ των οποίων και το Bank Marketing Dataset, και διεξάγεται σύγκριση μεταξύ αυτών των μεθόδων ως προς την ικανότητα εκπαίδευσης μίας ομάδας αλγορίθμων μηχανικής μάθησης. Η μέθοδος η οποία προτείνεται από τους ερευνητές είναι εμπνευσμένη από τις τεχνικές επεξεργασίας φυσικής γλώσσας και συγκεκριμένα χρησιμοποιώντας την τεχνική tokenization. Με την εφαρμογή αυτής της μεθόδου δημιουργείται μία μεταβλητή στην οποία εμπεριέχονται οι κατηγορίες/τιμές, όλων των κατηγορηματικών μεταβλητών της καταχώρησης, ως φυσικές λέξεις οι οποίες στη συνέχεια αντιπροσωπεύουν από ένα διάνυσμα. Για την εξέταση των αποτελεσμάτων εφαρμόστηκαν οι τεχνικές προετοιμασίας one-hot encoding, target encoding, binary encoding και η προτεινόμενη μέθοδος. Οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιήθηκαν για την εκπαίδευση είναι οι Logistic Regression, Multi-Layer Perceptron, Random Forest, Gradient

Boosting και LSTM. Σύμφωνα με τα αποτελέσματα το μεγαλύτερο accuracy και precision (73%, 83%) επιτεύχθηκε με τον αλγόριθμο LSTM και την τεχνική προετοιμασίας που προτείνεται στην έρευνα.

Ένα από τα προβλήματα που πολλές φορές συναντάμε σε σετ δεδομένων που πραγματεύονται προβλήματα κατηγοριοποίησης, όπως είναι το πρόβλημα της πρόβλεψης του αν ένας πελάτης θα αγοράσει ή όχι ένα προϊόν, είναι το πρόβλημα της ανισορροπίας (class imbalance) του σετ δεδομένων. Σύμφωνα με τον Verma (2019), οι αλγόριθμοι μηχανικής μάθησης έχουν καλύτερες επιδόσεις όταν η εκπαίδευσή τους γίνεται με τη χρήση ισορροπημένων δεδομένων. Για να λύσει το πρόβλημα της ανισορροπίας, ο ερευνητής εξετάζει τις επιδόσεις στις μεθόδους random undersampling, random oversampling, και synthetic minority oversampling technique (SMOTE).

Name	Method
random undersampling	διαγραφή τυχαίων δεδομένων του majority class
random oversampling	αναπαραγωγή τυχαίων δεδομένων του minority class
SMOTE	δημιουργία συνθετικών δεδομένων για το minority class

Πίνακας 1: Μέθοδοι επίλυσης ανισορροπίας δεδομένων

Στα συμπεράσματα αναφέρει ότι οι μέθοδοι αυτοί ανέβασαν σημαντικά τις μετρικές precision, recall, f1-score και AUCPR για το minority class. Η μέθοδος random oversampling προκαλεί υπερφόρτωση του μοντέλου και γι αυτό είναι προτιμότερη η μέθοδος SMOTE, ενώ η μέθοδος random undersampling, παρόλο που αφαιρεί ένα μεγάλο κομμάτι του σετ δεδομένων, φέρει πολύ καλά αποτελέσματα. Οι μεθοδολογίες αυτές εφαρμόστηκαν πάνω στο Bank Marketing Dataset του UC Irvine Machine Learning Repository με τη χρήση του εργαλείου WEKA.

Η επιλογή μεταβλητών (feature selection) είναι μία σημαντική διαδικασία της προετοιμασίας των δεδομένων, και ειδικά στα σετ δεδομένων με πολλές διαστάσεις. Σύμφωνα με τους Chen .ά. (2020), οι μέθοδοι επιλογής μεταβλητών έχουν την ικανότητα να βελτιώσουν την ακρίβεια αλλά και την ικανότητα πρόβλεψης των μοντέλων μηχανικής μάθησης. Στη συγκεκριμένη έρευνα πραγματοποιήθηκε η εξέταση και η σύγκριση μεθόδων επιλογής μεταβλητών οι οποίες εφαρμόστηκαν σε διάφορα σετ δεδομένων, εκ των οποίων και το Bank Marketing Dataset. Οι μέθοδοι επιλογής δεδομένων όπου εφαρμόστηκαν είναι οι Boruta, Recursive Feature Elimination (RFE) , και Random Forest Feature Importance (RF). Οι αλγόριθμοι μηχανικής μάθησης που χρησιμοποιήθηκαν για την εκπαίδευση είναι οι Support Vector Machines (SVM),

K-Nearest Neighbor (KNN), Latent Dirichlet Allocation (LDA), και Random Forest (RF). Στα αποτελέσματα της έρευνας όλες οι μέθοδοι επιλογής δεδομένων φαίνεται να έχουν καλά αποτελέσματα με τη μέθοδο Random Forest να ξεχωρίζει σε ελάχιστο βαθμό.

Στο επόμενο άρθρο που αναλύσαμε, ο Wang (2020) προσπαθεί να λύσει το πρόβλημα της ανισορροπίας, του Bank Marketing Dataset, εφαρμόζοντας την τεχνική random undersampling όπου αναφέραμε προηγουμένως. Μετά την εφαρμογή της μεθόδου random undersampling στο σετ δεδομένων, το ποσοστό των θετικών καταγραφών καταλάμβανε το 56.51% του σετ δεδομένων ενώ των αρνητικών καταγραφών το 43.49%. Αυτό σημαίνει ότι μετέτρεψε την ομάδα των θετικών καταγραφών από minority class σε majority class. Εκτός αυτού, πραγματοποίησε την επιλογή των μεταβλητών σύμφωνα με την στατιστική συσχέτιση που είχαν μεταξύ τους. Στις μεταβλητές age και balance εντοπίστηκαν outliers τους οποίους ο ερευνητής τους αντικατέστησε με τις πλησιέστερες τιμές αυτών των μεταβλητών. Στη συνέχεια έγινε η υλοποίηση του μοντέλου πρόβλεψης κάνοντας εκπαίδευση του αλγόριθμου C5.0 decision tree. Η τελική ακρίβεια προβλέψεων που επιτεύχθηκε στην έρευνα είναι 89.3%.

Ένα σημαντικό κομμάτι της ανάπτυξης ρεαλιστικών μοντέλων πρόβλεψης είναι η ικανότητα βελτίωσης των επιδόσεων αυτών των μοντέλων με μεθόδους επανεκπαίδευσης. Σύμφωνα με τους Kaiser .ά. (2024), η επανεκπαίδευση των μοντέλων πρόβλεψης που σχετίζονται με το τηλεφωνικό μάρκετινγκ είναι πολύ σημαντική μιας και η στάση των πελατών απέναντι στο προϊόν μπορεί να αλλάξει εξαιτίας τρίτων παραγόντων, πράγμα που καθιστά την συνεχής ανανέωση του μοντέλου απαραίτητη. Αυτή η επανεκπαίδευση του μοντέλου μπορεί να πραγματοποιηθεί με τη συλλογή νέων δεδομένων και με την εφαρμογή μεθόδων όπως το online learning. Στην συγκεκριμένη έρευνα χρησιμοποιήθηκε το σετ δεδομένων Bank Marketing, του UC Irvine Machine Learning Repository, πάνω στο οποίο εφαρμόστηκαν μέθοδοι προετοιμασίας όπως label encoding, one hot encoding, και feature scaling. Μετά πραγματοποιήθηκε η επιλογή των μεταβλητών, και στη συνέχεια η αντιμετώπιση της ανισορροπίας του σετ δεδομένων με τη μέθοδο SMOTE. Η ανάπτυξη των μοντέλων πρόβλεψης πραγματοποιήθηκε με την εκπαίδευση των αλγόριθμων RF, AdaBoost (ADA), GB και XGB. Τέλος, εφαρμόστηκε η μέθοδος online learning όπου το 20% του σετ δεδομένων χωρίστηκε σε δύο σύνολα δεδομένων τα οποία χρησιμοποιήθηκαν για την επανεκπαίδευση των μοντέλων πρόβλεψης. Κατά την εξέταση της επίδοσης των μοντέλων παρατηρήθηκε μία αύξηση του 4%-5% στις επιδόσεις των αλγορίθμων μετά την εφαρμογή της μεθόδου online learning, πράγμα που καθιστά αυτή τη μέθοδο σημαντική για την ανάπτυξη ρεαλιστικών μοντέλων πρόβλεψης.

Ένα φαινόμενο που παρατηρούμε στην πλειοψηφία της βιβλιογραφίας που σχετίζεται με το Bank Marketing Dataset, είναι η χρήση της μεταβλητής duration η οποία εμπεριέχεται μέσα στο σετ δεδομένων. Αυτή η μεταβλητή περιγράφει την διάρκεια που κράτησε η τηλεφωνική κλήση η οποία έγινε στον πελάτη. Σε ένα ρεαλιστικό πλαίσιο πρόβλεψης, η γνώση της τιμής αυτής της μεταβλητής δεν είναι δυνατή, και για αυτό το λόγο δεν θα τη συμπεριλάβουμε κατά την εκπαίδευση των αλγόριθμων μηχανικής μάθησης. Ταυτόχρονα, το duration αποτελεί τη μεταβλητή με τη μεγαλύτερη στατιστική συσχέτιση σχετικά με τα αποτελέσματα και για αυτό το λόγο οι μετρικές απόδοσης του τελικού μοντέλου που θα υλοποιήσουμε θα περιέχουν ασθενέστερες τιμές από αυτές της υπάρχουσας βιβλιογραφίας.

3 Περιγραφή Δεδομένων

Το πρώτο βήμα πριν την εκπαίδευση κάποιου αλγόριθμου μηχανικής μάθησης είναι η εύρεση ενός σετ δεδομένων στο οποίο θα πρέπει να πραγματοποιηθεί μία επεξηγηματική ανάλυσή του, και μετέπειτα η προεπεξεργασία του και η εξόρυξη πληροφορίας από αυτό. Για να το πετύχουμε αυτό, το σετ δεδομένων θα πρέπει να περιέχει επαρκή δεδομένα ως προς την ποσότητα αλλά και ως προς την ποιότητά τους. Λέγοντας ποσότητα, εννοούμε τον αριθμό καταχωρίσεων ή καταγραφών δεδομένων, ενώ με τη λέξη ποιότητα εννοούμε την αποτελεσματικότητα του συνόλου των μεταβλητών του σετ δεδομένων ως προς την περιγραφή και επίλυση του προβλήματος που προσπαθούμε να επιλύσουμε. Σκοπός της εφαρμογής που θα υλοποιήσουμε είναι η πρόβλεψη της πιθανής συνδρομής σε κάποιο καταθετικό τραπεζικό λογαριασμό από ένα σετ πελατών. Το αποτέλεσμα αυτών των προβλέψεων θα περιορίζεται σε δύο τιμές (0 αν ο εκάστοτε πελάτης δεν ανοίξει λογαριασμό, και 1 αν ο πελάτης ανοίξει λογαριασμό), αυτό σημαίνει ότι το πρόβλημα που θα προσπαθήσουμε να λύσουμε είναι πρόβλημα κατηγοριοποίησης (ή αλλιώς ταξινόμησης). Δηλαδή, το σετ δεδομένων που χρειαζόμαστε θα πρέπει να περιέχει χαρακτηριστικά που θα περιγράφουν τον κάθε πελάτη αλλά και στοιχεία σχετικά με την τηλεφωνική επαφή με τον πελάτη, και τέλος το αποτέλεσμα της εκάστοτε τηλεφωνικής επαφής με τον κάθε πελάτη, δηλαδή το αν ήταν επιτυχής ή αποτυχή.

3.1 Επιλογή Δεδομένων

Το σετ δεδομένων το οποίο επιλέχθηκε για την ανάλυση και εξόρυξη δεδομένων είναι το Bank Marketing Dataset από το UC Irvine Machine Learning Repository . Μέσα σε αυτό βρίσκονται καταγραφές σχετικές με κάποιες εκστρατίες μάρκετινγκ που πραγματοποιήθηκαν και καταγράφηκαν από ένα Πορτογαλικό τραπεζικό ινστιτούτο. Οι εκστρατίες αυτές έγιναν μέσω τηλεφωνικού μάρκετινγκ. Στόχος αυτών των εκστρατιών ήταν η συνδρομή πελατών σε κάποιον καταθετικό λογαριασμό που πρόσφερε η τράπεζα. Το σετ δεδομένων περιέχει 45211 καταγραφές οι οποίες είναι κατανεμημένες χρονολογικά (από τις παλαιότερες προς τις καινούργιες), και 17 μεταβλητές (εκ των οποίων και το τελικό αποτέλεσμα της κάθε σειράς) τα οποία περιγράφουν την κάθε καταγραφή. Μέσα στα δεδομένα δεν περιέχεται κανένα κενό στοιχείο.

3.2 Περιγραφή Μεταβλητών

Προτού προχωρήσουμε στην ανάλυση αυτών των δεδομένων θα πρέπει πρώτα να κάνουμε μία περιγραφή της κάθε μεταβλητής (feature) που υπάρχει μέσα στο σετ δεδομένων που επιλέξαμε και να την κατατάξουμε σε συγκεκριμένες κατηγορίες. Ο πρώτος διαχωρισμός που πρέπει να κάνουμε είναι σχετικά με την περιγραφική τους αξία, δηλαδή ως προς το αντικείμενο που περιγράφει η κάθε μεταβλητή. Οι μεταβλητές του σετ δεδομένων στην προκειμένη περίπτωση χωρίζονται σε τρεις κατηγορίες. Η πρώτη κατηγορία μεταβλητών σχετίζεται με δεδομένα τα οποία περιγράφουν τον πελάτη στον οποίο γίνεται το τηλεφώνημα.

Μεταβλητές	Περιγραφή
Προσωπικά	
age	Ηλικία.
job	Δουλειά/απασχόληση.
marital	Οικογενειακή κατάσταση.
education	Μορφωτικό επίπεδο.
Οικονομικά	
default	Περιγράφει το αν ο εκάστοτε πελάτης έχει πιστωτικό χρέος.
balance	Μέσο ετήσιο υπόλοιπο τραπεζικού λογαριασμού (σε ευρώ).
housing	Περιγράφει το αν ο εκάστοτε πελάτης έχει λάβει στεγαστικό δάνειο.
loan	Περιγράφει το αν ο εκάστοτε πελάτης έχει λάβει προσωπικό δάνειο.

Πίνακας 2: Μεταβλητές: Περιγραφή πελάτη

Η δεύτερη κατηγορία μεταβλητών σχετίζεται με δεδομένα τα οποία περιγράφουν στοιχεία σχετικά με τις τηλεφωνικές κλήσεις αυτές καθ' αυτές που έχουν γίνει στον εκάστοτε πελάτη.

Μεταβλητές	Περιγραφή
Τωρινή εκστρατεία	
contact	Είδος τηλεφωνικής συσκευής (σταθερό/κινητό) που χρησιμοποιήθηκε στην κλήση από την πλευρά του πελάτη.
day	Η μέρα στην οποία πραγματοποιήθηκε η τηλεφωνική κλήση (με αριθμό).
month	Ο μήνας στον οποίο πραγματοποιήθηκε η τηλεφωνική κλήση (με ονομασία).
duration	Διάρκεια της τηλεφωνικής κλήσης (σε δευτερόλεπτα).
campaign	Αριθμός τηλεφωνικών κλήσεων που πραγματοποιήθηκαν στην τωρινή εκστρατία μάρκετινγκ στον συγκεκριμένο πελάτη.
Προηγούμενη εκστρατεία	
pdays	Αριθμός ημερών που πέρασαν από την τελευταία τηλεφωνική κλήση που πραγματοποιήθηκε σε προηγούμενη εκστρατία μάρκετινγκ.
previous	Αριθμός τηλεφωνικών κλήσεων που πραγματοποιήθηκαν σε προηγούμενη εκστρατία μάρκετινγκ στον συγκεκριμένο πελάτη.
poutcome	Αποτέλεσμα της τηλεφωνικής κλήσης από την προηγούμενη εκστρατία μάρκετινγκ στον συγκεκριμένο πελάτη.

Πίνακας 3: Μεταβλητές: Περιγραφή τηλεφωνικής κλήσης

Τέλος, η τρίτη κατηγορία αποτελείται από μία μεταβλητή, στην προκειμένη περίπτωση είναι η τελική μεταβλητή (target variable) y , η οποία καθορίζει το τελικό αποτέλεσμα μιας τηλεφωνικής κλήσης και αποτελεί την πιο σημαντική μεταβλητή για την εκπαίδευση του αλγορίθμου μηχανικής μάθησης.

3.3 Τύποι δεδομένων

Πέρα από τον διαχωρισμό των μεταβλητών σχετικά με την περιγραφική τους αξία, πολύ σημαντικός είναι και ο διαχωρισμός τους σχετικά με τον τύπο δεδομένων που περιέχει η κάθε μια από αυτές μιας και η αντιμετώπιση του κάθε τύπου δεδομένων είναι διαφορετική ως προς την

διαδικασία προεπεξεργασίας που θα δούμε παρακάτω. Οι δύο βασικοί τύποι δεδομένων που συναντάμε στο σετ δεδομένων που έχουμε επιλέξει είναι τα κατηγορηματικά και τα αριθμητικά δεδομένα/μεταβλητές. Κατηγορηματικά δεδομένα ονομάζουμε τις μεταβλητές οι οποίες παίρνουν μη-αριθμητικές τιμές, ενώ αριθμητικά δεδομένα ονομάζουμε τις μεταβλητές οι οποίες παίρνουν αριθμητικές τιμές. Όπως θα δούμε και παρακάτω, αυτοί οι δύο τύποι δεδομένων χωρίζονται σε άλλες υποκατηγορίες σύμφωνα με τον τρόπο τον οποίο εκφράζουν τις τιμές τους.

3.3.1 Ονομαστικά δεδομένα

Τα ονομαστικά δεδομένα είναι ένα είδος κατηγορηματικών δεδομένων τα οποία περιγράφουν τις κατηγορίες/τιμές μιας μεταβλητής χωρίς να υπάρχει κάποια σειρά ή κατάταξη αυτών των κατηγοριών/τιμών.

Μεταβλητές	Περιγραφή τιμών μεταβλητής
job	Διαχειριστής (admin.), Άγνωστο (unknown), Άνεργος (unemployed), Διευθυντής (management), Οικιακή Βοηθός (housemaid), Επιχειρηματίας (entrepreneur), Εργάτης (blue-collar), Αυτοαπασχολούμενος (self-employed), Συνταξιούχος (retired), Τεχνικός (technician), Υπηρεσίες (services), Φοιτητής (student).
marital	Παντρεμένος (married), Χωρισμένος (divorced) (περιγράφει και τα άτομα που είναι χήροι), Ελεύθερος (single).
contact	Σταθερό (telephone), Κινητό (cellular), Άγνωστο (unknown).
poutcome	Αποτυχία (failure), Επιτυχία (success), Άλλο (other), Άγνωστο (unknown).

Πίνακας 4: Μεταβλητές: Περιγραφή ονομαστικών μεταβλητών

3.3.2 Τακτικά δεδομένα

Τα τακτικά δεδομένα είναι ένα άλλο είδος κατηγορηματικών δεδομένων τα οποία περιγράφουν τις κατηγορίες/τιμές μιας μεταβλητής αλλά, σε αντίθεση με τα ονομαστικά δεδομένα, φέρουν μία ξεκάθαρη σειρά ή κατάταξη των κατηγοριών/τιμών αυτών.

Μεταβλητές	Περιγραφή τιμών μεταβλητής
education	Πρωτοβάθμια εκπαίδευση (primary), δευτεροβάθμια εκπαίδευση (secondary), τριτοβάθμια εκπαίδευση (tertiary), Άγνωστο (unknown).
day	Παίρνει ακέραιες αριθμητικές τιμές μεταξύ του 1 και του 31 οι οποίες περιγράφουν την ημέρα του μήνα στην οποία πραγματοποιήθηκε η τηλεφωνική κλήση στον πελάτη. Παρόλο που οι τιμές που παίρνει η συγκεκριμένη μεταβλητή είναι αριθμητικές, οι τιμές αυτές στην ουσία συμβολίζουν τις ημέρες του μήνα.
month	Ιανουάριος (jan), Φεβρουάριος (feb), Μάρτιος (mar), Απρίλιος (apr), Μάιος (may), Ιούνιος (jun), Ιούλιος (jul), Αύγουστος (aug), Σεπτέμβριος (sep), Οκτώβριος (oct), Νοέμβριος (nov), Δεκέμβριος (dec).

Πίνακας 5: Μεταβλητές: Περιγραφή τακτικών μεταβλητών

3.3.3 Δυαδικά δεδομένα

Τα δυαδικά δεδομένα είναι το τελευταίο είδος κατηγορηματικών δεδομένων που θα συναντήσουμε σε αυτό το σετ δεδομένων. Αυτή η κατηγορία μεταβλητών μπορεί να πάρει μονάχα δύο κατηγορίες/τιμές. Αυτές οι δύο κατηγορίες/τιμές συνήθως συμβολίζουν αντίθετες ή αντιτιθέμενες καταστάσεις σχετικά με το αντικείμενο το οποίο περιγράφουν.

Μεταβλητές	Περιγραφή τιμών μεταβλητής
default	Ναι (yes), Όχι (no)
housing	Ναι (yes), Όχι (no)
loan	Ναι (yes), Όχι (no)
y	Ναι (yes), Όχι (no)

Πίνακας 6: Μεταβλητές: Περιγραφή δυαδικών μεταβλητών

3.3.4 Διακριτά δεδομένα

Τα διακριτά δεδομένα είναι ένα είδος αριθμητικών δεδομένων που αποτελούνται από ξεχωριστές, διακριτές τιμές που δεν μπορούν να υποδιαιρεθούν περαιτέρω. Αυτές οι τιμές είναι συνήθως ακέριοι αριθμοί και αντιπροσωπεύουν μετρήσεις ή κατηγορίες.

Μεταβλητές	Περιγραφή τιμών μεταβλητής
campaign	Παίρνει ακέραιες αριθμητικές τιμές οι οποίες προστίθενται για να περιγράψουν τον συνολικό αριθμό τηλεφωνικών κλήσεων που έγιναν στην τωρινή εκστρατία μάρκετινγκ στον συγκεκριμένο πελάτη. Η μέγιστη τιμή που συναντάται στη συγκεκριμένη μεταβλητή είναι 63, ενώ η ελάχιστη είναι 1.
pdays	Παίρνει ακέραιες αριθμητικές τιμές οι οποίες προστίθενται για να περιγράψουν τον συνολικό αριθμό μερών που έχουν περάσει από την τελευταία τηλεφωνική κλήση που πραγματοποιήθηκε σε προηγούμενη εκστρατία μάρκετινγκ στον συγκεκριμένο πελάτη. Η μέγιστη τιμή που συναντάται στη συγκεκριμένη μεταβλητή είναι 871, ενώ η ελάχιστη είναι -1. Στην περίπτωση όπου η τιμή της μεταβλητής είναι -1, όπως αναγράφεται στις οδηγίες του σετ δεδομένων, αυτό σημαίνει ότι δεν έχει πραγματοποιηθεί στο παρελθόν καμία τηλεφωνική κλήση στον συγκεκριμένο πελάτη συνεπώς δεν υπάρχει ούτε αριθμητικό σύνολο ημερών που έχουν περάσει από κάποια τηλεφωνική κλήση σε προηγούμενη εκστρατία μάρκετινγκ στον συγκεκριμένο πελάτη.
previous	Παίρνει ακέραιες αριθμητικές τιμές οι οποίες προστίθενται για να περιγράψουν τον συνολικό αριθμό τηλεφωνικών κλήσεων που έγιναν σε προηγούμενες εκστρατίες μάρκετινγκ στον συγκεκριμένο πελάτη. Η μέγιστη τιμή που συναντάται στη συγκεκριμένη μεταβλητή είναι 275, ενώ η ελάχιστη είναι 0.

Πίνακας 7: Μεταβλητές: Περιγραφή διακριτών μεταβλητών

3.3.5 Συνεχή δεδομένα

Τα συνεχή δεδομένα είναι ένα είδος αριθμητικών δεδομένων τα οποία αντιπροσωπεύουν μετρήσεις που μπορούν να λάβουν οποιαδήποτε τιμή μέσα σε ένα συγκεκριμένο εύρος. Αυτές οι τιμές συχνά εκφράζονται ως δεκαδικοί αριθμοί και μπορούν να υποδιαιρεθούν σε μικρότερες μονάδες.

Μεταβλητές	Περιγραφή τιμών μεταβλητής
age	Παίρνει ακέραιες αριθμητικές τιμές οι οποίες περιγράφουν την ηλικία του πελάτη. Στο συγκεκριμένο σετ δεδομένων υπάρχουν μόνο πελάτες όπου οι ηλικίες τους ξεκινάνε από τα 18 και άνω. Η μέγιστη ηλικία που συναντάται στη συγκεκριμένη μεταβλητή είναι 95 ετών, ενώ η ελάχιστη είναι 18 ετών.
balance	Παίρνει ακέραιες αριθμητικές τιμές οι οποίες περιγράφουν το μέσο τραπεζικό υπόλοιπο (της χρονιάς) του πελάτη σε ευρώ. Το μέγιστο υπόλοιπο που συναντάται σε αυτή τη μεταβλητή είναι τα 102127 ευρώ, ενώ το ελάχιστο είναι τα -8019 ευρώ. Η συγκεκριμένη μεταβλητή παίρνει και αρνητικές τιμές στην περίπτωση όπου το υπόλοιπο του πελάτη είναι αρνητικό.
duration	Παίρνει ακέραιες αριθμητικές τιμές οι οποίες περιγράφουν την διάρκεια της τηλεφωνικής κλήσης που πραγματοποιήθηκε στον πελάτη σε δευτερόλεπτα. Η μέγιστη τιμή που συναντάται στη συγκεκριμένη μεταβλητή είναι τα 4918 δευτερόλεπτα, ενώ η ελάχιστη τιμή είναι τα 0 δευτερόλεπτα.

Πίνακας 8: Μεταβλητές: Περιγραφή συνεχή μεταβλητών

4 Μεθοδολογία

Σε αυτό το σημείο, θα γίνει η περιγραφή της μεθοδολογίας την οποία ακολουθήσαμε για να πετύχουμε τους στόχους που έχουμε ορίσει. Θα αναφερθούμε στις μεθόδους στατιστικής ανάλυσης και προεπεξεργασίας του σετ δεδομένων, της εκπαίδευσης και σύγκρισης αλγορίθμων μηχανικής μάθησης, αλλά και της ανάπτυξης μιας εφαρμογής η οποία θα ενσωματώνει και θα αυτοματοποιεί αυτές τις μεθόδους. Ταυτόχρονα θα γίνεται αναφορά των εργαλείων των οποίων χρησιμοποιήθηκαν για την εφαρμογή αυτών των μεθόδων.

4.1 Στατιστική ανάλυση

Ξεκινώντας με την στατιστική ανάλυση των δεδομένων που μας δίνονται θα πρέπει να μελετήσουμε πρώτα την τελική μεταβλητή y (target variable) η οποία μας δίνει ουσιαστικά το αποτέλεσμα του κάθε τηλεφωνήματος που πραγματοποιήθηκε σε κάθε πελάτη. Το άμα δηλαδή η τηλεφωνική κλήση προς τον πελάτη ήταν επιτυχής ή αποτυχής. Η συγκεκριμένη μεταβλητή είναι δυαδική, δηλαδή παίρνει μόνο δύο τιμές (yes/no). Αυτές τις τιμές θα τις χαρακτηρίζουμε ως θετικές (επιτυχής τηλεφωνική κλήση) όταν το αποτέλεσμα είναι yes, ενώ όταν είναι no, θα τις χαρακτηρίζουμε ως αρνητικές (αποτυχής τηλεφωνική κλήση).

Όπως αναφέρθηκε και στην Επιλογή Δεδομένων, ο συνολικός αριθμός καταχωρήσεων του σετ δεδομένων είναι 45211 καταγραφές. Από αυτές, οι 39922 έχουν αρνητικό αποτέλεσμα, ενώ οι υπόλοιπες 5289 έχουν θετικό αποτέλεσμα. Αυτό σημαίνει ότι παρατηρούμε εξ αρχής μια μεγάλη ανισορροπία μεταξύ των δύο κλάσεων του αποτελέσματος μιας και τα θετικά αποτελέσματα κατέχουν μονάχα το 11,7% του συνόλου καταχωρήσεων του σετ δεδομένων, ενώ το υπόλοιπο 88.3% ανήκει στα αρνητικά αποτελέσματα. Αυτή η μεγάλη ανισορροπία ενδεχομένως θα πρέπει να αντιμετωπιστεί μετά το κομμάτι της προεπεξεργασίας δεδομένων που θα δούμε παρακάτω.

Target	Total	Percentage
Yes	5289	11,7%
No	39922	88.3%

Πίνακας 9: Ανάλυση τελικής μεταβλητής

Εφόσον έχουμε αναλύσει ξεχωριστά την τελική μεταβλητή y , θα πρέπει να κάνουμε στατιστική ανάλυση όλες τις υπόλοιπες μεταβλητές του σετ δεδομένων. Η ανάλυση αυτή θα πρέπει να γίνει ξεχωριστά σε όλες τις μεταβλητές, αλλά και σε σύγκριση με την τελική μεταβλητή y ,

για να διαπιστώσουμε το ενδεχόμενο του να υπάρχει κάποια σημαντική στατιστική συσχέτιση των μεταβλητών αυτών με την τελική μεταβλητή. Οι μέθοδοι που θα χρησιμοποιήσουμε για να το πετύχουμε αυτό είναι διαφορετικές για τον κάθε τύπο δεδομένων.

4.1.1 Κατηγορηματικές μεταβλητές

Για να διαπιστώσουμε το αν μία κατηγορηματική μεταβλητή έχει κάποια στατιστική συσχέτιση με την τελική μεταβλητή y θα εφαρμόσουμε την μέθοδο χ^2 (chi-square). Οι τιμές που θα μας δωθούν από την εφαρμογή αυτής της μεθόδου θα χρησιμοποιηθούν για να γίνει ο υπολογισμός του p-value για την κάθε κατηγορηματική μεταβλητή την οποία εξετάζουμε. Προτού εφαρμόσουμε αυτή τη μέθοδο, θα πρέπει να ορίσουμε τη μηδενική υπόθεση H_0 (null hypothesis) η οποία μας λέει ότι η κατηγορηματική μεταβλητή που εξετάζουμε αλλά και η τελική μεταβλητή είναι ανεξάρτητες μεταξύ τους. Αυτό πρακτικά σημαίνει ότι υποθέτουμε ότι αυτές οι κατηγορίες/τιμές τις οποίες θα εξετάσουμε θα μοιράζονται την ίδια κατανομή με αυτή του συνόλου αποτελεσμάτων της τελικής μεταβλητής, δηλαδή 11.7% για τα θετικά αποτελέσματα και 88.3% για τα αρνητικά. Με τα αποτελέσματα της μεθόδου χ^2 θα μάθουμε το ποσό απόκλισης που υπάρχει στις κατανομές αυτών των κατηγοριών/τιμών. Τέλος, με τον υπολογισμό του p-value θα συμπεράνουμε το αν η μηδενική υπόθεση που δημιουργήσαμε θα απορριφθεί ή όχι.

Αφού επιλέξουμε την μεταβλητή την οποία θέλουμε να εξετάσουμε, δημιουργούμε έναν πίνακα (contingency table) ο οποίος περιέχει όλες τις κατηγορίες/τιμές που εμπεριέχονται στη συγκεκριμένη μεταβλητή, τον συνολικό αριθμό εμφάνισής τους στο σετ δεδομένων, και τον αριθμό θετικών και αρνητικών αποτελεσμάτων για την κάθε κατηγορία/τιμή της μεταβλητής. Αφού γίνει αυτό, υπολογίζουμε την προβλεπόμενη συχνότητα των θετικών και αρνητικών αποτελεσμάτων για την κάθε κατηγορία/τιμή της μεταβλητής.

$$E_{ij} = \frac{(\text{Row Total} \times \text{Column Total})}{\text{Grand Total}}$$

Στη συνέχεια υπολογίζουμε το χ^2 (chi-square) όπου O_{ij} είναι οι παρατηρούμενες συχνότητες των κατηγοριών/τιμών ενώ το E_{ij} είναι οι προβλεπόμενες συχνότητες των κατηγοριών/τιμών, τις οποίες υπολογίσαμε προηγουμένως:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Μετά που υπολογίσουμε το χ^2 (chi-square) θα πρέπει να υπολογίσουμε τον βαθμό ελευθερίας της συγκεκριμένης μεταβλητής, όπου r είναι ο συνολικός αριθμός σειρών του πίνακα, ενώ c

είναι ο συνολικός αριθμός στηλών του πίνακα:

$$df = (r - 1) \times (c - 1)$$

Προτού υπολογίσουμε το p-value, θα πρέπει να θέσουμε ένα επίπεδο σημαντικότητας. Στην προκειμένη περίπτωση το επίπεδο σημαντικότητας που θα δώσουμε είναι το 0.05 το οποίο είναι το πιο διαδεδομένο επίπεδο σημαντικότητας που χρησιμοποιείται σε στατιστικές αναλύσεις. Στην περίπτωση που το p-value που θα βγάλουμε είναι μικρότερο από το επίπεδο σημαντικότητας που έχουμε θέσει, τότε αυτό σημαίνει ότι η μηδενική υπόθεση ακυρώνεται και το χαρακτηριστικό που εξετάσαμε έχει στατιστική συσχέτιση με την τελική μεταβλητή y. χρησιμοποιώντας το χ^2 chi-square και τον βαθμό ελευθερίας που βρήκαμε μπορούμε να υπολογίσουμε το p-value της μεταβλητής που εξετάζουμε.

$$p\text{-value} = 1 - F(\chi^2, df)$$

Ο υπολογισμός του βαθμού ελευθερίας (degrees of freedom), της προβλεπόμενης συχνότητας, του chi-square και του p-value έγινε χρησιμοποιώντας τη βιβλιοθήκη SciPy της python. Παρακάτω βρίσκεται ο πίνακας των αποτελεσμάτων με σειρά στατιστικής σημασίας:

Feature	Chi-square	Degrees of Freedom	P-value
month	3061.839	11	0.000e+00
poutcome	4391.507	3	0.000e+00
contact	1035.714	2	1.252e-225
housing	874.822	1	2.919e-192
job	836.105	11	3.337e-172
day	574.051	30	6.896e-102
education	238.924	3	1.627e-51
loan	209.617	1	1.665e-47
marital	196.496	2	2.145e-43
default	22.202	1	2.454e-06

Πίνακας 10: Στατιστική σημαντικότητα: Κατηγορηματικές μεταβλητές

Όπως βλέπουμε στον πίνακα των αποτελεσμάτων, παρατηρούμε ότι όλες οι κατηγορηματικές μεταβλητές φαίνεται να έχουν στατιστική συσχέτιση συγκριτικά με την τελική μεταβλητή y, μιας και το p-value όλων των μεταβλητών που εξετάσαμε μέχρι στιγμής δείχνει να είναι

μικρότερο από το επίπεδο σημαντικότητας το οποίο θέσαμε (0.05) πράγμα που σημαίνει ότι απορρίπτεται η μηδενική υπόθεση για αυτές τις μεταβλητές. Τέλος, παρατηρούμε ότι οι μεταβλητές με την μεγαλύτερη στατιστική σημασία από το σετ των κατηγορηματικών μεταβλητών, φαίνεται να είναι το month και το poutcome, όπου το p-value τους ισούται με το μηδέν.

4.1.2 Αριθμητικές μεταβλητές

Για να διαπιστώσουμε το αν μια αριθμητική μεταβλητή έχει κάποια στατιστική συσχέτιση με την τελική μεταβλητή y θα εφαρμόσουμε την μέθοδο independent t-test. Προτού ξεκινήσουμε θα πρέπει να ορίσουμε την μηδενική υπόθεση H_0 με την οποία, στην προκειμένη περίπτωση, υποθέτουμε ότι οι μέσος όρος των τιμών που είχαν θετικό αποτέλεσμα θα είναι ίδιος με τον μέσο όρο των τιμών που είχαν αρνητικό αποτέλεσμα. Η εναλλακτική υπόθεση H_1 είναι το ότι αυτοί οι μέσοι όροι δεν θα είναι ίδιοι. Για την κάθε ομάδα αποτελεσμάτων (Yes/No) θα πρέπει να υπολογίσουμε τους μέσους όρους ($\bar{X}$) και την τυπική απόκλιση (s), της κάθε μεταβλητής που εξετάζουμε.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}}$$

$$p\text{-value} = 2 \times \text{CDF}(-|t|, df)$$

Αφού υπολογίσουμε το t-test, υπολογίζουμε τον βαθμό ελευθερίας και μετέπειτα το p-value για το οποίο θέτουμε το επίπεδο σημαντικότητας στο 0.05.

4.2 Ανάλυση και επεξεργασία μεταβλητών

Οι πρώτη μεταβλητή που θα πρέπει να αναλύσουμε είναι το month το οποίο μας δίνει τον μήνα τον οποίο πραγματοποιήθηκε μία τηλεφωνική κλήση σε πελάτη. Η μεταβλητή αυτή περιέχει όλους τους μήνες του χρόνου, χωρίς να λείπει κάποιος από το σετ δεδομένων. Είναι σημαντικό να έχουμε στο νου μας το ότι το σετ δεδομένων εκτυλίσσεται σε ένα χρονικό πλαίσιο το οποίο ξεπερνάει το ένα έτος, με αποτέλεσμα να επαναλαμβάνονται κάποιοι μήνες του χρόνου. Αυτό

εν μέρει δημιουργεί μια διαφορά στον αριθμό καταγραφών σε κάποιους μήνες αλλά, όπως θα δούμε και παρακάτω το φαινόμενο αυτό δεν αποτελεί τον κύριο λόγο για αυτή τη διαφορά.

Categories	Counts	Percentages	Success rate
apr	2932	6.48%	19.67%
aug	6247	13.81%	11.01%
dec	214	0.47%	46.72%
feb	2649	5.85%	16.64%
jan	1403	3.10%	10.12%
jul	6895	15.25%	9.09%
jun	5341	11.81%	10.22%
mar	477	1.05%	51.99%
may	13766	30.44%	6.71%
nov	3970	8.78%	10.15%
oct	738	1.63%	43.76%
sep	579	1.28%	46.45%

Πίνακας 11: Μετρικές εμφάνισης και ποσοστά επιτυχίας της μεταβλητής month

Όπως βλέπουμε στον πίνακα, υπάρχει μία μεγάλη ανισορροπία ως προς τα ποσοστά επιτυχίας μεταξύ των κατηγοριών της μεταβλητής month. Αυτό μπορεί να δικαιολογηθεί εν μέρει από την επίσης μεγάλη ανισορροπία που υπάρχει μεταξύ των ποσοστών εμφάνισης των κατηγοριών αυτών μιας και παρατηρούμε ότι οι μήνες με τις λιγότερες καταγραφές τείνουν να έχουν μεγαλύτερα ποσοστά επιτυχίας από τους μήνες με περισσότερες καταγραφές. Σε άλλες περιπτώσεις, αυτή η ανισορροπία ενδεχομένως να δημιουργούσε προβλήματα ως προς τα συμπεράσματα που θα μπορούσαμε να βγάλουμε για την μεταβλητή αυτή. Σε αυτή την περίπτωση όμως, υποθέτουμε ότι τα θετικά αποτελέσματα φέρουν μεγαλύτερη κατά μέσο όρο τιμή στη μεταβλητή duration από τα αρνητικά αποτελέσματα, δηλαδή οι θετικές τηλεφωνικές κλήσεις έχουν μεγαλύτερη διάρκεια με αποτέλεσμα να μειώνεται ο συνολικός αριθμός τηλεφωνικών κλήσεων που πραγματοποιήθηκαν μέσα στη μέρα ή και στον μήνα. Χωρίζοντας το σετ δεδομένων σε δύο υποσύνολα όπου το πρώτο υποσύνολο περιέχει μόνο τις θετικές καταγραφές ενώ το δεύτερο τις αρνητικές, μπορούμε να επαληθεύσουμε την υποθεσή μας συγκρίνοντας τον μέσο όρο και τον ενδιάμεσο των δύο υποσυνόλων.

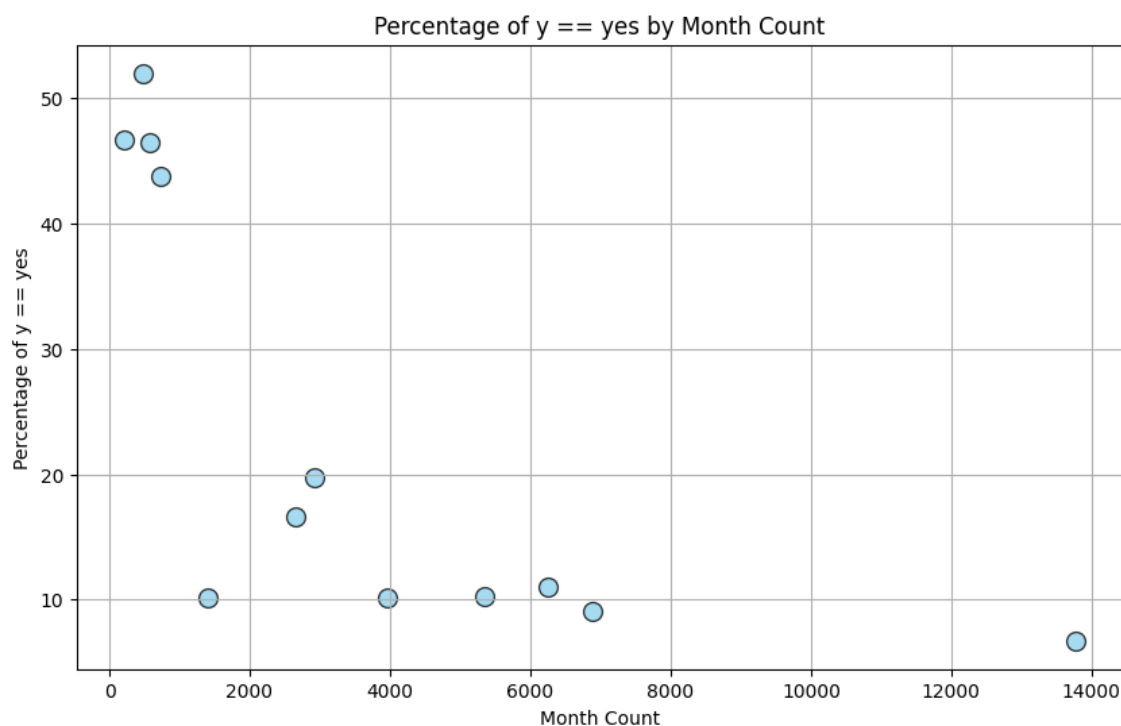
Categories	Duration mean	Duration median
All	258.16	180
Positive	537.29	426
Negative	221.18	164

Πίνακας 12: Μέσος όρος και ενδιάμεσος του duration ανάλογα με το αποτέλεσμα

Όπως βλέπουμε και στον πίνακα, είναι ξεκάθαρο το ότι οι τηλεφωνικές κλήσεις με θετικό αποτέλεσμα διαρκούν (σε δευτερόλεπτα) κατά κανόνα περισσότερο από αυτές με αρνητικό αποτέλεσμα. Χωρίζοντας το σετ δεδομένων σε 12 υποσύνολα (ένα για κάθε μήνα) και συγκρίνοντας το γενικό μέσο όρο και ενδιάμεσο της μεταβλητής duration για κάθε υποσύνολο, μπορούμε να διαπιστώσουμε ότι οι μετρήσεις θα είναι παρόμοιες μεταξύ τους, και σε κάποιες περιπτώσεις ακόμη και μεγαλύτερες στους μήνες με τις λιγότερες καταγραφές.

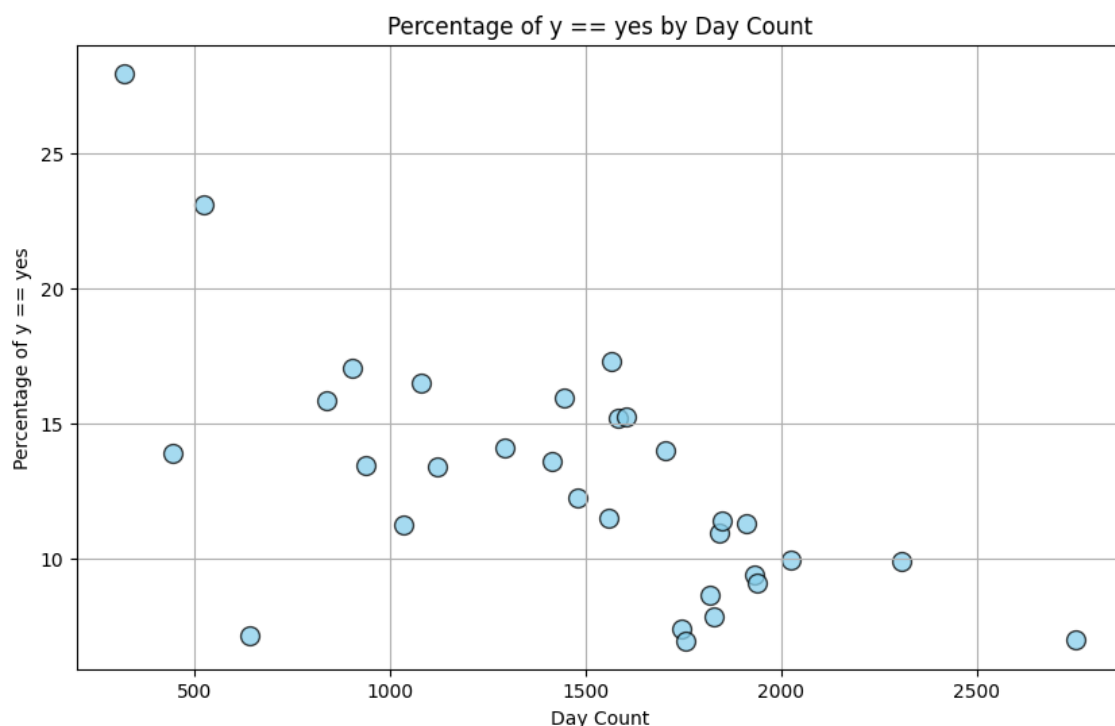
Categories	Count	Duration mean	Duration median
may	13766	260.92	190
jun	5341	243.08	164
jul	6895	267.97	180
aug	6247	232.40	156
oct	738	288.30	207
nov	3970	253.14	168
dec	214	329.32	267
jan	1403	268.22	194
feb	2649	248.29	171
mar	477	244.40	187
apr	2932	298.09	224
sep	579	292.25	224

Πίνακας 13: Μέσος όρος και ενδιάμεσος του duration ανάλογα με τον μήνα



Σχήμα 1: Month count plot: $y == \text{yes}$

Μετά από αυτές τις διαπιστώσεις που κάναμε για την μεταβλητή *month*, μπορούμε να συμπεράνουμε το ότι αυτή η αυξομείωση των θετικών αποτελεσμάτων ανάλογα με τον αριθμό καταγραφών της κάθε κατηγορίας δεν είναι τυχαία και είναι εμφανές ότι υπάρχει μία αριθμητική σχέση μεταξύ του αριθμού καταγραφών, της μεταβλητής *month*, και του τελικού αποτελέσματος. Τα ίδια γενικά συμπεράσματα μπορούμε βγάλουμε φυσικά και για την μεταβλητή *day* που σχετίζεται με τη μεταβλητή *month*. Τα τελικά αποτελέσματα του *day* ήταν πιο ομαλά, και αυτό ευθύνεται στο γεγονός ότι η κάθε μέρα που εξετάζουμε εμφανίζεται μέσα σε σχεδόν όλους τους μήνες, πράγμα που ενισχύει την σημασία της μεταβλητής *month* στο τελικό αποτέλεσμα αλλά και επιβεβαιώνει για ακόμη μία φορά την υπόθεση σχετικά με το *duration*.



Σχήμα 2: Day count plot: y == yes

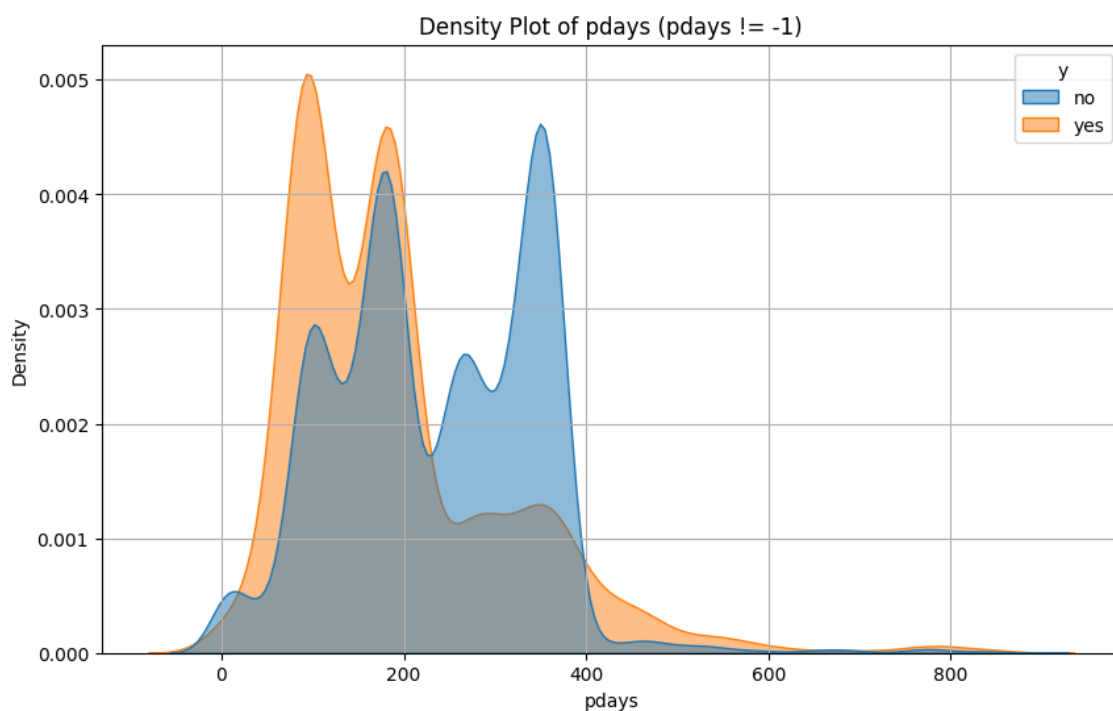
Η μεταβλητή `pdays`, όπως αναφέρθηκε προηγουμένως, περιέχει τον αριθμό των ημερών που έχουν περάσει από την τελευταία τηλεφωνική κλήση που πραγματοποιήθηκε στον πελάτη από προηγούμενη εκστρατεία μάρκετινγκ, είναι συνεπώς μία αριθμητική μεταβλητή. Ωστόσο, οι καταγραφές οι οποίες δεν περιέχουν προηγούμενη τηλεφωνική κλήση, περιγράφονται με την τιμή -1, η οποία περιγράφει ένα διαφορετικό γεγονός από τις υπόλοιπες τιμές της μεταβλητής. Αυτό ενδεχομένως μπορεί να μας δημιουργεί πρόβλημα ως προς την ακεραιότητα της πληροφορίας που θα εξορύξουμε από αυτή τη μεταβλητή. Για αυτό το λόγο θα πρέπει να κάνουμε μία περαιτέρω ανάλυση και επεξεργασία στη συγκεκριμένη μεταβλητή.

Categories	Counts	Percentages	Success rate
-1	36954	81.73%	9.15%
not -1	8257	18.26%	23.07%

Πίνακας 14: Μετρικές εμφάνισης και ποσοστό επιτυχίας για το `pdays`

Βλέποντας και στο πίνακάκι, παρατηρούμε ότι η μεγαλύτερη διαφορά που υπάρχει στη μετα-

βλητή είναι μεταξύ των πελατών που δεν έχουν λάβει στο παρελθόν κάποια κλήση σε σύγκριση με τους πελάτες που έχουν λάβει. Οι πελάτες που καλούνται πρώτη φορά είναι με συντηρητική πλειοψηφία περισσότεροι μιας και αγγίζουν το 82% των κλήσεων που πραγματοποιήθηκαν, αλλά και έχουν αρκετά μικρότερο ποσοστό επιτυχίας συγκριτικά με τους υπόλοιπους πελάτες.

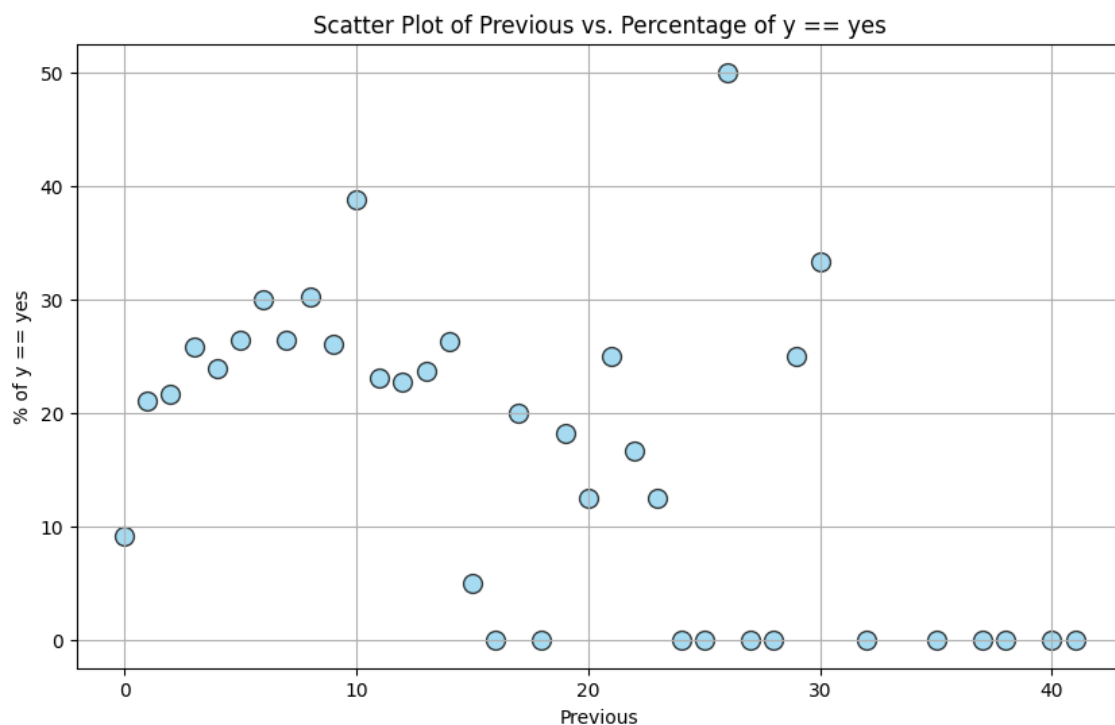


Σχήμα 3: Density plot of pdays (without -1)

Στο παραπάνω σχήμα έχουμε απομονώσει τους πελάτες στους οποίους έχει ξαναγίνει τηλεφωνική κλήση στο παρελθόν, δηλαδή δεν έχουμε συμπεριλάβει τους πελάτες που έχουν την τιμή -1 στη μεταβλητή pdays. Παρόλο που η κατανομή τους δεν είναι πολύ ξεκάθαρη, μπορούμε να βγάλουμε κάποια συμπεράσματα. Τα περισσότερα τηλεφωνήματα που πραγματοποιήθηκαν έγιναν σε πελάτες οι οποίοι τηλεφωνήθηκαν πριν από 90-390 ημέρες. Επίσης, παρατηρούμε δύο απότομες κορυφώσεις θετικών αποτελεσμάτων σε συγκεκριμένα σημεία του διαγράμματος. Για να αποφύγουμε την υπερφόρτωση του μοντέλου μετατρέψαμε τη συγκεκριμένη μεταβλητή σε μια κατηγορηματική μεταβλητή.

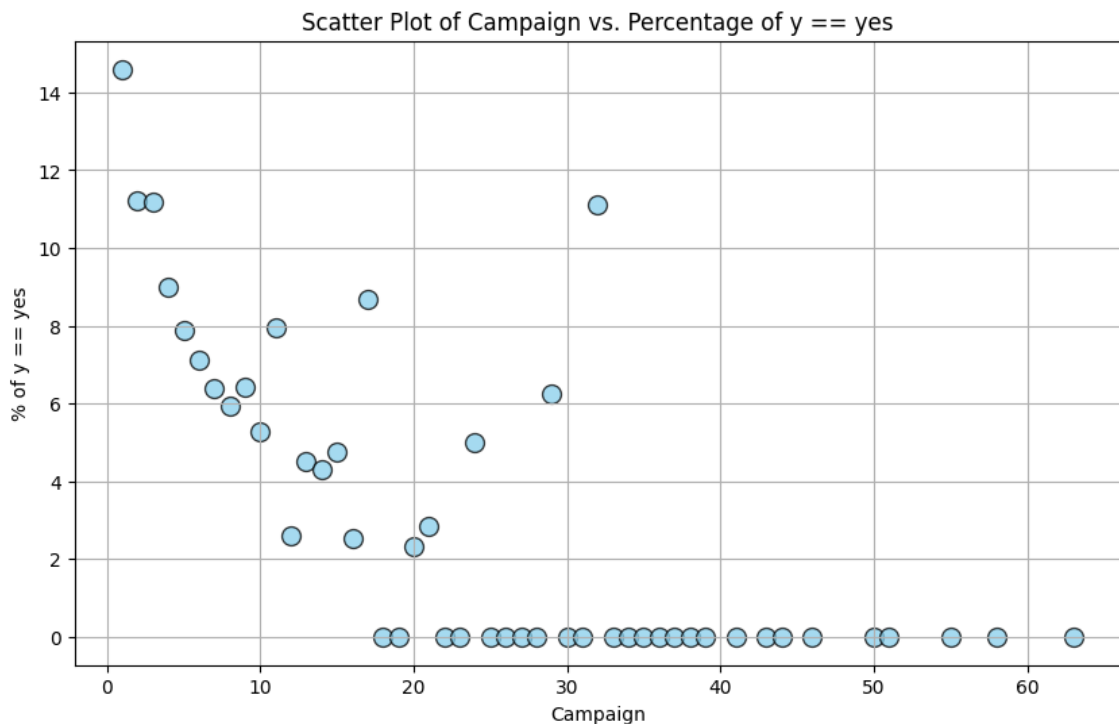
Οι επόμενες μεταβλητές που θα πρέπει να αναλύσουμε είναι το previous και το campaign. Για την καλύτερη ανάλυση αλλά και εκπαίδευση των αλγορίθμων έγινε αφαίρεση κάποιων outliers

από τη μεταβλητή previous η οποία βοήθησε στην βελτίωση της γενίκευσης του αλγορίθμου και στην αποφυγή της υπερφόρτωσής του.



Σχήμα 4: Previous plot: y == yes

Η μεταβλητή previous στις περιπτώσεις που έχει ως τιμή το μηδέν, εκφράζει την ίδια ομάδα πελατών που εκφάζει και η μεταβλητή pdays όταν φέρει την τιμή -1, δηλαδή τους πελάτες στους οποίους δεν έχει γίνει προηγούμενη τηλεφωνική κλήση πριν της καταγεγραμμένης. Αυτό σημαίνει ότι σε αυτές τις περιπτώσεις έχει το ίδιο ποσοστό επιτυχίας με το pdays. Στις υπόλοιπες τιμές, όπως φαίνεται και στο σχήμα, υπάρχει μία αύξηση του ποσοστού επιτυχίας από την τιμή 1 μέχρι την τιμή 10, ενώ στη συνέχεια υπάρχει μείωση του ποσοστού αυτού.



Σχήμα 5: Campaign plot: $y == \text{yes}$

Στη μεταβλητή *campaign* φαίνεται να υπάρχει μία πιο ξεκάθαρη εικόνα, μιας και το ποσοστό επιτυχίας μειώνεται όσο οι τιμές του *campaign* αυξάνονται, πράγμα που μας δείχνει ότι υπάρχει μια αντιστρόφως ανάλογη σχέση με την τελική μεταβλητή *y*.

Μια άλλη μεταβλητή που μας απασχολεί είναι το *outcome*, το οποίο περιγράφει το αν η προηγούμενη τηλεφωνική κλήση που έγινε στον πελάτη ήταν επιτυχής. Η μεταβλητή αυτή περιέχει τις κατηγορίες *success* και *failure*, οι οποίες συμβολίζουν την επιτυχία ή αποτυχία της προηγούμενης κλήσης αντίστοιχα, και τις κατηγορίες *other* και *unknown*, για τις οποίες δεν υπάρχει κάποια επεξήγηση ως προς τον συμβολισμό τους.

Όπως παρατηρούμε και στον πίνακα, η κατηγορία με το μεγαλύτερο ποσοστό θετικών αποτελεσμάτων είναι το *success*, ενώ με το μικρότερο είναι το *unknown*. Ταυτόχρονα, το *unknown* έχει με συντριπτική πλειοψηφία τις περισσότερες καταχωρήσεις της μεταβλητής. Για αυτό το λόγο θα πρέπει να την αναλύσουμε καλύτερα για να μπορέσουμε να την ερμηνεύσουμε. Λόγω της πολλαπλής της εμφάνισης και σε άλλες μεταβλητές, το πρώτο συμπέρασμα που θα μπορούσαμε να βγάλουμε για την κατηγορία *unknown* είναι το ότι πρόκειται για μία κατηγορία η

Categories	Count (positive and negative entries)	Success rate
unknown	36959	9.16%
failure	4901	12.61%
other	1840	16.68%
success	1511	64.73%

Πίνακας 15: Αριθμός εμφάνισης και ποσοστά επιτυχίας του poutcome

οποία υποδουλώνει την έλλειψη πληροφορίας, όπως ισχύει για τις υπόλοιπες μεταβλητές. Αντιθέτως, υποθέσαμε ότι η συγκεκριμένη κατηγορία είναι αποκλειστικά συνδεδεμένη με τους πελάτες στους οποίους δεν έχει πραγματοποιηθεί καμία κλήση πριν την καταγεγραμμένη. Για να επαληθεύσουμε αυτή την υπόθεση χρησιμοποιήσαμε τη μεταβλητή pdays για να φτιάξουμε δύο υποσύνολα από το σετ δεδομένων. Το πρώτο υποσύνολο περιέχει μόνο τις καταγραφές πελατών στους οποίους δεν έχει γίνει προηγούμενη τηλεφωνική κλήση, ενώ το δεύτερο υποσύνολο περιέχει μόνο τις καταγραφές πελατών στους οποίους έχει γίνει προηγούμενη κλήση. Στη συνέχεια, μετρήσαμε τον αριθμό καταχωρήσεων για την κάθε κατηγορία της μεταβλητής στα συγκεκριμένα υποσύνολα αλλά και στο συνολικό σετ δεδομένων.

Categories	All	Pdays = -1	Pdays != -1
unknown	36959	36954	5
failure	4901	0	4901
other	1840	0	1840
success	1511	0	1511

Πίνακας 16: Αριθμός εμφάνισης unknown ανάλογα με την τιμή του pdays.

Όπως βλέπουμε και στο πίνακάκι, η κατηγορία unknown εμφανίζεται μονάχα στην ομάδα πελατών που περιγράψαμε προηγουμένως, με εξαίρεση 5 καταχωρήσεις οι οποίες λόγω του μικρού αριθμητικού μεγέθους τους μπορούν να θεωρηθούν αμελητέες. Συνεπώς, η υπόθεση που κάναμε ισχύει, το οποίο σημαίνει ότι θα θεωρήσουμε την κατηγορία unknown ως μία ξεχωριστή κατηγορία και όχι ως την έλλειψη πληροφορίας. Για λόγους ακριβείας, έγινε αφαίρεση αυτών των 5 καταχωρήσεων που δεν συμβαδίζουν με τα υπόλοιπα δεδομένα.

Η επόμενη μεταβλητή που θα πρέπει να κοιτάξουμε είναι η μεταβλητή contact, η οποία περιγράφει το είδος συσκευής που χρησιμοποίησε ο πελάτης στην τηλεφωνική κλήση, δηλαδή

το αν χρησιμοποίησε σταθερό ή κινητό τηλέφωνο. Εκτός από τις δύο βασικές κατηγορίες (telephone/cellular) που περιέχει η μεταβλητή, υπάρχει και η κατηγορία unknown η οποία συμβολίζει την έλλειψη πληροφορίας. Όπως έχουμε αναφέρει, ο τελικός στόχος αυτής της εργασίας είναι η ανάπτυξη ενός μοντέλου πρόβλεψης και κατ' επέκταση μιας εφαρμογής που θα μπορεί να ανταποκρίνεται σε ρεαλιστικές περιπτώσεις πρόβλεψης. Σε μία τέτοια περίπτωση φυσικά, η μη επίγνωση του είδους συσκευής που χρησιμοποιεί ο πελάτης είναι αδύνατη μιας και χωρίς αυτή την πληροφορία δεν θα μπορούσε να πραγματοποιηθεί η τηλεφωνική κλήση εξαρχής. Για αυτό το λόγο, καλό θα ήταν να αφαιρέσουμε τις καταχωρήσεις που περιέχουν την κατηγορία unknown στη μεταβλητή contact. Αυτή η διαδικασία θα πρέπει να γίνει με προσοχή επειδή υπάρχει κίνδυνος να επηρεαστεί η ακεραιότητα του σετ δεδομένων.

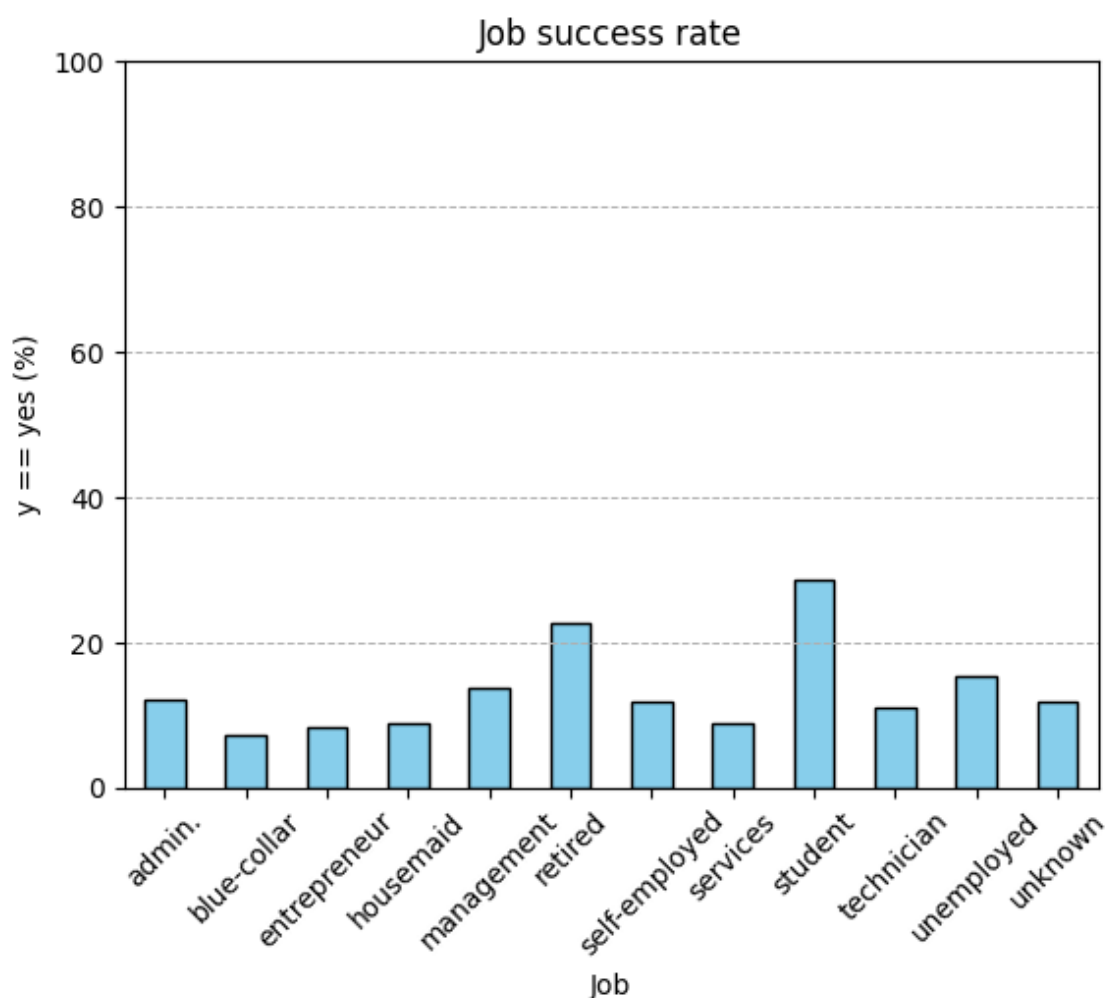
Categories	Count (positive and negative entries)	Success rate
cellular	29285	14.92%
unknown	13020	4.07%
telephone	2906	13.42%

Πίνακας 17: Ποσοστά επιτυχίας της μεταβλητής contact

Όπως μπορούμε να παρατηρήσουμε στον πίνακα, οι κατηγορίες cellular και telephone έχουν παρόμοια ποσοστά επιτυχίας (13-15%). Μια λογική πρόβλεψη που θα μπορούσαμε να κάνουμε για την κατηγορία unknown είναι το να έχει και αυτή περίπου το ίδιο ποσοστό επιτυχίας με τις άλλες δύο κατηγορίες. Αντιθέτως, στην πραγματικότητα το ποσοστό που έχει είναι 4.07%, το οποίο διαφέρει αρκετά με τις άλλες δύο μεταβλητές. Χωρίζοντας το σετ δεδομένων σε δύο υποσύνολα τα οποία συμβολίζουν το σετ δεδομένων πριν και μετά την αφαίρεση της κατηγορίας unknown, κάναμε συγκρίσεις με άλλες μεταβλητές για να ελέγξουμε την επιρροή που έχει αυτή η κατηγορία στην ακεραιότητα του σετ δεδομένων. Τα συμπεράσματα που βγάλαμε για αυτή τη κατηγορία είναι το ότι η συντηρητική πλειοψηφία των πελατών (99.46%) που ανήκουν σε αυτή την κατηγορία δεν είχαν δεχτεί κάποια προηγούμενη τηλεφωνική κλήση στο παρελθόν. Επίσης, οι τηλεφωνικές κλήσεις που πραγματοποιήθηκαν σε αυτά τα άτομα έγιναν με συντηρητική πλειοψηφία στους μήνες Μάιο με ποσοστό εμφάνισης 61.25%, και Ιούνιο με ποσοστό 34.80%, πράγμα που επηρέασε με αρνητικό τρόπο τα αποτελέσματα σε αυτούς τους μήνες.

Μια ακόμη μεταβλητή που θα πρέπει να αναλύσουμε είναι η μεταβλητή job η οποία περιγράφει

το είδος της επαγγελματικής κατάστασης του πελάτη. Μία από τις κατηγορίες που περιέχει η συγκεκριμένη μεταβλητή είναι η κατηγορία unknown η οποία συμβολίζει την έλλειψη της πληροφορίας. Στη συγκεκριμένη περίπτωση αυτή η κατηγορία δεν μας δημιουργεί πρόβλημα μιας και μοιράζεται πρακτικά το ίδιο ποσοστό επιτυχίας με το γενικό ποσοστό επιτυχίας του σετ δεδομένων (unknown: 11.8%, y: 11.7%). Για αυτό το λόγο δεν είναι απαραίτητη η αφαίρεση του unknown από τη μεταβλητή job.



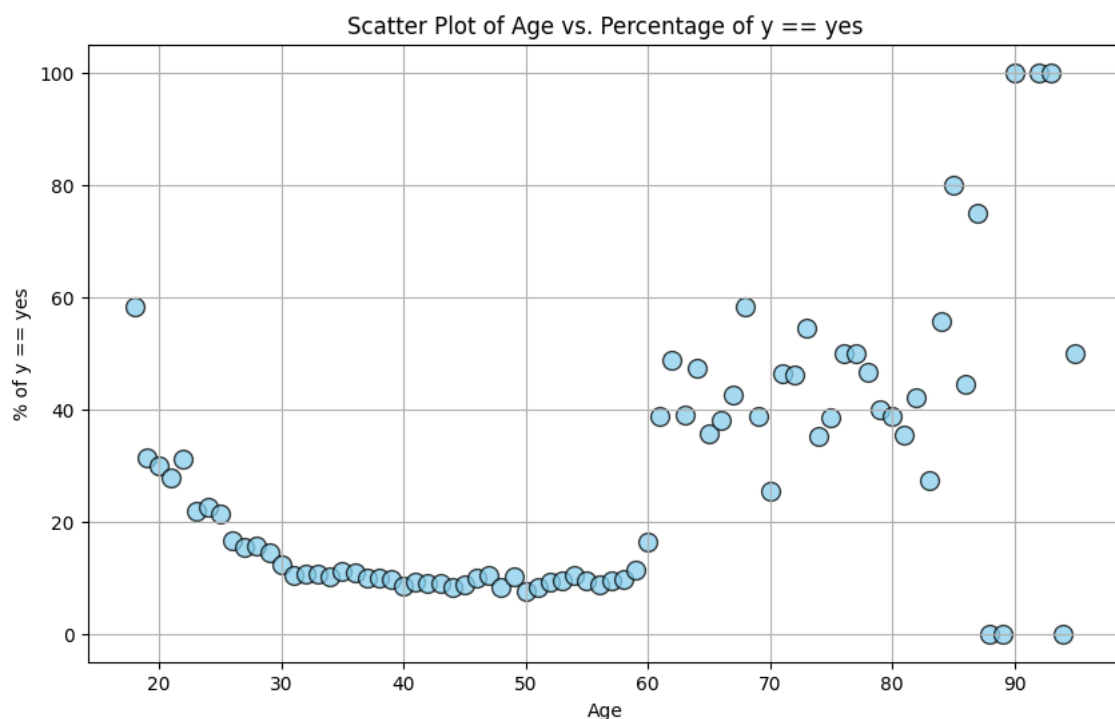
Σχήμα 6: Job plot: y == yes

Όπως παρατηρούμε στο σχήμα, τα ποσοστά επιτυχίας των κατηγοριών της μεταβλητής job είναι αρκετά ομοιόμορφα. Οι μόνες κατηγορίες που ξεχωρίζουμε στο σχήμα είναι οι retired

και student οι οποίες είναι οι μόνες κατηγορίες της μεταβλητής όπου ξεπερνάνε το 20% επιτυχίας. Στη συνέχεια εξετάσαμε τα ποσοστά εμφάνισης του job σε σύγκριση με τους μήνες οι οποίοι εμφανίζουν τα μεγαλύτερα ποσοστά επιτυχίας. Αυτό το κάναμε για να εξετάσουμε το ενδεχόμενο του να επηρεάζουν αυτές οι κατηγορίες (retired, student) τους μήνες με τη μεγαλύτερη επιτυχία. Ακολουθώντας αυτή τη διαδικασία παρατηρούμε ότι υπάρχει μία μικρή αύξηση του ποσοστού εμφάνισης αυτών των κατηγοριών στους μήνες που αναφέραμε. Τέλος, συγκρίναμε τα ποσοστά επιτυχίας του job σε αυτούς τους μήνες ξεχωριστά και διαπιστώσαμε ότι παρόλο που τα ποσοστά εμφάνισης των κατηγοριών retired και student είναι ελαφρώς ανεβασμένα, τα ποσοστά επιτυχίας που εμφανίζουν, σε σύγκριση με τις άλλες κατηγορίες, δεν είναι αρκετά ψηλά για να βγάλουμε το συμπέρασμα ότι επηρεάζουν τα ποσοστά επιτυχίας στους μήνες όπου εξετάσαμε.

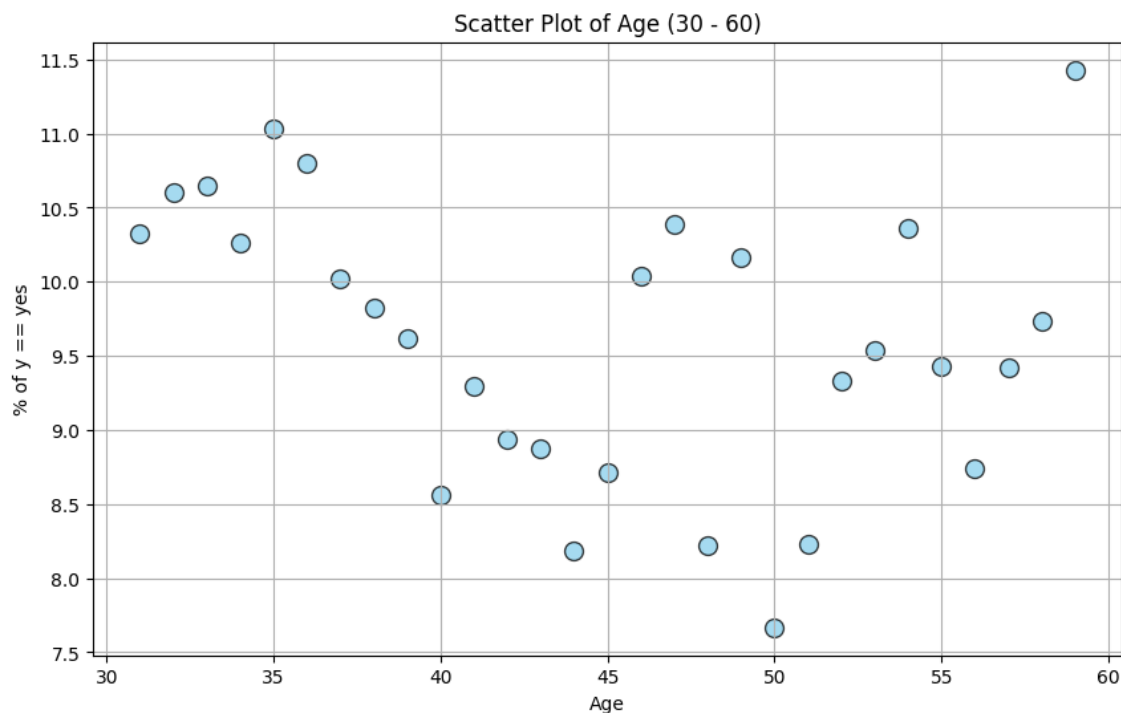
Οι επόμενες μεταβλητές που μελετήσαμε είναι οι μεταβλητές marital και education. Για την μεταβλητή marital, η κατηγορία single εμφανίζει το μεγαλύτερο ποσοστό επιτυχίας γύρω κοντά στο 15%, ενώ οι κατηγορίες married και divorced φέρουν ποσοστά επιτυχίας κοντά στο 10% και 12% αντίστοιχα.

Στη συνέχεια κάναμε ανάλυση στις δυαδικές μεταβλητές μεταβλητές default, loan και housing. Παρατηρήσαμε και στις τρεις μεταβλητές ότι η κατηγορία no έχει περίπου το διπλάσιο ποσοστό επιτυχίας από την κατηγορία yes. Για τις μεταβλητές default και loan τα ποσοστά επιτυχίας του no είναι 11.5-12.5% και του yes είναι 6.4-6.7%, ενώ για τη μεταβλητή housing τα ποσοστά είναι 16.7% και 7.7% αντίστοιχα.



Σχήμα 7: Age plot: y == yes

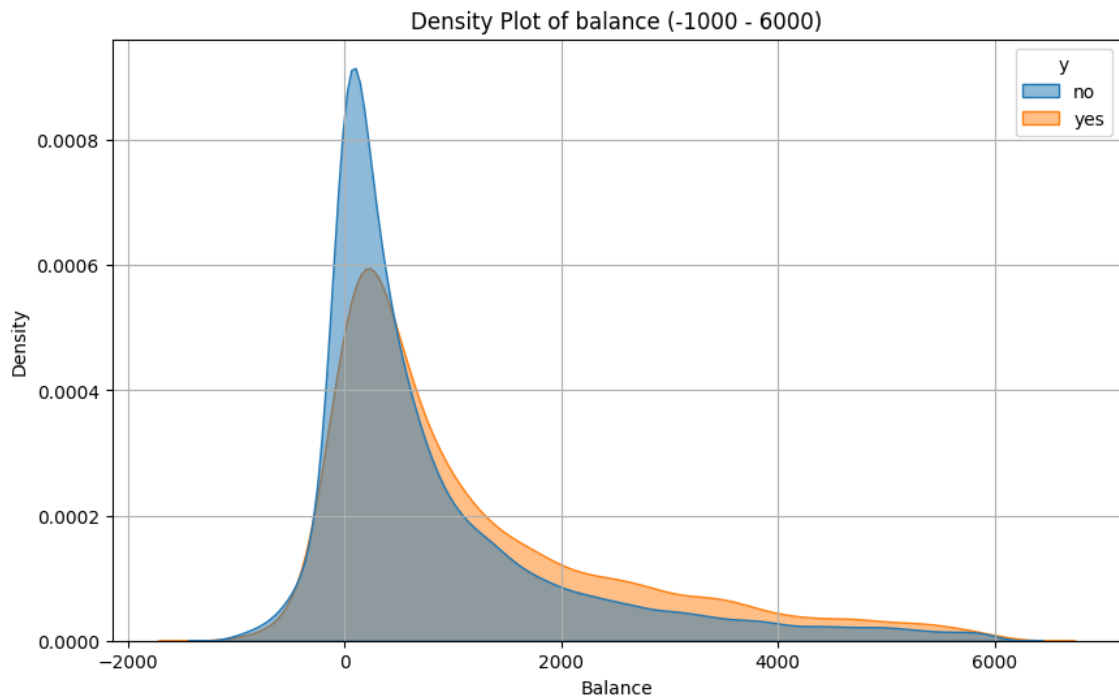
Η επόμενη μεταβλητή η οποία αναλύσαμε είναι η αριθμητική μεταβλητή age. Όπως παρατηρούμε στο σχήμα παραπάνω σχήμα, μπορούμε να δούμε ότι στις ηλικίες κάτω των 20 και άνω των 60 ετών η κατανομή στα ποσοστά επιτυχίας είναι διάσπαρτη. Αυτό ενδεχομένως να ευθύνεται στο ότι οι ομάδες αυτές έχουν μικρό αριθμό καταχωρήσεων. Επίσης παρατηρούμε ότι τα ποσοστά επιτυχίας μειώνονται σταδιακά από τις ηλικίες 18 μέχρι 30, ενώ από τις ηλικίες 60 και άνω τα ποσοστά επιτυχίας αυξάνονται με απότομο και διάσπαρτο τρόπο. Στις ηλικίες 30 μέχρι 60 ετών η κατανομή είναι πιο σταθερή.



Σχήμα 8: Age plot (30-60): $y == \text{yes}$

Απομονώνοντας τις ηλικίες 30 με 60 ετών, μπορούμε να δούμε στο σχήμα, ότι υπάρχει μια μικρή αύξηση των ποσοστών επιτυχίας από τις ηλικίες 30 μέχρι 35, και στη συνέχεια, από τις ηλικίες 35 μέχρι 45 υπάρχει πάλι μείωση αυτών των ποσοστών. Τέλος, από τις ηλικίες 45 μέχρι 60 παρατηρούμε μία πιο διάσπαρτη αύξηση αυτών των ποσοστών. Για να αποφύγουμε την υπερφόρτωση του μοντέλου θα χρειαστεί να κάνουμε μετατροπή της μεταβλητής από αριθμητική σε κατηγορηματική χωρίζοντας τις τιμές της μεταβλητής σε ηλικιακές ομάδες.

Η τελευταία μεταβλητή που αναλύσαμε είναι η μεταβλητή *balance* όπου μας δίνει τον μέσο όρο χρημάτων (σε ευρώ) ετησίως που διαθέτει ο πελάτης στον τραπεζικό λογαριασμό του. Βλέποντας το σχήμα παρακάτω, παρατηρούμε ότι η πλειοψηφία των πελατών έχει *balance* κάτω των 1000 ευρώ. Επίσης παρατηρούμε ότι οι πελάτες οι οποίοι έχουν *balance* άνω των 500 ευρώ έχουν μεγαλύτερο ποσοστό επιτυχίας από τους υπόλοιπους.



Σχήμα 9: Balance density plot

4.3 Προετοιμασία δεδομένων

Για να πραγματοποιηθεί μια επιτυχημένη εκπαίδευση και υλοποίηση ενός μοντέλου μηχανικής μάθησης, το σετ δεδομένων θα πρέπει να έχει μία συγκεκριμένη μορφή για να μπορέσει να διαβαστεί από τον εκάστοτε αλγόριθμο μηχανικής μάθησης. Αυτή τη μορφή δεδομένων μπορούμε να την πετύχουμε εφαρμόζοντας κάποιες μεθόδους προετοιμασίας δεδομένων. Προτού γίνει αυτό, εφαρμόσαμε τεχνικές καθαρισμού δεδομένων όπως ο έλεγχος για τυχόν ελλειπούσες τιμές από τις μεταβλητές του σετ δεδομένων, έλεγχος για επαναλαμβανόμενες σειρές δεδομένων, αλλά και έλεγχος για δομικά και λογικά σφάλματα του σετ δεδομένων.

4.3.1 Αφαίρεση μεταβλητών

Για να ξεκινήσουμε την προετοιμασία και μετατροπή των μεταβλητών θα πρέπει να επιλέξουμε τις μεταβλητές τις οποίες θα χρησιμοποιήσουμε για την εκπαίδευση του μοντέλου. Αυτή η επιλογή θα πρέπει να γίνει κρατώντας υπόψιν ότι σκοπός είναι η υλοποίηση ενός ρεαλιστικού μοντέλου πρόβλεψης. Για αυτό το λόγο, θα πρέπει να αφαιρέσουμε την μεταβλητή *duration*, η οποία μας δίνει την διάρκεια (σε δευτερόλεπτα) των τηλεφωνικών κλήσεων που πραγματο-

ποιούνται στους πελάτες. Παρόλο που ως μεταβλητή έχει πολύ μεγάλη στατιστική σημασία ως προς το αποτέλεσμα, όπως είδαμε προηγουμένως, θα πρέπει να την αφαιρέσουμε από το σετ δεδομένων μιας και είναι αδύνατον να γνωρίζουμε την διάρκεια μιας τηλεφωνικής κλήσης προτού αυτή πραγματοποιηθεί.

4.3.2 Δυαδικές μεταβλητές

Οι δυαδικές μεταβλητές που υπάρχουν στο σετ δεδομένων είναι το default, το housing, το loan, και η τελική μεταβλητή y. Οι μεταβλητές αυτές μοιράζονται τις ίδιες τιμές (yes/no). Η κωδικοποίηση που κάναμε σε αυτές τις μεταβλητές είναι το να μετατρέψουμε τις τιμές τους σε αριθμητικές, το οποίο έγινε μετατρέποντας το yes/no σε 1 και 0.

4.3.3 One-hot encoding

Το one-hot encoding είναι μία μέθοδος προετοιμασίας κατηγορηματικών δεδομένων στην οποία μετατρέπουμε τις κατηγορίες μίας κατηγορηματικής μεταβλητής σε ξεχωριστές μεταβλητές οι οποίες παίρνουν τις τιμές 0 ή 1 ανάλογα με το αν η συγκεκριμένη κατηγορία εμφανίζεται στην συγκεκριμένη καταχώρηση. Αυτή η μέθοδος εφαρμόστηκε για τις μεταβλητές marital και contact απ' τις οποίες δημιουργήθηκαν 5 μεταβλητές (single, married, divorced, cellular, telephone).

4.3.4 Binning

Το binning είναι μία μέθοδος προετοιμασίας αριθμητικών δεδομένων στην οποία μετατρέπουμε τις αριθμητικές μεταβλητές που επιλέξαμε σε κατηγορηματικές. Οι μεταβλητές στις οποίες θα εφαρμόσουμε αυτή τη μέθοδο είναι το age και το pdays. Παρατηρώντας την κατανομή στα ποσοστά επιτυχίας του age, έγινε διαχωρισμός των τιμών του στις ομάδες (18-25), (25-30), (30-35), (35-40), (40-45), (45-50), (50-55), (55-60), και (60+). Για τη μεταβλητή pdays, παρατηρώντας την κατανομή των ποσοστών επιτυχίας, έγινε ομαδοποίηση των τιμών τις στις κατηγορίες (-1), (0-60), (60-230), και (230+).

4.3.5 Label encoding

Το label encoding είναι μία μέθοδος προετοιμασίας κατηγορηματικών δεδομένων στην οποία μετατρέπουμε τις κατηγορίες μίας κατηγορηματικής μεταβλητής σε ακέραιους αριθμούς ξεκινώντας από το μηδέν. Αυτή την μέθοδο την εφαρμόσαμε σε κάποια τακτικά κατηγορηματικά δεδομένα, δηλαδή σε μεταβλητές οι οποίες φέρουν μία σειρά ή ιεραρχία ως προς τη δομή τους.

Οι μεταβλητές αυτές είναι το education, το age και το pdays (αφού εφαρμοστεί binning στο age και στο pdays).

4.3.6 Μεταβλητές day και month

Στην ανάλυση που κάναμε στις μεταβλητές day και month βγάλαμε το συμπέρασμα ότι ο αριθμός συχνότητας καταχωρήσεων μίας ημέρας ή ενός μήνα, έχει αριθμητική σχέση με το ποσοστό επιτυχίας το οποίο φέρει. Για αυτό το λόγο θα μετατρέψουμε αυτές τις δύο κατηγορηματικές μεταβλητές σε αριθμητικές μεταβλητές, όπου η τιμή που θα περιέχει η μεταβλητή θα ισούται με τον αριθμό καταχωρήσεων του αντίστοιχου μήνα ή ημέρας.

4.3.7 Target mean encoding

Για τις μεταβλητές job και poutcome εφαρμόσαμε τη μέθοδο target mean encoding στην οποία αντικαθιστούμε την κάθε κατηγορία της μεταβλητής με το ποσοστό επιτυχίας το οποίο φέρει η αντίστοιχη κατηγορία. Σε αυτή τη μεθοδολογία υπάρχει ο κίνδυνος υπερφόρτωσης του μοντέλου από κατηγορίες οι οποίες, λόγω του μικρού αριθμού καταχωρήσεων που έχουν, υπάρχει περίπτωση να εκφράζουν ακραία ποσοστά επιτυχίας ή και αποτυχίας. Για να το αποφύγουμε αυτό εφαρμόσαμε μία μέθοδο απάλυνσης αυτών των ποσοστών σε αυτές τις κατηγορίες προσθέτοντας μία παράμετρο βαρύτητας.

$$\text{smooth} = (\text{counts} * \text{means} + \text{weight} * \text{mean}) / (\text{counts} + \text{weight})$$

4.3.8 Κανονικοποίηση

Η τελευταία μέθοδος που θα εφαρμόσουμε για την προετοιμασία του σετ δεδομένων είναι η κανονικοποίηση των τιμών που περιέχουν οι μεταβλητές. Αυτή η μέθοδος δεν είναι υποχρεωτική για όλους τους αλγορίθμους μηχανικής μάθησης, αλλά για την σωστή εκπαίδευση αλγορίθμων όπως τα SupportVectorMachines που θα δούμε παρακάτω είναι απαραίτητη. Για να το κάνουμε αυτό, χρησιμοποιήσαμε την συνάρτηση MinMaxScaler της βιβλιοθήκης sklearn. Η κανονικοποίηση έγινε θέτοντας ως min=0 και max=1, δηλαδή όλες οι τιμές των μεταβλητών ανήκουν στο φάσμα μεταξύ του 0 και του 1.

4.3.9 Διόρθωση ανισορροπίας

Όπως αναφέραμε προηγουμένως, το συγκεκριμένο σετ δεδομένων έχει τεράστια ανισορροπία καταχωρήσεων μεταξύ των δύο κλάσεων (yes/no). Οι θετικές καταχωρήσεις κατέχουν το 11.7% των καταχωρήσεων. Για να αποφύγουμε την υπερφόρτωση του μοντέλου από τις αρνητικές καταχωρήσεις, εφαρμόσαμε μέθοδο sampling για να διορθώσουμε την ανισορροπία των δεδομένων. Συγκεκριμένα χρησιμοποιήσαμε τη μέθοδο random undersampling με την οποία αφαιρέσαμε τυχαίες αρνητικές καταχωρήσεις από το σετ δεδομένων μέχρι τα ποσοστά των δύο κλάσεων να είναι της μορφής 1:1. Μετά την εφαρμογή της μεθόδου, ο συνολικός αριθμός των καταχωρήσεων του σετ δεδομένων κατέβηκε στις 9052 καταχωρήσεις.

4.3.10 Επιλογή αλγόριθμων μηχανικής μάθησης

Για την ανάπτυξη του μοντέλου πρόβλεψης εκπαιδεύσαμε μία ομάδα αλγόριθμων μηχανικής μάθησης εκ των οποίων οι δύο είναι συμβατικοί ενώ οι υπόλοιποι ακολουθούν μεθόδους εκπαίδευση συνδυαστικών μοντέλων πρόβλεψης (ensemble methods). Οι δύο συμβατικοί αλγόριθμοι που χρησιμοποιήσαμε είναι οι Decision Trees (DT) και Support Vector Machines (SVM). Οι αλγόριθμοι που χρησιμοποιούν τη συνδυαστική μέθοδο boosting και τους επιλέξαμε για την εκπαίδευση είναι οι AdaBoost (ADA), GradientBoosting (GB) και Extreme Gradient Boosting (XGB), ενώ για τη συνδυαστική μέθοδο bagging εκπαιδεύσαμε τον αλγόριθμο Random Forest (RF). Οι αλγόριθμοι που χρησιμοποιήσαμε για την εκπαίδευση και εξέταση των επιδόσεων πάρθηκαν από τη βιβλιοθήκη της Python sklearn, με εξαίρεση τον αλγόριθμο Extreme Gradient Boosting (XGB) ο οποίος πάρθηκε από τη βιβλιοθήκη της Python xgboost.

4.3.11 K-fold cross validation

Για την εκπαίδευση και σύγκριση των αλγόριθμων μηχανικής μάθησης χρησιμοποιήσαμε την μέθοδο k-fold cross validation η οποία μας επιτρέπει να εξετάσουμε την επίδοση αλλά και την ικανότητα γενίκευσης του κάθε αλγόριθμου με σκοπό την αποφυγή της υπερφόρτωσης των αλγόριθμων. Αυτό το πετυχαίνουμε χωρίζοντας το σετ δεδομένων σε $k=10$ ομάδες. Στη συνέχεια πραγματοποιείται η εκπαίδευση του αλγορίθμου χρησιμοποιώντας τις $k-1$ ομάδες δεδομένων και χρησιμοποιώντας την υπολοιπόμενη ομάδα για την εξέταση των επιδόσεων. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να χρησιμοποιηθούν όλες οι ομάδες ως δεδομένα εξέτασης του αλγόριθμου.

4.3.12 Παραμετροποίηση του μοντέλου

Μετά την εξέταση και σύγκριση των αποτελεσμάτων των αλγόριθμων μηχανικής μάθησης έγινε η επιλογή και παραμετροποίηση του βέλτιστου αλγόριθμου. Για την τελική εκπαίδευση και εξέταση του βέλτιστου αλγόριθμου, πάρθηκε το σετ δεδομένων στη μορφή που είχε πριν την διόρθωση της ανισορροπίας των κλάσεων και έγινε ο διαχωρισμός του σετ δεδομένων σε δεδομένα εκπαίδευσης τα οποία χρησιμοποιήσαμε για την εκπαίδευση αυτών των αλγόριθμων, και σε δεδομένα τεστ για την εξέταση των επιδόσεων του αλγόριθμου σε δεδομένα στα οποία δεν έχει εκπαιδευτεί. Εφαρμόσαμε διαχωρισμό 20/80 όπου το σετ με το 80% των δεδομένων χρησιμοποιήθηκε για την εκπαίδευση. Ο διαχωρισμός αυτός πραγματοποιήθηκε με τη συνάρτηση `train_test_split`, της βιβλιοθήκης `sklearn`, προσθέτοντας την παράμετρο `stratify=y` ώστε να υπάρχει το ίδιο γενικό ποσοστό επιτυχίας και στα δύο σετ δεδομένων. Αφού έγινε αυτό, χρησιμοποιώντας την συνάρτηση `RandomUnderSampler` της βιβλιοθήκης `imblearn`, διορθώσαμε την ανισορροπία των ποσοστών αποτελεσμάτων (από 12/88 σε 50/50) του σετ δεδομένων μόνο για τα δεδομένα εκπαίδευσης. Στα δεδομένα τεστ, δεν εφαρμόστηκε μέθοδος διόρθωσης της ανισορροπίας των κλάσεων διότι θέλαμε να εξετάσουμε τις επιδόσεις του μοντέλου πρόβλεψης σε μία πιο ρεαλιστική κατανομή δεδομένων.

5 Αποτελέσματα

Σε αυτό το κεφάλαιο θα εξετάσουμε τα αποτελέσματα τα οποία βγάλαμε μετά την εκπαίδευση των αλγόριθμων μηχανικής μάθησης και θα τα συγκρίνουμε μεταξύ τους με σκοπό την επιλογή και παραμετροποίηση του βέλτιστου αλγόριθμου ο οποίος θα χρησιμοποιηθεί για την υλοποίηση του συστήματος υποστήριξης αποφάσεων.

5.1 Μετρικές εκτίμησης

Παρακάτω γίνεται η περιγραφή των μετρικών εκτίμησης των επιδόσεων του μοντέλου.

- True positive (TP): Επιτυχημένη θετική πρόβλεψη
- True negative (TN): Επιτυχημένη αρνητική πρόβλεψη
- False positive (FP): Αποτυχημένη θετική πρόβλεψη
- False negative (FN): Αποτυχημένη αρνητική πρόβλεψη

Accuracy: Το ποσοστό των επιτυχημένων προβλέψεων του συνόλου προβλέψεων.

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

Precision: Το ποσοστό των σωστών θετικών προβλέψεων σε σχέση με όλες τις θετικές προβλέψεις.

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

Recall: Το ποσοστό των σωστών θετικών προβλέψεων σε σχέση με το σύνολο των πραγματικών θετικών περιπτώσεων.

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

ROC AUC: Το εμβαδόν κάτω από την καμπύλη ROC (Receiver Operating Characteristic). Μετρά την ικανότητα διάκρισης μεταξύ θετικών και αρνητικών κλάσεων.

Std (Standard deviation of accuracy): Μετρά τη διασπορά των τιμών της μετρικής Accuracy.

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$$

5.2 Συμπεράσματα αποτελεσμάτων

Μετά τη διαδικασία του 10-fold cross validation παρατηρούμε ότι ο αλγόριθμος με το μεγαλύτερο accuracy (0.72) είναι ο Gradient Boosting (GB). Το μικρότερο standard deviation (0.008) και το μεγαλύτερο recall (0.66) το έχει ο αλγόριθμος Extreme Gradient Boosting (XGB), ενώ το μεγαλύτερο ROC AUC (0.77) το μοιράζονται οι αλγόριθμοι Random Forest (RF), Gradient Boosting (GB) και Extreme Gradient Boosting (XGB).

Results	DT	SVM	RF	ADA	GB	XGB
Accuracy	0.62	0.69	0.71	0.71	0.72	0.71
Std	0.01	0.01	0.01	0.01	0.009	0.008
AUC	0.62	0.74	0.77	0.76	0.77	0.77
Recall	0.61	0.56	0.65	0.62	0.62	0.66
Precision	0.62	0.76	0.74	0.75	0.77	0.73

Πίνακας 18: Αποτελέσματα 10-fold cross validation

Για την επιλογή του βέλτιστου αλγόριθμου δώσαμε βάση στην μετρική recall μιας και σε ένα ρεαλιστικό πλαίσιο μας ενδιαφέρει το να συμπεριλάβουμε όσους περισσότερους πιθανούς πελάτες γίνεται. Για αυτό το λόγο, ο αλγόριθμος που επιλέχθηκε για την υλοποίηση του συστήματος υποστήριξης αποφάσεων είναι ο Extreme Gradient Boosting (XGB).

XGB	Accuracy	Precision	Recall	AUC
Balanced				
train	0.72	0.59	0.80	0.72
test	0.72	0.58	0.79	0.71
Unbalanced				
test	0.79	0.60	0.38	0.71

Πίνακας 19: Αποτελέσματα παραμετροποίησης XGB

Μετά την παραμετροποίηση του αλγόριθμου που επιλέξαμε (Extreme Gradient Boosting) και προσπαθώντας να μεγιστοποιήσουμε την μετρική recall χωρίς να κάνουμε υπερφόρτωση του μοντέλου, παρατηρούμε ότι οι μετρικές precision και ROC AUC παραμένουν σταθερές εξετάζοντας και τα δεδομένα εκπαίδευσης αλλά και τα δεδομένα τεστ ανεξαρτήτως του αν έχει διορθωθεί η ανισορροπία του σετ δεδομένων. Αντιθέτως, η μετρική recall, στις περιπτώσεις όπου δεν έχουμε διορθώσει την ανισορροπία στα δεδομένα τεστ, πέφτει στο 0.38, πράγμα που σημαίνει ότι στις πιο ρεαλιστικές κατανομές δεδομένων το 62% των πιθανών πελατών με θετικά αποτελέσματα δεν υπολογίζονται στις προβλέψεις του μοντέλου.

6 Υλοποίηση του συστήματος

Σε αυτό το κεφάλαιο θα γίνει η περιγραφή της ανάπτυξης του συστήματος υποστήριξης αποφάσεων το οποίο θα είναι υπεύθυνο για την πρόβλεψη πιθανών συνδρομητών σε τραπεζικούς καταθετικούς λογαριασμούς μίας τραπεζικής οντότητας. Επίσης θα γίνει η περιγραφή της υλοποίησης μίας εφαρμογής διαχείρισης πελατολογίου η οποία θα χρησιμοποιηθεί για την προσομοίωση της συνεργασίας του συστήματος υποστήριξης αποφάσεων που αναπτύξαμε με ενδεχόμενες τρίτες εφαρμογές διαχείρισης πελατολογίου σε ένα ρεαλιστικό πλαίσιο εκστρατείας μάρκετινγκ.

6.1 Ανάπτυξη διεπαφής προγραμματισμού

Μετά την επιλογή του βέλτιστου μοντέλου πρόβλεψης, θα πρέπει να γίνει η ενσωμάτωσή του σε ένα ολοκληρωμένο σύστημα υποστήριξης αποφάσεων. Η ανάπτυξη του συστήματος υποστήριξης αποφάσεων έγινε στο πλαίσιο της ενσωμάτωσης του μοντέλου πρόβλεψης σε μία διεπαφή προγραμματισμού η οποία θα μπορεί να χρησιμοποιηθεί από τρίτες εφαρμογές διαχείρισης υποψήφιων πελατών. Το περιβάλλον ανάπτυξης που χρησιμοποιήθηκε για την υλοποίηση της διεπαφής είναι το Flask, το οποίο αποτελεί ένα ιδανικό framework για την ανάπτυξη ηλεκτρονικών διαδικτυακών μικροϋπηρεσιών. Η συγγραφή της διεπαφής πραγματοποιήθηκε με τη χρήση της γλώσσας προγραμματισμού Python αλλά και χρησιμοποιώντας κάποια έτοιμα εργαλεία από βιβλιοθήκες της Python για την επιτέλεση συγκεκριμένων λειτουργιών. Παρακάτω βρίσκεται η πλήρης λίστα από βιβλιοθήκες της Python όπου χρησιμοποιήθηκαν:

Όπως αναφέραμε προηγουμένως, η χρησιμότητα της διεπαφής βασίζεται στην συνεργασία της με τρίτες εφαρμογές διαχείρισης πελατολογίου τράπεζας. Αυτή η συνεργασία επιτυγχάνεται με την ανταλλαγή δεδομένων μεταξύ της εφαρμογής διαχείρισης και της διεπαφής. Η ανταλλαγή αυτών των δεδομένων πραγματοποιείται μέσω του διαδικτυακού πρωτόκολλου HTTP μέσω δικτυακών διαδρομών (network routes) οι οποίες εξυπηρετούν διαφορετικές διαδικασίες. Η διεπαφή που αναπτύξαμε περιέχει δύο δικτυακές διαδρομές. Η πρώτη (/prediction) είναι υπεύθυνη για την διαδικασία προβλέψεων και αποστολής αποτελεσμάτων, ενώ η δεύτερη (/training/new) είναι υπεύθυνη για την επανεκπαίδευση του μοντέλου πρόβλεψης. Πριν την αποστολή δεδομένων σε αυτές τις δικτυακές διαδρομές απαιτείται η μετατροπή των αποστέλλομενων δεδομένων σε μορφή json. Η αποστολή των δεδομένων γίνεται μέσω αιτημάτων POST, και για την επιτυχημένη διεξαγωγή αυτών των αιτημάτων έχει γίνει εφαρμογή διαδικασιών αποφυγής σφάλματος. Κατόπιν της αποστολής δεδομένων, ανεξάρτητα της δικτυακής

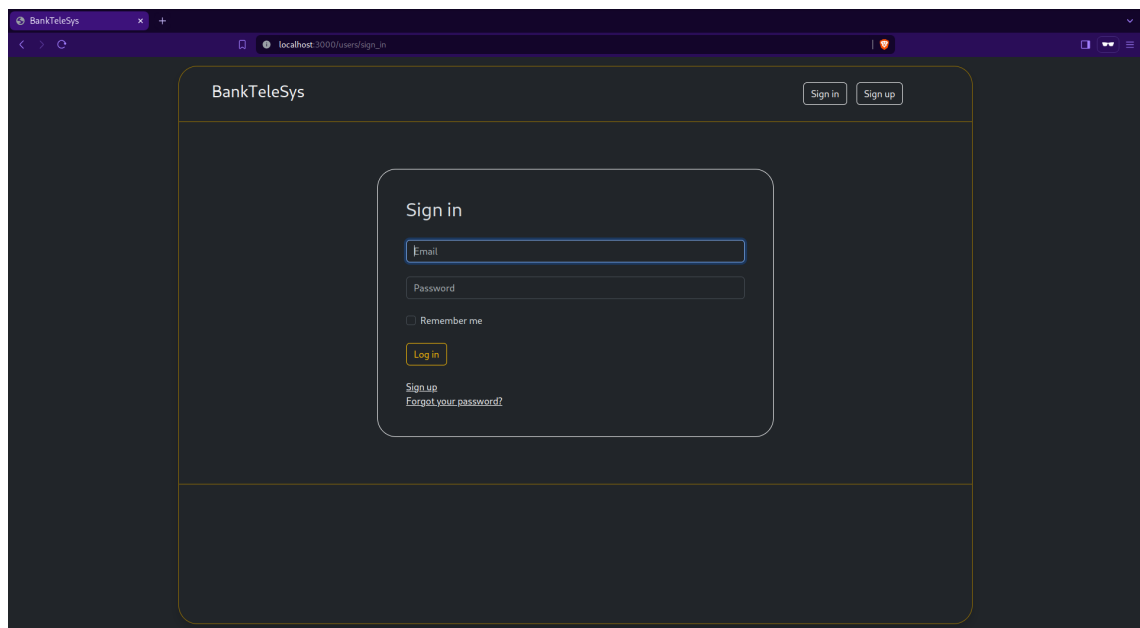
διαδρομής στην οποία πραγματοποιείται, γίνεται έλεγχος των μεταβλητών του σετ δεδομένων αλλά και της γενικής μορφής του σετ δεδομένων από τη συνάρτηση `schema_validation`.

Για την αποστολή δεδομένων στην πρώτη δικτυακή διαδρομή (`/prediction`) θα πρέπει να περιέχεται στα δεδομένα η μεταβλητή `id` η οποία θα αποδίδει έναν ξεχωριστό ακέραιο αριθμό σε κάθε καταχώρηση. Αυτή η απαίτηση υπάρχει για να γίνεται η αναγνώριση του κάθε πελάτη πριν και μετά τη διαδικασία πρόβλεψης των αποτελεσμάτων αποφεύγοντας τη μεταφορά ευαίσθητων δεδομένων όπως είναι το ονοματεπώνυμο ή τα στοιχεία επικοινωνίας του κάθε πελάτη. Αφού γίνει η αποστολή των δεδομένων στη συγκεκριμένη δικτυακή διαδρομή, ακολουθεί η επεξεργασία και η προετοιμασία των δεδομένων, σύμφωνα με τις τεχνικές που αναλύσαμε στο κεφάλαιο της μεθοδολογίας, και μετέπειτα πραγματοποιούνται οι προβλέψεις αποτελεσμάτων για την κάθε καταχώρηση του σετ δεδομένων από το μοντέλο πρόβλεψης που αναπτύξαμε προηγουμένως. Τέλος, γίνεται η αποστολή των `id` των καταχωρήσεων μαζί με τις προβλέψεις (`predictions`) της κάθε καταχώρησης στον αρχικό πομπό.

Για την αποστολή δεδομένων στην δεύτερη δικτυακή διαδρομή (`/training/new`), ο αριθμός των καταχωρήσεων που περιέχονται θα πρέπει να ξεπερνάει τις 999 καταχωρήσεις για να ακολουθήσει η διαδικασία της επανεκπαίδευσης του μοντέλου. Σε περίπτωση που οι καταχωρήσεις είναι λιγότερες από το απαιτούμενο μέγεθος, η διεπαφή επιστρέφει ένα μήνυμα που περιγράφει τον λόγο που δεν έγινε δεκτό το αίτημα. Αφού πραγματοποιηθεί μία επιτυχημένη αποστολή δεδομένων, ακολουθούν διαδικασίες επεξεργασίας και προετοιμασίας των δεδομένων, σύμφωνα με αυτές που αναλύσαμε στο κεφάλαιο της μεθοδολογίας, και στη συνέχεια γίνεται η επανεκπαίδευση του αλγόριθμου. Για γίνει δεκτή η καθιέρωση του καινούργιου μοντέλου, θα πρέπει να υπάρχει βελτίωση στις μετρικές επιδόσεων σε σύγκριση με το προηγούμενο μοντέλο πρόβλεψης.

6.2 Ανάπτυξη εφαρμογής διαχείρισης πελατολόγιου

Η ανάπτυξη της συγκεκριμένης εφαρμογής πραγματοποιήθηκε με τις προδιαγραφές ενός web application το οποίο θα χρησιμοποιείται από τους υπαλλήλους, μίας ενδεχόμενης τράπεζας, οι οποίοι εργάζονται στον τομέα του τηλεφωνικού μάρκετινγκ για την προώθηση τραπεζικών προϊόντων. Το περιβάλλον ανάπτυξης στο οποίο πραγματοποιήθηκε η υλοποίηση της εφαρμογής είναι το Rails, το οποίο αποτελεί ένα ιδανικό framework για την ανάπτυξη διαδικτυακών εφαρμογών. Η συγκεκριμένη εφαρμογή περιέχει δικό της γραφικό περιβάλλον και προσφέρει διάφορες λειτουργίες, που βοηθάνε στην διαχείριση πελατών, τις οποίες θα αναλύσουμε παρακάτω.



Σχήμα 10: Sign in page

Για να αποκτήσει ο χρήστης πρόσβαση στην εφαρμογή θα πρέπει να δημιουργήσει έναν προσωπικό λογαριασμό και στη συνέχεια θα πρέπει να συνδέεται στην εφαρμογή μέσω του λογαριασμού που έφτιαξε. Όταν ο υπάλληλος είναι συνδεδεμένος στον λογαριασμό του θα μπορεί αποκτήσει πρόσβαση στα εργαλεία και λειτουργίες της εφαρμογής.

BankTeleSys

Customers

ID	First Name	Last Name	Contact	Calls	Prediction	Outcome
216	Fname1	Lname1	+30 211 111 1111	0	Negative	Unknown
217	Fname2	Lname2	+30 222 222 2222	0	Negative	Unknown
218	Fname3	Lname3	+30 233 333 3333	0	Negative	Unknown
219	Fname4	Lname4	+30 244 444 4444	0	Positive	Unknown
220	Fname6	Lname6	+30 691 111 1111	0	Negative	Unknown
221	Fname7	Lname7	+30 692 222 2222	0	Negative	Unknown
222	Fname8	Lname8	+30 693 333 3333	0	Positive	Unknown

New customer All entries Positive No calls No outcome Reset Delete all

Query Tool: Column Value Search ID Tool: Customer ID Show

Choose File No file chosen Import CSV API Tool: Prediction Action

Σχήμα 11: Customer list page

Στην κύρια σελίδα της εφαρμογής βρίσκεται η λίστα με τους υποψήφιους πελάτες που έχει αποθηκεύσει ο χρήστης. Σε αυτή τη λίστα εμφανίζονται μονάχα συγκεκριμένες πληροφορίες για τον κάθε πελάτη. Οι πληροφορίες αυτές περιέχουν στοιχεία ταυτοποίησης του πελάτη (id, ονοματεπώνυμο), στοιχεία επικοινωνίας του πελάτη (αριθμός τηλεφώνου), ο αριθμός των τηλεφωνικών κλήσεων που έχουν πραγματοποιηθεί στον συγκεκριμένο πελάτη στην τωρινή εκστρατεία, η πρόβλεψη του συστήματος υποστήριξης αποφάσεων για τον συγκεκριμένο πελάτη, και τέλος το πραγματικό αποτέλεσμα της εκστρατείας για τον συγκεκριμένο πελάτη. Πιο κάτω βρίσκονται εργαλεία που βοηθάνε στην οργάνωση αυτής της λίστας. Με το εργαλείο Query tool, ο χρήστης μπορεί να αναζητήσει συγκεκριμένες ομάδες πελατών σύμφωνα με τη μεταβλητή και την τιμή που τοποθετεί στο εργαλείο. Οι μεταβλητές που μπορεί να χρησιμοποιήσει ο χρήστης στο συγκεκριμένο εργαλείο δεν περιορίζονται σε αυτές που εμφανίζονται στη λίστα πελατών, αλλά μπορεί να χρησιμοποιήσει όλες τις διαθέσιμες μεταβλητές που περιγράφουν τον κάθε πελάτη όπου θα τις δούμε στη συνέχεια. Πάνω από το Query tool βρίσκονται τέσσερα πλήκτρα τα οποία με τη χρήση τους πραγματοποιούν κάποιες προκαθορισμένες συχνές αναζητήσεις μέσα στη λίστα των πελατών.

Πλήκτρο	Λειτουργία
All entries	Εμφάνιση όλων των πελατών.
Positive	Απομόνωση των πελατών με θετικές προ-βλέψεις.
No calls	Απομόνωση των πελατών που δεν έχουν δε-χτεί τηλεφωνική κλήση στην τωρινή εκστρα-τεία μάρκετινγκ.
No outcome	Απομόνωση των πελατών στους οποίους δεν έχει καθοριστεί ακόμη το πραγματικό απο-τέλεσμα της εκστρατείας.

Πίνακας 20: Query buttons

BankTeleSys

Customers Edit Profile Sign out

Customer: 216

First Name:	Fname1	Last Name:	Lname1	Date of Birth:	1960-01-01
Contact:	+30 211 111 1111	Email:	onemail@mail.gr	Balance:	1500
Job:	Retired	Marital Status:	Single	Education:	Tertiary
Credit in Default:	No	Housing Loan:	Yes	Personal Loan:	No
Calls:	0	Last Call:	2024-06-02	Previous Calls:	0
Previous Outcome:	Success	Prediction:	Negative	Outcome:	Unknown

Notes:

Edit this customer Back to customers Delete this customer

Query Tool: Column Value Search ID Tool: Customer ID Show

Choose File No file chosen Import CSV API Tool: Prediction Action

Σχήμα 12: Customer entry page

Στην περίπτωση όπου ο χρήστης επιθυμεί να αποκτήσει πρόσβαση σε περισσότερα στοιχεία ενός πελάτη μπορεί να το κάνει μέσω του πλήκτρου Show. Πατώντας το πλήκτρο Show, το οποίο βρίσκεται δίπλα από κάθε καταχώρηση πελάτη, ο χρήστης μπορεί να εισέλθει στη σελίδα profile του κάθε πελάτη στην οποία εμφανίζονται όλα τα διαθέσιμα στοιχεία του εκάστοτε πελάτη. Αυτό μπορεί να πραγματοποιηθεί και μέσω του ID tool πληκτρολογώντας το id του

πελάτη. Μέσα στη σελίδα profile του πελάτη, ο χρήστης μπορεί να κάνει αλλαγές στα στοιχεία, να γράψει σημειώσεις για τον πελάτη οι οποίες θα τον βοηθήσουν σε επόμενες εκστρατείες, και τέλος μπορεί να καταχωρήσει το αποτέλεσμα της εκστρατείας για τον πελάτη.

Με κάθε τηλεφωνική κλήση που πραγματοποιείται στον πελάτη, αυξάνεται αυτόματα ο αριθμός των καταγεγραμμένων κλήσεων στον πελάτη (campaign), αλλά και ανανεώνεται η ημερομηνία της τελευταίας κλήσης που πραγματοποιήθηκε στον πελάτη (για τον υπολογισμό του rdays). Στο τέλος της εκστρατείας, το τελικό αποτέλεσμα του πελάτη μετατρέπεται στο αποτέλεσμα της προηγούμενης εκστρατείας (roulcome), και ο αριθμός τηλεφωνικών κλήσεων στον πελάτη της τωρινής εκστρατείας (campaign) προσθέτεται στο συνολικό αριθμό κλήσεων στον πελάτη (previous) και στη συνέχεια μηδενίζεται.

Ο χρήστης έχει τη δυνατότητα διαγραφής καταγραφών πελατολόγιου, αλλά και τη δυνατότητα εισαγωγής νέων καταχωρήσεων πελατών. Η εισαγωγή νέων καταχωρήσεων υποψήφιων πελατών μπορεί να πραγματοποιηθεί με δύο τρόπους. Ο πρώτος τρόπος είναι μέσω της λειτουργίας new customer όπου ο χρήστης μπορεί να εισάγει πελάτες συμπληρώνοντας μία φόρμα με τα στοιχεία του πελάτη. Σε περίπτωση που ο χρήστης επιθυμεί να εισάγει μαζικά καταχωρήσεις πελατών, μπορεί να το κάνει μέσω της λειτουργίας Import CSV όπου μπορεί να εισάγει αρχείο μορφής CSV το οποίο περιέχει τα στοιχεία των πελατών.

The screenshot shows the 'New customer' page in the BankTeleSys application. The page is a web form for adding a new customer. It includes fields for First Name, Last Name, Date of Birth, Email, Contact, Balance, Job, Marital, Education, and various loan status checkboxes (Has credit in default, Has housing loan, Has personal loan). There are also fields for Last Call, Calls, Previous Calls, Previous Outcome, Prediction, and Outcome. At the bottom, there is a 'Back to customers' button and a 'Query Tool' section with a search bar and an 'Import CSV' button.

Σχήμα 13: New customer page

7 Συμπεράσματα

Το τηλεφωνικό μάρκετινγκ αποτελεί μία από τις πιο διαδεδομένες μεθόδους άμεσου μάρκετινγκ και χρησιμοποιείται από οντότητες όπως οι τραπεζικοί οργανισμοί για την διαφήμιση τραπεζικών προϊόντων και υπηρεσιών. Η ραγδαία αύξηση της μάζας των δεδομένων αλλά και η επιθυμία των τραπεζικών οντοτήτων να αυξήσουν την αποτελεσματικότητα των εκστρατειών μάρκετινγκ που διεξάγουν έχει δημιουργήσει την ανάγκη για βελτιώσεις στον τομέα του στοχευμένου μάρκετινγκ. Στόχος της εργασίας ήταν η υλοποίηση ενός ρεαλιστικού συστήματος υποστήριξης αποφάσεων το οποίο θα ανταποκρίνεται σε αυτή την ανάγκη. Η κύρια λειτουργία του συστήματος είναι η δημιουργία προβλέψεων σχετικά με την επιτυχία μίας εκστρατείας μάρκετινγκ η οποία εφαρμόζεται πάνω σε μία ομάδα υποψήφιων πελατών, δηλαδή η πρόβλεψη για το ποιοί πελάτες πρόκειται να γίνουν συνδρομητές ενός τραπεζικού καταθετικού λογαριασμού. Η ανάπτυξη αυτού του συστήματος ολοκληρώθηκε μελετώντας και εφαρμόζοντας σύγχρονες μεθόδους ανάλυσης και εξόρυξης δεδομένων από την ήδη υπάρχουσα διαθέσιμη βιβλιογραφία του κλάδου της επιστήμης δεδομένων.

Οι τεχνικές ανάλυσης και εξόρυξης δεδομένων, και μετέπειτα οι τεχνικές προετοιμασίας δεδομένων, μας βοηθάνε να αποκτήσουμε μία καλύτερη κατανόηση του σετ δεδομένων που χρησιμοποιούμε για την υλοποίηση μοντέλων μηχανικής μάθησης και κατέπεκτα να πετύχουμε μεγαλύτερα ποσοστά επιτυχίας στις προβλέψεις που προσπαθούμε να κάνουμε με αυτά τα μοντέλα. Παρόλα αυτά, το σημαντικότερο κομμάτι αυτής της διαδικασίας παραμένει η ποιότητα του ίδιου του σετ δεδομένων που χρησιμοποιούμε. Πολλές φορές, σε δεδομένα τέτοιου τύπου όπως το το σετ δεδομένων που χρησιμοποιήσαμε, παρατηρούμε κάποιες ανισορροπίες ή ενδεχομένως ακόμη και κάποιες προτιμήσεις εντός των δεδομένων από την πλευρά της οντότητας που έχει συλλέξει αυτά τα δεδομένα. Είναι αδύνατον να γνωρίζουμε το αν το συγκεκριμένο ινστιτούτο, κατά τη διάρκεια της συλλογής αυτών των δεδομένων, χρησιμοποιούσε τεχνικές στοχευμένου μάρκετινγκ. Για την υλοποίηση μοντέλων πρόβλεψης υψηλής ποιότητας, η διαδικασία της συλλογής των δεδομένων θα πρέπει να γίνεται οργανωμένα και με μεγάλη προσοχή ώστε να αποφεύγονται πιθανές ανισορροπίες και προτιμήσεις από την πλευρά της οντότητας που συλλέγει τα δεδομένα.

Μιλώντας για την ποιότητα των δεδομένων, το συγκεκριμένο σετ δεδομένων που αναλύσαμε και χρησιμοποιήσαμε για την υλοποίηση του συστήματος υποστήριξης αποφάσεων δεν περιέχει πληροφορίες οι οποίες σχετίζονται με το ίδιο το προϊόν το οποίο διαφημίζεται, δηλαδή τους ίδιους τους καταθετικούς λογαριασμούς. Το είδος αυτών των πληροφοριών αποτελεί ένα

πολύ σημαντικό κομμάτι για την πρόβλεψη του αποτελέσματος αλλά και για την κατανόηση κάποιων πιθανών τάσεων των δημογραφικών που εξετάζουμε προς τα προϊόντα αυτά. Με λίγα λόγια, η συμπερίληψη αυτού του είδους της πληροφορίας είναι απαραίτητη για την υλοποίηση συστημάτων πρόβλεψης υψηλής ποιότητας. Το ίδιο φυσικά ισχύει και για άλλου είδους πληροφορίες, όπως μεταβλητές που περιγράφουν το οικονομικό ή κοινωνικοπολιτικό πλαίσιο στο οποίο διεξάγεται η εκστρατεία μάρκετινγκ, οι οποίες απουσιάζουν από το σετ δεδομένων.

Σύμφωνα με τα αποτελέσματα που αποκτήσαμε εξετάζοντας την ικανότητα πρόβλεψης μίας ομάδας συγκεκριμένων αλγόριθμων μηχανικής μάθησης, μπορούμε να συμπεράνουμε ότι το ενδεχόμενο της απόλυτης ακρίβειας στα συγκεκριμένα μοντέλα πρόβλεψης δεν είναι ρεαλιστικό με το συγκεκριμένο σετ δεδομένων όπου χρησιμοποιήθηκε. Μάλιστα, η ακρίβεια προβλέψεων η οποία πιθανόν θα μας δινόταν στα πλαίσια μίας ρεαλιστικής εφαρμογής αυτών των μοντέλων, ενδέχεται να είναι διαφορετική από αυτή που εξετάσαμε. Αυτό συμβαίνει διότι η ισχύ των συγκεκριμένων δεδομένων μπορεί να διαφέρει από χώρα σε χώρα, από χρονιά σε χρονιά, ή ακόμη και από τράπεζα σε τράπεζα. Αυτό φυσικά δεν ακυρώνει τη χρησιμότητα ή την ανάγκη του συστήματος υποστήριξης αποφάσεων που αναπτύξαμε. Όπως αναφέραμε και προηγουμένως οι οργανισμοί οι οποίοι επιτελούν πρακτικές τηλεφωνικού μάρκετινγκ έρχονται αντιμέτωποι με τεράστια μεγέθη δεδομένων τα οποία πρέπει να διαχειριστούν. Η χρησιμότητα του μοντέλου που αναπτύξαμε δεν βρίσκεται στην απόλυτη ακρίβεια προβλέψεων αλλά στην απομόνωση μίας ομάδας πελατών από το αρχικό σύνολο των πελατών. Αυτή η ομάδα πελατών η οποία συλλέγεται από το μοντέλο περιέχει μεγαλύτερα ποσοστά επιτυχίας από το αρχικό σύνολο πελατών. Με αυτό τον τρόπο, πραγματοποιούμε μία μορφή στοχευμένου μάρκετινγκ αποφεύγοντας πιθανά κόστη και δαπάνες που θα προκαλούνταν στον κλάδο του μάρκετινγκ της τράπεζας.

Το σύστημα το οποίο αναπτύξαμε σε αυτή την εργασία αποτελεί μία καλή βάση στον τομέα των συστημάτων υποστήριξης αποφάσεων που σχετίζονται με το στοχευμένο άμεσο μάρκετινγκ και συγκεκριμένα με το τηλεφωνικό μάρκετινγκ από τραπεζικούς οργανισμούς για τη διαφήμιση τραπεζικών καταθετικών λογαριασμών. Η ικανότητα πρόβλεψης αυτού του συστήματος έχει την ικανότητα να αυξήσει την αποδοτικότητα στις τηλεφωνικές εκστρατείες μάρκετινγκ που διεξάγουν οι τράπεζες. Με την συλλογή και εισαγωγή νέων δεδομένων στο συγκεκριμένο σύστημα, κατόπιν καινούργιων εκστρατειών τηλεφωνικού μάρκετινγκ, είναι δυνατή η βελτίωσή του ως προς τις ικανότητες πρόβλεψης που κατέχει αλλά ταυτόχρονα η βελτίωση αυτή είναι εμπρόκειτο να είναι περιορισμένη. Για την καταπολέμηση αυτού του περιορισμού απαραίτητη είναι η εισαγωγή περισσότερων μεταβλητών στα δεδομένα που συλλέγονται με σκοπό την

ανάπτυξη ενός μοντέλου πρόβλεψης το οποίο θα προσφέρει αυξημένη ικανότητα προβλέψεων μέσα από πιο πολύπλοκες διαδικασίες υποστήριξης αποφάσεων.

8 Κώδικας

```
1 from flask import Flask, jsonify, request
2 import pandas as pd
3 import jsonschema
4 from jsonschema import validate
5 import os
6 from datetime import datetime
7 import json
8 import xgboost as xgb
9 import joblib
10 from sklearn.preprocessing import MinMaxScaler
11 from sklearn.model_selection import cross_val_score, StratifiedKFold
12
13 app = Flask(__name__)
14
15 # Import validation schemas
16 with open("schema/prediction.json") as file:
17     prediction_schema = json.load(file)
18
19 with open("schema/training.json") as file:
20     training_schema = json.load(file)
21
22 # Import feature preprocessing and encoding data
23 with open(f"preprocessing/target_mean_encoding.json", "r") as file:
24     encoding = json.load(file)
25
26 with open("preprocessing/columns_order.txt", "r") as file:
27     order = [line.strip() for line in file]
28
29 # Import model data
30 with open("model/metrics.json", "r") as file:
31     metrics = json.load(file)
32
33 model = joblib.load("model/xgboost_model.joblib")
34
35 model_blueprint = xgb.XGBClassifier(max_depth = 3,
36                                     subsample = 0.3,
37                                     n_estimators = 230,
38                                     learning_rate = 0.03,
39                                     min_child_weight = 1.5,
40                                     booster = "gbtree",
```

```

41         objective = "binary:hinge",
42         random_state = 42)
43
44 def schema_validation(data: list, schema: dict) -> bool:
45     if not isinstance(data, list):
46         return False
47     for index, entry in enumerate(data):
48         try:
49             validate(instance=entry, schema=schema)
50         except jsonschema.exceptions.ValidationError as e:
51             print(f"Validation failed for item {index + 1}: {e}")
52             return False
53     return True
54
55 def target_mean_encoding(data: pd.Series, target: pd.Series) -> pd.Series:
56     weight = 5
57     data = data.astype(object)
58     mean = target.mean()
59     agg = target.groupby(data).agg(["count", "mean"])
60     counts = agg["count"]
61     means = agg["mean"]
62     smooth = (counts * means + weight * mean) / (counts + weight)
63     column = data.map(smooth)
64     return column
65
66 def dict_encoding(keys: pd.Series, values: pd.Series) -> dict:
67     keys = keys.astype(str)
68     enc_dict = dict(zip(keys.unique(), values.unique()))
69     return enc_dict
70
71 def map_encoding(data: pd.Series, encoding: dict) -> pd.Series:
72     column = data.copy()
73     column = column.astype(str)
74     column = column.map(encoding)
75     column = fill_na(column, encoding)
76     return column
77
78 def fill_na(data: pd.Series, encoding: dict) -> pd.Series:
79     column = data.copy()
80     if column.isnull().values.any():
81         val = list(encoding.values())

```

```

82         avg = sum(val) / len(val)
83         column = column.fillna(avg)
84     return column
85
86 @app.route('/prediction', methods=['POST'])
87 def prediction():
88     try:
89         # Data validation
90         if not schema_validation(request.json["data"], prediction_schema):
91             return jsonify({"Error": "Validation error."})
92
93         # Data to dataframe
94         data = pd.DataFrame(request.json["data"])
95         customer_id = data.id
96         data = data.drop(columns=["id"])
97
98         # Create age from date_of_birth
99         data['date_of_birth'] = pd.to_datetime(data['date_of_birth'])
100        data.date_of_birth = data.date_of_birth.apply(lambda x: datetime.
101            now().year - x.year - ((datetime.now().month, datetime.now().
102            day) < (x.month, x.day)))
103        data = data.rename(columns={"date_of_birth": "age"})
104
105        # Transform contact
106        data.contact = data.contact.apply(lambda x: 'Cellular' if x.
107            startswith("+30 69") else 'Telephone')
108
109        # Create day and month
110        data["month"] = datetime.now().month
111        data["day"] = datetime.now().day
112
113        # Binary features encoding
114        binary_mapping = {"Yes": 1, "No": 0}
115        data.default = data.default.map(binary_mapping)
116        data.housing = data.housing.map(binary_mapping)
117        data.loan = data.loan.map(binary_mapping)
118
119        # Label encoding education
120        education_mapping = {'Primary': 0, 'Secondary': 1, 'Tertiary': 3}
121        data['education'] = data['education'].map(education_mapping)

```

```

120 # One hot encoding contact and marital
121 contact_unique = ['contact_Cellular', 'contact_Telephone']
122 dummies = pd.get_dummies(data['contact'], prefix="contact").astype(
123     int)
124 for val in contact_unique:
125     if val not in dummies.columns:
126         dummies[val] = 0
127 data = pd.concat([data, dummies], axis=1)
128 data = data.drop(columns='contact')
129
130 marital_unique = ['marital_Married', 'marital_Divorced', '
131     marital_Single']
132 dummies = pd.get_dummies(data['marital'], prefix="marital").astype(
133     int)
134 for val in marital_unique:
135     if val not in dummies.columns:
136         dummies[val] = 0
137 data = pd.concat([data, dummies], axis=1)
138 data = data.drop(columns='marital')
139 print(data.columns)
140
141 # Binning and label encoding age
142 data['age'] = data['age'].apply(lambda x: 0 if 18 <= x <= 25 else 1
143     if 25 < x <= 30 else 2 if 30 < x <= 35 else 3 if 35 < x <= 40
144     else 4 if 40 < x <= 45 else 5 if 45 < x <= 50 else 6 if 50 < x
145     <= 55 else 7 if 55 < x <= 60 else 8 if x > 60 else None)
146
147 # Apply target encoding to job and previous outcome
148 for column in ['job', 'previous_outcome']:
149     data[column] = map_encoding(data[column], encoding[column])
150
151 # Scaling and order
152 scaler = MinMaxScaler()
153 scaled_data = scaler.fit_transform(data)
154 data = pd.DataFrame(scaled_data, columns=data.columns)
155 data = data[order]
156
157 # Make predictions
158 y_pred = model.predict(data)
159 y_pred = pd.Series(y_pred, name="prediction")
160 result = pd.concat([customer_id, y_pred], axis=1)

```

```

155         result = result.replace({0: "Negative", 1: "Positive"})
156         response = result.to_dict(orient="records")
157         return jsonify(response)
158     except Exception as e:
159         return jsonify({"Error": str(e)}), 400
160
161 @app.route('/training/new', methods=['POST'])
162 def training_new():
163     try:
164         # Number of entries check
165         if not len(request.json["data"]) > 999:
166             return jsonify({"message": "Not enough data."})
167
168         # Dataset schema validation
169         if not schema_validation(request.json["data"], prediction_schema):
170             return jsonify({"message": "Validation error."})
171
172         # Data to dataframe and transformations
173         data = pd.DataFrame(request.json["data"])
174         data["calls"] = data["calls"] - 1
175         data['last_call'] = pd.to_datetime(data['last_call'])
176         data['month'] = data['last_call'].dt.month
177         data['day'] = data['last_call'].dt.day
178         data['date_of_birth'] = pd.to_datetime(data['date_of_birth'])
179         data.date_of_birth = data.date_of_birth.apply(lambda x: datetime.
180             now().year - x.year - ((datetime.now().month, datetime.now().
181             day) < (x.month, x.day)))
182
183         data = data.rename(columns={"date_of_birth": "age"})
184         data.contact = data.contact.apply(lambda x: 'Cellular' if x.
185             startswith("+30 69") else 'Telephone')
186
187         # Target variable
188         target = data['outcome']
189         target = target.map({"Success": 1, "Failure": 0})
190
191         data = data.drop(columns=["id", "outcome", "last_call"])
192
193         # Binary features encoding
194         binary_mapping = {"Yes": 1, "No": 0}
195         data.default = data.default.map(binary_mapping)
196         data.housing = data.housing.map(binary_mapping)

```

```

193     data.loan = data.loan.map(binary_mapping)
194
195     # Label encoding education
196     education_mapping = {'Primary': 0, 'Secondary': 1, 'Tertiary': 3}
197     data['education'] = data['education'].map(education_mapping)
198
199     # One hot encoding contact and marital
200     contact_unique = ['contact_Cellular', 'contact_Telephone']
201     dummies = pd.get_dummies(data['contact'], prefix="contact").astype(
202         int)
203     for val in contact_unique:
204         if val not in dummies.columns:
205             dummies[val] = 0
206     data = pd.concat([data, dummies], axis=1)
207     data = data.drop(columns='contact')
208
209     marital_unique = ['marital_Married', 'marital_Divorced', '
210         marital_Single']
211     dummies = pd.get_dummies(data['marital'], prefix="marital").astype(
212         int)
213     for val in marital_unique:
214         if val not in dummies.columns:
215             dummies[val] = 0
216     data = pd.concat([data, dummies], axis=1)
217     data = data.drop(columns='marital')
218
219     # Binning and label encoding age
220     data['age'] = data['age'].apply(lambda x: 0 if 18 <= x <= 25 else 1
221         if 25 < x <= 30 else 2 if 30 < x <= 35 else 3 if 35 < x <= 40
222         else 4 if 40 < x <= 45 else 5 if 45 < x <= 50 else 6 if 50 < x
223         <= 55 else 7 if 55 < x <= 60 else 8 if x > 60 else None)
224
225     # Create target encoding dictionary
226     encoding_dict = {}
227     for column in ['job', 'previous_outcome']:
228         column_encoding = target_mean_encoding(data[column], target)
229         column_encoding_dictionary = dict_encoding(data[column],
230             column_encoding)
231         encoding_dict[column] = column_encoding_dictionary
232
233     # Apply feature's encodings to the dataset

```

```

227     for column in ['job', 'previous_outcome']:
228         data[column] = map_encoding(data[column], encoding_dict[column]
229                                     ])
229
230     # Comparing mean accuracy and standard deviation with default model
231     # 's metrics
232     data = data[order]
233     kfold = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
234     cv_scores = cross_val_score(model_blueprint, data, target, cv=kfold
235                                 , scoring="accuracy")
236     if metrics["mean"] > cv_scores.mean() or metrics["std"] < cv_scores
237         .std():
238         print("Model did not change.")
239         return jsonify({"message": "Model did not change."})
240
241     # Creating and saving new model, new model's metrics, and feature
242     # encodings dictionary
243     model = model_blueprint.fit(data, target)
244     joblib.dump(model, "model/xgboost_model.joblib")
245     metrics["mean"] = cv_scores.mean()
246     metrics["std"] = cv_scores.std()
247
248     # saving new model metrics and encodings
249     with open("model/metrics.json", "w") as file:
250         json.dump(metrics, file, indent=4)
251
252     with open("preprocessing/target_mean_encoding.json", "w") as file:
253         json.dump(encoding_dict, file, indent=4)
254
255     return jsonify({"message": "New model created."})
256 except Exception as e:
257     return jsonify({"Error": str(e)}), 400
258
259 if __name__ == '__main__':
260     app.run(host="0.0.0.0", port=5000, debug=True)

```

Listing 1: Flask API

References

- Chen, Rung-Ching et al. (2020). “Selecting critical features for data classification based on machine learning methods”. In: *Journal of Big Data* 7.1, p. 52.
- Dahouda, Mwamba Kasongo and Inwhee Joe (2021). “A deep-learned embedding technique for categorical features encoding”. In: *IEEE Access* 9, pp. 114381–114391.
- Govindarajan, Muthukumarasamy (2015). “Comparative study of ensemble classifiers for direct marketing”. In: *Intelligent Decision Technologies* 9.2, pp. 141–152.
- Hossam, Meshref (2020). “Predicting loan approval of bank direct marketing data using ensemble machine learning algorithms”. In: *International Journal of Circuits, Systems and Signal Processing*.
- Kaisar, Shahriar et al. (2024). “Enhancing Telemarketing Success Using Ensemble-Based Online Machine Learning”. In: *Big Data Mining and Analytics* 7.2, pp. 294–314.
- Moro, Sergio, Raul Laureano, Paulo Cortez, et al. (2011). “Using data mining for bank direct marketing: An application of the crisp-dm methodology”. In: *Proceedings of the European Simulation and Modelling Conference-ESM*. Vol. 2011.
- Page, Carole and Ye Luding (2003). “Bank managers’ direct marketing dilemmas—customers’ attitudes and purchase intention”. In: *International Journal of Bank Marketing* 21.3, pp. 147–163.
- Patwary, Muhammed JA et al. (2021). “Bank deposit prediction using ensemble learning”. In: *Artificial Intelligence Evolution*, pp. 42–51.
- Tang, Xing and Yusi Zhu (2024). “Enhancing bank marketing strategies with ensemble learning: Empirical analysis”. In: *Plos one* 19.1, e0294759.
- Verma, Archit (2019). “Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA”. In: *International Research Journal of Engineering and Technology* 5.13, pp. 54–60.
- Wang, Deli (2020). “Research on bank marketing behavior based on machine learning”. In: *Proceedings of the 2nd international conference on artificial intelligence and advanced manufacture*, pp. 150–154.