# Πρόβλεψη αποχώρησης πελατών στον τραπεζικό τομέα με χρήση μηχανικής μάθησης

Σπύρος Σάββας
Πρόγραμμα Μεταπτυχιακών Σπουδών (Π.Μ.Σ.)
Τμήμα Ψηφιακών Συστημάτων
Μεγάλα Δεδομένα και Αναλυτική
Πανεπιστήμιο Πειραιώς
Πειραιάς, Ελλάδα
spyros_savvas@hotmail.com

ABSTRACT

Στο ανταγωνιστικό περιβάλλον του τραπεζικού κλάδου, η διατήρηση των πελατών είναι ζωτικής σημασίας για τη βιώσιμη ανάπτυξη και την κερδοφορία. Η παρούσα εργασία παρουσιάζει ένα νέο πλαίσιο μηχανικής μάθησης που έχει σχεδιαστεί για την πρόβλεψη της αποχώρησης πελατών, επιτρέποντας στις τράπεζες να σχεδιάζουν στρατηγικές διατήρησης πελατών πριν από την αποχώρηση. Χρησιμοποιούμε ένα πλούσιο σύνολο δεδομένων, που περιλαμβάνει δημογραφικά, συναλλακτικά και συμπεριφορικά χαρακτηριστικά, για να εκπαιδεύσουμε μια σειρά ταξινομητών, συμπεριλαμβανομένων των Logistic Regression, Random Forest και Naïve Bayes. Η μεθοδολογία μας περιλαμβάνει προεπεξεργασία δεδομένων, μετασχηματισμό χαρακτηριστικών και την επιλογή υπερπαραμέτρων για την ενίσχυση της απόδοσης του μοντέλου.

## I. Εισαγωγη

Σε μια εποχή με πιο πολυπληθείς αγορές και εντονότερο ανταγωνισμό μεταξύ των επιχειρήσεων, η απώλεια των πελατών αποτελεί σημαντικό πρόβλημα. Σύμφωνα με πολλές έρευνες, το κόστος απόκτησης νέων καταναλωτών είναι το 1/5 του κόστους διατήρησης των υπαρχόντων. Για το λόγο αυτό, οι επιχειρήσεις προτιμούν να διατηρούν τους υπάρχοντες πελάτες παρά να προσθέτουν νέους και εφαρμόζουν πολιτικές προς αυτή την κατεύθυνση. Ένα από τα πολυτιμότερα χαρακτηριστικά στις στρατηγικές που αποσκοπούν στη μείωση ή την αποτροπή της απομάκρυνσης πελατών είναι τα δεδομένα της συμπεριφοράς των πελατών της τρέχουσας πελατειακής βάσης. Κατά συνέπεια, στο πλαίσιο ενός στρατηγικού σχεδίου για τους καταναλωτές που στοχεύει στη μείωση της αποχώρησης πελατών, η ανακάλυψη και η διερεύνηση πελατών με έντονη επιθυμία να εγκαταλείψουν τον οργανισμό ή η πρόβλεψη της αποχώρησης πελατών είναι απαραίτητη.

Ο τραπεζικός τομέας είναι ένας από τους τομείς όπου η ανάλυση της συμπεριφοράς των πελατών και η εκτίμηση της απομάκρυνσης των πελατών με βάση αυτές τις συμπεριφορές αποτελεί ουσιαστικό θέμα έρευνας. Τα αποτελέσματα της ανάλυσης της απομάκρυνσης πελατών έχουν μεγάλο αντίκτυπο στην πολιτική της τράπεζας. Διότι επιτρέπουν στις τράπεζες να αναπτύξουν νέες στρατηγικές για τους πελάτες ή να βελτιώσουν τις υπάρχουσες. Επιπλέον, οι τράπεζες είναι ζωτικής σημασίας για τη χρηματοοικονομική ανάπτυξη και εξέλιξη μιας χώρας. Επειδή δεν είναι πάντοτε δυνατόν να αποκτηθούν νέοι πελάτες στην ανταγωνιστική τραπεζική αγορά, πρωταρχικός στόχος των τραπεζών είναι να διασφαλίσουν τη διατήρηση των υφιστάμενων πελατών. Επειδή οι τράπεζες, όπως και όλες οι επιχειρήσεις στον τομέα των υπηρεσιών, είναι πελατοκεντρικές, οι πελατειακές σχέσεις με τις τράπεζες αποτελούν προτεραιότητα για τη μακροπρόθεσμη επιχειρηματική τους επίτευξη. Μελέτες που έχουν διεξαχθεί για τον τραπεζικό τομέα διαφόρων χωρών έχουν αποκαλύψει ότι, λόγω της ανταγωνιστικής και δυναμικής φύσης του τραπεζικού τομέα, η διασφάλιση της ικανοποίησης των πελατών αποτελεί σημαντική πολιτική για την αποτροπή της απομάκρυνσης των πελατών. Η ανάπτυξη ισχυρών σχέσεων με τους πελάτες τους πλεονεκτεί στην υψηλή ικανοποίηση των πελατών και, συνεπώς, στην αφοσίωσή τους και μετατρέπεται στον πιο σημαντικό παράγοντα για τη σταθερότητα, την ανάπτυξη και την κερδοφορία τους.

## II. Μεθοδολογια

### *A. Λογιστική Παλινδρομηση*

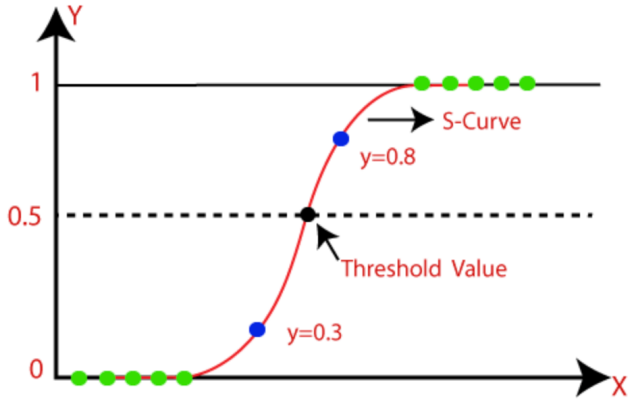
Η λογιστική παλινδρόμηση είναι ένας από τους από τους παλαιότερους αλγόριθμους ΜΜ και παρουσιάστηκε το 1958 από τον Cox. Με τον αλγόριθμο της λογιστικής παλινδρόμησης δημιουργούνται γραμμικά μοντέλα. Στο μοντέλο που δημιουργείται οι πιθανότητες που περιγράφουν τις πιθανές εξόδους μοντελοποιούνται χρησιμοποιώντας τη λογιστική συνάρτηση, που είναι γνωστή και ως σιγμοειδής συνάρτηση. Η σιγμοειδής συνάρτηση, δέχεται ως παράμετρο μια πραγματική τιμή $z$ και εξάγει ως αποτέλεσμα μία πραγματική τιμή που ανήκει στο διάστημα [0,1]. Η σιγμοειδής συνάρτηση δίνεται από τον τύπο:

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

Στο μοντέλο της λογιστικής παλινδρόμησης, υπολογίζεται το σταθμισμένο άθροισμα $z = w_0x_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$
των χαρακτηριστικών της εισόδου, και στο z εφαρμόζεται η σιγμοειδής συνάρτηση, η οποία επιστρέφει τιμές στο εύρος [0,1], οπότε υπολογίζεται η πιθανότητα $p = \sigma(z)$, όπου z το σταθμισμένο άθροισμα κάθε εισόδου.

Για ένα στιγμιότυπο του συνόλου εκπαίδευσης $x$ η πρόβλεψη της εξόδου $\hat{y}$ είναι:

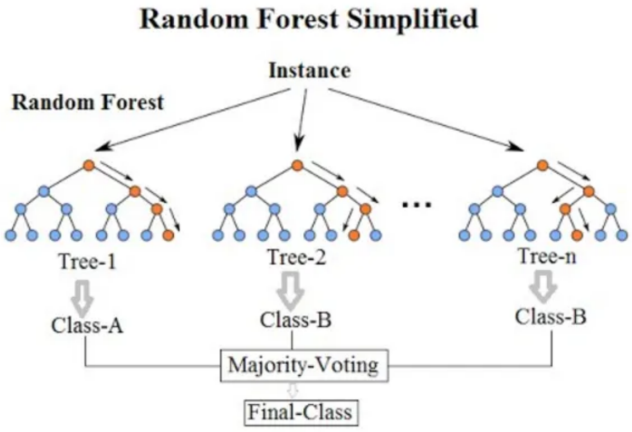
$$\hat{y} = \begin{cases} 0 \text{ εάν } \sigma(z) < 0.5 \\ 1 \text{ εάν } \sigma(z) \geq 0.5 \end{cases}$$




Εικόνα 1: Γραφικής Παράσταση Λογιστικής Παλινδρόμησης

### *B. Τυχαίο Δασος*

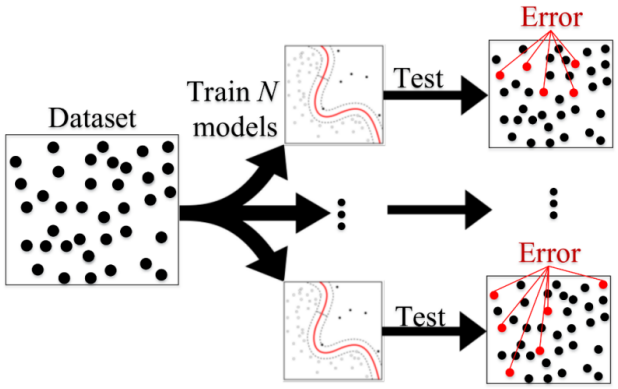
Το Τυχαίο Δάσος είναι μία μέθοδος συλλογικής μάθησης που βασίζεται στον αλγόριθμο του δέντρου αποφάσεων και στην τεχνική του bagging. Ο αλγόριθμος προτάθηκε αρχικά από τον Ho το 1996 και στη συνέχεια επεκτάθηκε και βελτιώθηκε από τον Breiman. Ο αλγόριθμος δέντρου αποφάσεων είναι ένας από τους παλαιότερους αλγόριθμους ταξινόμησης. Αρχικά αναπτύχθηκε από τον Quinlan, βελτιώθηκε από τον ίδιο το 1993 και αποτελεί μία μη παραμετρική μέθοδο επιβλεπόμενης μηχανικής μάθησης. Τα δέντρα απόφασης ονομάζονται έτσι επειδή, παρόμοια με ένα δέντρο, ο αλγόριθμος ξεκινά από έναν κόμβορίζα που αναπτύσσεται σε κλαδιά και έτσι δημιουργείται μια δομή που μοιάζει με δέντρο, όπως φαίνεται στο Σχήμα . Ο στόχος του αλγόριθμου είναι η δημιουργία ενός μοντέλου που προβλέπει την τιμή της μεταβλητής στόχου μαθαίνοντας απλούς κανόνες απόφασης που συμπεραίνονται από τα χαρακτηριστικά των δεδομένων. Με δοσμένα τα διανύσματα εκπαίδευσης $x$ και το διάνυσμα ετικέτας $y$, το δέντρο απόφασης διαχωρίζει αναδρομικά τον χώρο των χαρακτηριστικών έτσι ώστε δείγματα με την ίδια ετικέτα να ομαδοποιούνται μαζί. Υπάρχουν πολλοί διαφορετικοί αλγόριθμοι δέντρων απόφασης, όπως για παράδειγμα ο ID3, ο CART κλπ..




Εικόνα 2: Ο αλγόριθμος του Τυχαίου Δάσους

Για να διασφαλιστεί ότι η συμπεριφορά κάθε μεμονωμένου δένδρου δεν σχετίζεται με τη συμπεριφορά οποιουδήποτε άλλου δέντρου μέσα στο τυχαίο δάσος, χρησιμοποιείται η τεχνική του bagging. Ο αλγόριθμος του bagging προτάθηκε από τον Breiman το 1996 ως μέθοδος ταξινόμησης με ψηφοφορία (voting classifier). Ο αλγόριθμος δημιουργείται από διαφορετικά δείγματα bootstrap. Δοθέντος ενός συνόλου δεδομένων εκπαίδευσης $D$ με $K$ παραδείγματα, το bagging δημιουργεί $N$ νέα σύνολα εκπαίδευσης $D_i$ με δειγματοληψία από το $D$ με ομοιόμορφη κατανομή και με αντικατάσταση. Κατά τη δειγματοληψία κάποια παραδείγματα μπορεί να επαναλαμβάνονται σε κάθε $D_i$ . Κάθε ένα από αυτά τα δείγματα είναι ένα δείγμα εκκίνησης (bootstrap sample). Για κάθε ένα από αυτά τα δείγματα εφαρμόζεται ένας ταξινομητής $C_i$ για την εκπαίδευσή τους. Τελικά, ο ταξινομητής $C$ περιέχει ή

δημιουργείται από τους $C1$, $C2$, … , $CN$ των οποίων η έξοδος είναι η κλάση που προβλέπεται συχνότερα από τους υπο-κατηγοριοποιητές. Η τεχνική του bagging φαίνεται στο παρακάτω σχήμα.




Εικόνα 3: Η τεχνική του Bagging

### C. *Απλοϊκός Bayes*

Ο απλοϊκός Bayes (Naive Bayes) εμφανίζεται στη δεκαετία του 1990 και βασίζεται στην εφαρμογή του θεωρήματος του Bayes με την «αφελή» υπόθεση της υπό όρους ανεξαρτησίας μεταξύ του διανύσματος των χαρακτηριστικών $\boldsymbol{x}$ και της αντίστοιχης μεταβλητής κλάσης $y$.

Το θεώρημα του Bayes δηλώνει την ακόλουθη σχέση πιθανότητας, με δεδομένη της μεταβλητής κλάσης ταξινόμησης $y$ και το διάνυσμα με τις τιμές των n χαρακτηριστικών $x1$ έως $xn$:

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

Χρησιμοποιώντας την απλοϊκή υπόθεση της υπό όρους ανεξαρτησίας η οποία είναι:

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

Για όλα τα $i$ η παραπάνω σχέση απλοποιείται σε:

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

Καθώς η $P(x_1, \dots, x_n)$ είναι σταθερή δοθείσης της εισόδου $x$, εφαρμόζεται ο ακόλουθος κανόνας ταξινόμησης για την πρόβλεψη $\hat{y}$:

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

Στη συνέχεια χρησιμοποιείται η μέγιστη εκ των υστέρων εκτίμηση για την εκτίμηση της πιθανότητας $P(y)$ που είναι η σχετική συχνότητα της κλάσης $y$ στο σύνολο εκπαίδευσης και την εκτίμηση της $P(x_i|y)$.

Ο Naive Bayes επιτυγχάνει στην ταξινόμηση, εφόσον πραγματικά ισχύει η υπόθεση ότι με δεδομένη την κλάση οι μεταβλητές εισόδου είναι ανεξάρτητες. Οι διαφορετικοί Naive Bayes ταξινομητές διαφέρουν κυρίως με βάση την υπόθεση σε σχέση με την κατανομή της πιθανότητας $P(x_i|y)$ , δηλαδή εάν είναι Gaussian, Multinomial ή Bernulli. Σε προβλήματα ταξινόμησης κειμένου χρησιμοποιούνται συνήθως ο Multinomial NB, ο οποίος απαιτεί ως είσοδο ακέραια δεδομένα και ο Bernulli NB ο οποίος απαιτεί ως είσοδο δυαδικά δεδεμένα.

### D. *Μετρικές Απόδοσης των Μοντέλων Ταξινόμησης*

Προκειμένου να προσδιοριστεί ποια από τα εφαρμοζόμενα μοντέλα ταξινόμησης μηχανικής μάθησης είναι πιο επιτυχημένα τόσο μεμονωμένα όσο και μεταξύ τους, πρέπει να εξεταστούν ορισμένες μετρικές απόδοσης. Οι μετρικές χρησιμοποιούνται για την αξιολόγηση της αποτελεσματικότητας της χρησιμοποιούμενης μεθόδου ταξινόμησης και για τη σύγκριση των μοντέλων ταξινόμησης. Θα πρέπει να εξετάζονται πολλαπλές μετρικές των μοντέλων, διότι η αξιολόγηση αυτών των τιμών ως ενιαίο κριτήριο επιτυχίας θα ήταν εσφαλμένη. Όλες οι παρατηρήσεις στο σύνολο δεδομένων δοκιμής αντικαθίστανται στο μοντέλο που δημιουργείται με το σύνολο δεδομένων εκπαίδευσης στα μοντέλα ταξινόμησης και επιτυγχάνονται βαθμολογίες πρόβλεψης ταξινόμησης. Τα αποτελέσματα της σύγκρισης των προβλεπόμενων τιμών με τις πραγματικές τιμές χρησιμοποιούνται για να προσδιοριστεί πόσο καλά προβλέπει αυτό το μοντέλο, καθώς και η επιτυχία και η απόδοσή του. Ο πίνακας σύγχυσης (confusion matrix) συνοψίζει τα αποτελέσματα της ακρίβειας του μοντέλου στην πρόβλεψη, καθώς και τα συμπεράσματα της αξιολόγησης των επιδόσεων του μοντέλου ταξινόμησης μηχανικής μάθησης.

| Confusion Matrix | | Actual Values | |
|---|---|---|---|
| | | 1 | 0 |
| Predicted Values | 1 | True Positive TP | False Positive FP |
| | 0 | False Negative FN | True Negative TN |

Πίνακας 1: Πίνακας Σύγχυσης (Confusion Matrix)

Αληθώς θετική (TP): Δείχνει ότι οι παρατηρήσεις με πραγματική τιμή κλάσης 1 προβλέφθηκαν σωστά ως 1.

Αληθώς αρνητική (TN): Δείχνει ότι οι παρατηρήσεις με πραγματική τιμή κλάσης 0 προβλέπονται σωστά ως 0.

Ψευδώς αρνητική (FN): Δείχνει ότι οι παρατηρήσεις με πραγματική τιμή κλάσης 1 αξιολογούνται εσφαλμένα ως 0 ως αποτέλεσμα της πρόβλεψης.

Ψευδώς θετική (FP): Δείχνει ότι παρατηρήσεις με πραγματική τιμή κλάσης 0 αξιολογούνται εσφαλμένα ως 1 ως αποτέλεσμα της πρόβλεψης.

Η ακρίβεια (accuracy) ή αλλιώς ορθότητα, είναι το πιο συνηθισμένο μέτρο απόδοσης ενός μοντέλου. Η ορθότητα είναι η αναλογία των παραδειγμάτων για τα οποία έχει γίνει η σωστή πρόβλεψη επί του συνόλου των προβλέψεων και με βάση τις παραπάνω συμβάσεις και ισοδυναμεί με τον δείκτη λάθους (error rate) του μοντέλου και δίνεται από τον τύπο:

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

Η ανάκληση (recall), γνωστή και ως ευαισθησία (sensitivity) ή δείκτης αληθώς θετικών (True Positive Rate – TPR), είναι η αναλογία κάθε θετικού παραδείγματος που είναι πραγματικά θετικό. Αφορά την ικανότητα του μοντέλου να αναγνωρίζει ένα παράδειγμα της θετικής κλάσης και υπολογίζεται από τον τύπο:

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

Το Precision, επίσης γνωστό ως δείκτης αληθώς θετικών προβλέψεων, είναι η αναλογία των πραγματικά θετικών παραδειγμάτων μεταξύ όλων των παραδειγμάτων που έχουν προβλεφθεί ως θετικά. Αυτό το μέτρο αναφέρεται στην ικανότητα του μοντέλου να μην κατατάσσει ένα αρνητικό παράδειγμα ως θετικό. Ο τύπος για το Precision είναι:

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

Σε πολλές περιπτώσεις, εάν θέλουμε να συνοψίσουμε την απόδοση του μοντέλου, επιδιώκοντας ένα είδος ισορροπίας μεταξύ της ακρίβειας και την ανάκλησης, χρησιμοποιείται ένα μέτρο απόδοσης που συνδυάζει τα δύο αυτά μέτρα, τον αρμονικό μέσο όρο της ακρίβειας και της ανάκλησης που ονομάζεται F1 αποτέλεσμα (F1 score) και υπολογίζεται από τον τύπο:

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{TP}{TP + \frac{FN + FP}{2}}$$

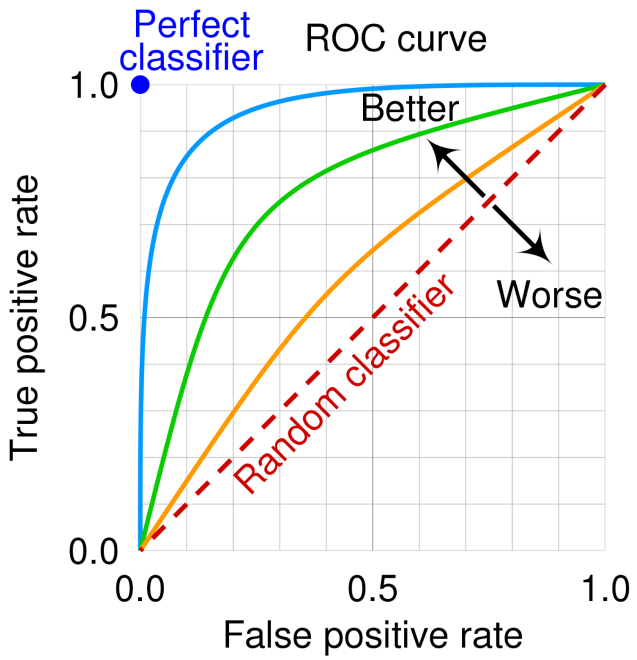
Η προσδιοριστικότητα (specificity), γνωστή και ως δείκτης αληθώς αρνητικών (True Negative Rate – TNR) μετρά τις ορθά αρνητικές προβλέψεις στο σύνολο των ορθών αρνητικών παραδειγμάτων και υπολογίζεται από τον τύπο:

$$Specificity = TNR = \frac{TN}{TN + FP}$$

Το ψευδώς θετικό ποσοστό (False Positive Rate - FPR) αντιστοιχεί στην αναλογία των αρνητικών παραδειγμάτων που θεωρήθηκαν ως θετικά, σε σχέση με όλα τα αρνητικά παραδείγματα. Όσο μεγαλύτερος είναι ο FPR, πόσα περισσότερα αρνητικά παραδείγματα έχουν ταξινομηθεί λάθος. Ο FPR υπολογίζεται από τον τύπο:

$$FPR = \frac{FP}{FP + TN} = 1 - specificity = 1 - TNR$$

Ένας εύκολος σχετικά γραφικός τρόπος και ο πιο συνηθισμένος για την ποιοτική εκτίμηση της απόδοσης ενός ταξινομητή είναι η λήψη χαρακτηριστικής καμπύλης λειτουργίας (Receiving Operating Characteristic curve – ROC curve). Η ROC καμπύλη είναι η σχεδίαση της μετρικής TPR, σε σχέση με την μετρική FPR για διάφορα κατώφλια. Η ROC καμπύλη χρησιμοποιείται ως μια γενική μετρική του μοντέλου. Όσο καλύτερο είναι ένα μοντέλο, τόσο υψηλότερα είναι η καμπύλη και συνεπώς τόσο μεγαλύτερη η επιφάνεια κάτω από την καμπύλη. Η επιφάνεια κάτω από την ROC καμπύλη (Area Under the Curve – AUC) ονομάζεται ROC AUC ή απλά AUC. Ο τέλειος ταξινομητής έχει ROC AUC ίση με 1, ενώ ένας κακός ταξινομητής θα έχει 0.5, ίσως και λιγότερο. Αυτό γιατί, όταν η τιμή είναι ίση με 1, αυτό σημαίνει ότι ο ταξινομητής μπορεί να διαχωρίσει πλήρως τις θετικές από τις αρνητικές κλάσεις. Εάν έχει την τιμή 0,5, τότε ο ταξινομητής δεν είναι σε θέση να διακρίνει μεταξύ θετικών και αρνητικών σημείων κλάσης και τότε ο ταξινομητής προβλέπει είτε τυχαία κλάση, είτε σταθερή κλάση για όλα τα σημεία δεδομένων. Συνεπώς, όσο πιο κοντά η ROC AUC στο 1, τόσο καλύτερος είναι ο ταξινομητής.



Εικόνα 4: Γραφική Παράσταση ROC

### *E. GridSearchCV*

Η τεχνική GridSearchCV είναι μια ισχυρή μέθοδος αυτοματοποιημένης αναζήτησης που χρησιμοποιείται για την εύρεση των βέλτιστων παραμέτρων για ένα μοντέλο μηχανικής μάθησης. Αυτό επιτυγχάνεται μέσω της εξέτασης ενός εύρους συνδυασμών παραμέτρων και της αξιολόγησης της απόδοσης του μοντέλου για κάθε έναν από αυτούς τους συνδυασμούς, χρησιμοποιώντας την τεχνική της διασταυρωμένης επικύρωσης (Cross Validation). Μετά την ολοκλήρωση της διαδικασίας, η GridSearchCV παρέχει το σετ παραμέτρων που επιτυγχάνει την καλύτερη απόδοση, βοηθώντας έτσι στην βελτιστοποίηση του μοντέλου.

### *F. Προβλήματα μη ισορροπημένης ταξινόμησης*

Για την ανάλυση της αποχώρησης πελατών, μελέτες έχουν εντοπίσει μια μη ισορροπημένη κατανομή κλάσεων σε σύνολα δεδομένων πελατών. Επειδή το μέγεθος του δείγματος των πελατών που αποχωρούν είναι σημαντικά μικρότερο από εκείνο των πελατών που δεν αποχωρούν, μπορεί να συμβεί το εξής σενάριο: η ακρίβεια της ταξινόμησης είναι υψηλή, ενώ η ακρίβεια πρόβλεψης των πελατών που αποχωρούν είναι χαμηλή. Έτσι, το πρόβλημα με τα μη ισορροπημένα σύνολα δεδομένων είναι ότι οι τυπικές τεχνικές εκμάθησης ταξινόμησης είναι συνήθως προκατειλημμένες προς τις πλειοψηφικές κλάσεις (που αναφέρονται ως "αρνητικές"), με αποτέλεσμα μεγαλύτερο ποσοστό λανθασμένης ταξινόμησης στις εμφανίσεις των μειοψηφικών κλάσεων (που αναφέρονται ως "θετικές" κλάσεις). Η πιο συνηθισμένη προσέγγιση σε αυτό το πρόβλημα είναι η χρήση μιας τεχνικής επαναδειγματοληψίας για την εξισορρόπηση της κατανομής των κλάσεων του συνόλου εκπαίδευσης πριν από την εκπαίδευση ενός μοντέλου ταξινόμησης. Η τυχαία υπερδειγματοληψία (RandomOverSampling - ROS) και η τυχαία υποδειγματοληψία (RandomUnderSampling - RUS) είναι δύο προσεγγίσεις για την επαναδειγματοληψία (RUS). Η ROS, η οποία συνίσταται στη μείωση των δεδομένων διαγράφοντας παραδείγματα που ανήκουν στην πλειοψηφούσα κλάση με στόχο την εξισορρόπηση του αριθμού των παραδειγμάτων κάθε κλάσης, και η RUS, η οποία σκοπεύει να αναπαράγει ή να δημιουργήσει νέα θετικά παραδείγματα προκειμένου να αποκτήσει σημασία. Το κύριο μειονέκτημα της τυχαίας υποδειγματοληψίας είναι ότι μπορεί να χαθούν δυνητικά σημαντικά δεδομένα που θα μπορούσαν να είναι σημαντικά στη διαδικασία επαγωγής. Η εξάλειψη δεδομένων είναι μια σημαντική απόφαση που πρέπει να ληφθεί υπόψιν. Η τυχαία υπερδειγματοληψία, από την άλλη πλευρά, μπορεί να αυξήσει την πιθανότητα υπερπροσαρμογής (overfitting), καθώς αντιγράφει ακριβώς τις περιπτώσεις της μειονοτικής κλάσης. Με αυτόν τον τρόπο, ένας ταξινομητής μπορεί να παράγει κανόνες που φαίνονται ακριβείς αλλά καλύπτουν μόνο μία αναπαραγόμενη περίπτωση.

Άλλη μια τεχνική είναι η SMOTE (Τεχνική Συνθετικής Μείωσης Υπερδειγματοληψίας Μειονοτήτων) όπου δημιουργήθηκε από τους Chawla, Bowyer, Hall και Kegelmeyer και αποτελεί μια στρατηγική που επικεντρώνεται στην τεχνητή δημιουργία νέων δειγμάτων της μειονοτικής κλάσης (θετικές κλάσεις) με βάση τα υπάρχοντα δείγματα. Η διαδικασία της SMOTE περιλαμβάνει τον εντοπισμό των k πλησιέστερων γειτόνων για κάθε δείγμα της μειονοτικής κλάσης στον χώρο των χαρακτηριστικών. Στη συνέχεια, για τη δημιουργία ενός νέου συνθετικού δείγματος, επιλέγεται τυχαία ένας από αυτούς τους γείτονες και λαμβάνεται ένας συνδυασμός των χαρακτηριστικών του αρχικού δείγματος και του επιλεγμένου γείτονα, μέσω ενός τυχαίου βάρους, για να παραχθεί ένα νέο δείγμα. Αυτή η μέθοδος οδηγεί στην ενίσχυση της μειονοτικής κλάσης και στην εξισορρόπηση της κατανομής των κλάσεων, χωρίς να χάνονται σημαντικά δεδομένα, όπως συμβαίνει με την τυχαία υποδειγματοληψία, και αποφεύγοντας τον κίνδυνο υπερπροσαρμογής που συνδέεται με την τυχαία υπερδειγματοληψία. Η SMOTE έχει αποδειχθεί ιδιαίτερα χρήσιμη σε πολλές εφαρμογές ταξινόμησης, βελτιώνοντας την απόδοση των μοντέλων σε μη ισορροπημένα σύνολα δεδομένων.

## III. Ανάλυση των Δεδομένων

### *A. Περιγραφή των Δεδομένων*

Το σύνολο δεδομένων που χρησιμοποιήθηκε για την υλοποίηση των μοντέλων μηχανικής μάθησης προέρχεται από το Kaggle https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn το οποίο περιέχει 10000 εγγραφές και τα εξής χαρακτηριστικά:

| **Variable** | **Definition** |
|---|---|
| RowNumber | Correspond to the record number and has no effect on the output. |
| CustomerId | Contains random values and has no effect on customer leaving the bank. |
| Surname | The surname of a customer has no impact on their decision to leave the bank. |
| CreditScore | Can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank. |
| Geography | A customer's location can affect their decision to leave the bank. |
| Gender | It's interesting to explore whether gender plays a role in a customer leaving the bank. |
| Age | This is certainly relevant, since older customers are less likely to leave their bank than younger ones. |
| Tenure | Refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank. |
| Balance | Also a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances. |
| NumOfProducts | Refers to the number of products that a customer has purchased through the bank. |
| HasCrCard | Denotes whether or not a customer has a credit card.This column is also relevant, since people with a credit card are less likely to leave the bank. |
| IsActiveMember | Active customers are less likely to leave the bank. |
| EstimatedSalary | As with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries. |
| Exited | Whether or not the customer left the bank. |
| Complain | Customer has complaint or not. |
| Satisfaction Score | Score provided by the customer for their complaint resolution. |
| Card Type | Type of card hold by the customer. |
| Points Earned | The points earned by customer for using credit card |

Πίνακας 2: Περιγραφή των Δεδομένων

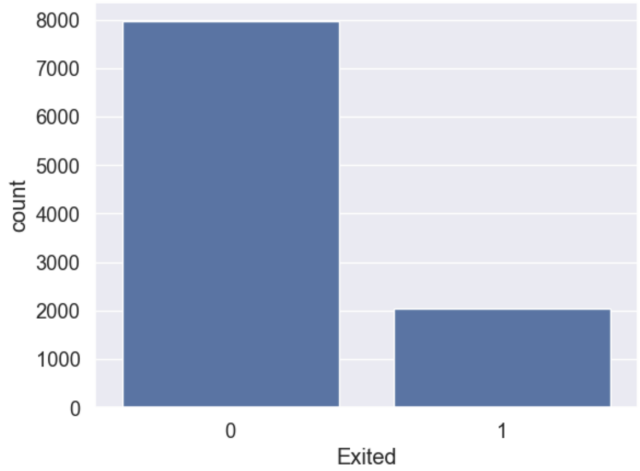
### *B. Προεπεξεργασία Δεδομένων*

Η προεπεξεργασία των δεδομένων αποτελεί ένα κρίσιμο βήμα στη διαδικασία της μηχανικής μάθησης καθώς βελτιώνει την ποιότητα των δεδομένων και εξασφαλίζει την αποδοτικότητα και την ακρίβεια των μοντέλων. Περιλαμβάνει διάφορες διαδικασίες όπως η αντιμετώπιση των χαμένων τιμών, η κανονικοποίηση των τιμών για την εξασφάλιση συνοχής, καθώς και η επιλογή και μετασχηματισμός χαρακτηριστικών για την αποφυγή περιττής πολυπλοκότητας και την ενίσχυση της

ερμηνευσιμότητας του μοντέλου. Αυτό το βήμα είναι θεμελιώδες για την κατασκευή αξιόπιστων μοντέλων που μπορούν να προβλέψουν με ακρίβεια και να παρέχουν έγκυρα αποτελέσματα.

Αρχικά με την βοήθεια της γλώσσας προγραμματισμού Python και της βιβλιοθήκης pandas διαπιστώνουμε ότι δεν υπάρχουν χαμένες τιμές.

Στο ακόλουθο σχήμα βλέπουμε ότι το σύνολο δεδομένων μας είναι unbalanced καθώς η κλάση Exited που θέλουμε να προβλέψουμε δεν είναι διαμοιρασμένη ισόποσα:




Γράφημα 1: Αριθμός παρατηρήσεων ανά κατηγορία

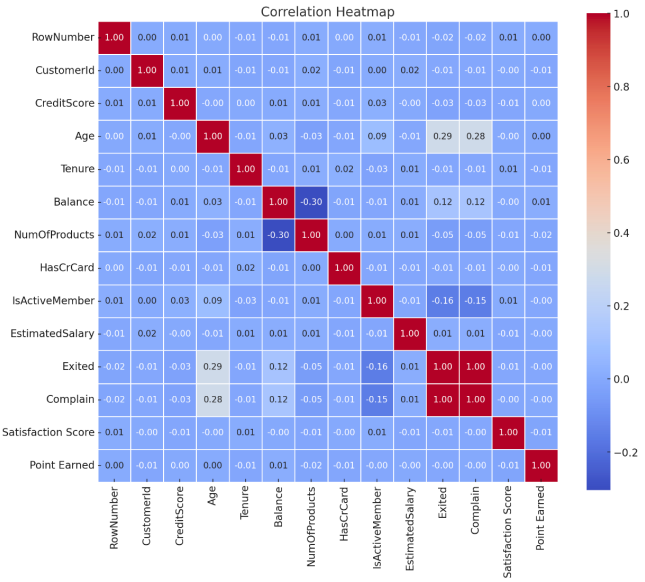
Στον παρακάτω πίνακα παρατίθεται κατάλογος των μεθόδων μετασχηματισμού που χρησιμοποιούμε για κάθε χαρακτηριστικό του συνόλου δεδομένων.

| **Feature** | **Transformation Method** |
|---|---|
| RowNumber | Removed |
| CustomerId | Removed |
| Surname | Removed |
| CreditScore | Standardization |
| Geography | One Hot Encoding |
| Gender | Label Encoding |
| Age | Standardization |
| Tenure | Standardization |
| Balance | Standardization |
| NumOfProducts | Standardization |
| HasCrCard | Label Encoding |
| IsActiveMember | Label Encoding |
| EstimatedSalary | Standardization |
| Satisfaction Score | Standardization |
| Card Type | One Hot Encoding |
| Points Earned | Standardization |

Πίνακας 3: Μέθοδοι Μετασχηματισμών

Αρχικά τα χαρακτηριστικά RowNumber, CustomerId, Surname αφαιρέθηκαν καθώς δεν μας προσφέρουν κάποια χρήσιμη πληροφορία. Στην συνέχεια στα κατηγορικά χαρακτηριστικά όπως Geography και Card Type έγινε χρήση του One Hot Encoding, ενώ στα δυαδικά (binary) χαρακτηριστικά Geography και Card Type έγινε χρήση του LabelEncoding.

### *C. Ανάλυση Συσχέτισης*




Γράφημα 2: Θερμικός Χάρτης Συσχέτισης

Η χρήση του correlation heatmap (θερμικού χάρτη συσχέτισης) προσφέρει μια άμεση και οπτικά προσβάσιμη μέθοδο για την ανάλυση της σχέσης μεταξύ δύο ή περισσότερων μεταβλητών σε ένα σύνολο δεδομένων. Επιτρέπει την εύκολη αναγνώριση και ερμηνεία των σχέσεων μεταξύ μεταβλητών, αναδεικνύοντας θετικές ή αρνητικές συσχετίσεις μέσω διαφορετικών χρωμάτων ή εντάσεων του χρώματος. Επίσης μπορεί να χρησιμοποιηθεί για την επιλογή χαρακτηριστικών, δηλαδή την επιλογή των πιο σημαντικών μεταβλητών για ένα πρόβλημα πρόβλεψης

ή ταξινόμησης, αποφεύγοντας τη χρήση μεταβλητών που παρέχουν παρόμοια πληροφορία.

Απο το παραπάνω γράφημα βλέπουμε ότι η μεταβλητή Complain έχει τέλεια συσχέτιση με τη μεταβλητή-στόχο Exited και είναι σημαντικό να εξεταστεί το ενδεχόμενο απόρριψης αυτού του χαρακτηριστικού για να αποφευχθεί η διαρροή δεδομένων (data leakage). Η διαρροή δεδομένων αναφέρεται στην κατάσταση στην οποία πληροφορίες από το σύνολο δεδομένων ελέγχου διαρρέουν στο σύνολο δεδομένων εκπαίδευσης προκαλώντας έτσι το μοντέλο να φαίνεται ότι έχει καλύτερη απόδοση στην πρόβλεψη από ό,τι πραγματικά έχει όταν εφαρμόζεται σε νέα, άγνωστα δεδομένα. Για τον λόγο αυτόν η αφαίρεση της στήλης Complain είναι ένα κρίσιμο βήμα για τη διασφάλιση της αξιοπιστίας και της δυνατότητας γενίκευσης του μοντέλου, επιτρέποντάς του να κάνει ακριβείς προβλέψεις σε πραγματικά σενάρια.

### *D. Συνολική Απόδοση των Μοντέλων Μηχανικής Μάθησης*

Η μεταβλητή στόχος του συνόλου δεδομένων είναι η Exited παίρνοντας τις τιμές ‘0’ αν ο πελάτης παραμένει στην τράπεζα και ‘1’ αν ο πελάτης αποχωρεί. Σκοπός είναι να συγκρίνουμε τους αλγόριθμους κατηγοριοποίησης και να επιλέξουμε τον βέλτιστο με βάση τις μετρικές απόδοσης που αναφέρθηκαν στην ενότητα της Μεθοδολογίας.

Συγκεκριμένα από αλγορίθμους κατηγοριοποίησης χρησιμοποιήσαμε:

- Logistic Regression
- Random Forest
- Naïve Bayes

Από τεχνικές υπερδειγματοληψίας και υποδειγματοληψίας χρησιμοποιήσαμε:

- RandomOverSampling (ROS)
- RandomUnderSampling (RUS)
- SMOTE

Τέλος τα δεδομένα από 10000 πελάτες χωρίστηκαν σε δύο μέρη, το 80% του συνόλου δεδομένων (8000 πελάτες) χρησιμοποιήθηκε για την εκπαίδευση των μοντέλων μηχανικής μάθησης, ενώ το 20% (2000 πελάτες) χρησιμοποιήθηκε για τη δοκιμή τους καθώς επίσης και η τεχνική της GridSearchCV για την επιλογή των βέλτιστων παραμέτρων. Η αποτελεσματικότητα κάθε επιμέρους μοντέλου προσδιορίστηκε με τη χρήση ορισμένων μετρικών απόδοσης. Τα αποτελέσματα αυτά τα αξιολογήσαμε μέσω των μετρικών απόδοσης των μοντέλων (π.χ. Accuracy, Recall, Precision, ROC curve, κτλ.)

Για τον Random Forest είχαμε τα εξής αποτελέσματα:

| Μετρικές | ROS | RUS | SMOTE |
|---|---|---|---|
| Accuracy | 0.866 | 0.796 | 0.861 |
| Precision | 0.801 | 0.713 | 0.792 |
| Recall | 0.742 | 0.790 | 0.731 |
| F1 Score | 0.765 | 0.732 | 0.755 |
| AUC | 0.742 | 0.790 | 0.731 |

Πίνακας 4: Αποτελέσματα Random Forest

Για την Logistic Regression είχαμε τα εξής αποτελέσματα:

| Μετρικές | ROS | RUS | SMOTE |
|---|---|---|---|
| Accuracy | 0.720 | 0.715 | 0.814 |
| Precision | 0.646 | 0.645 | 0.698 |
| Recall | 0.713 | 0.714 | 0.604 |
| F1 Score | 0.651 | 0.715 | 0.622 |
| AUC | 0.713 | 0.714 | 0.604 |

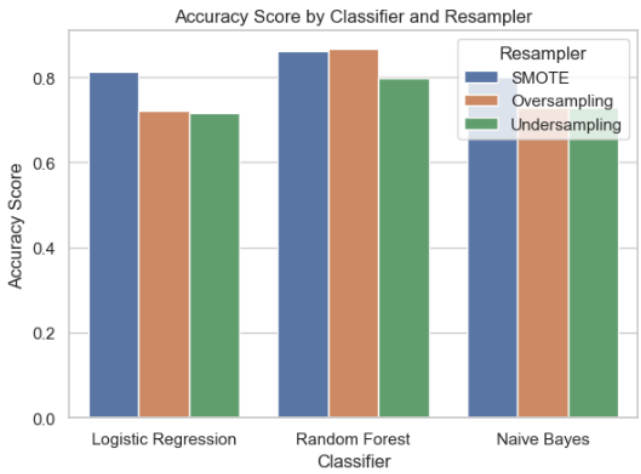
Πίνακας 5: Αποτελέσματα Logistic Regression

Για τον Naïve Bayes είχαμε τα εξής αποτελέσματα:

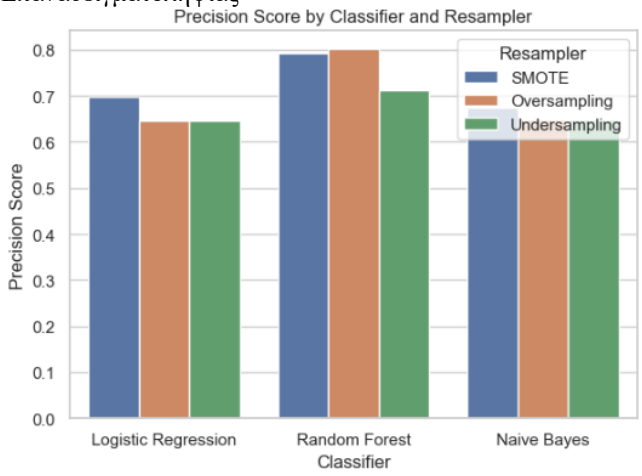
| Μετρικές | ROS | RUS | SMOTE |
|---|---|---|---|
| Accuracy | 0.728 | 0.729 | 0.814 |
| Precision | 0.648 | 0.648 | 0.698 |
| Recall | 0.712 | 0.712 | 0.604 |
| F1 Score | 0.655 | 0.656 | 0.622 |
| AUC | 0.712 | 0.712 | 0.604 |

Πίνακας 6: Αποτελέσματα Naïve Bayes

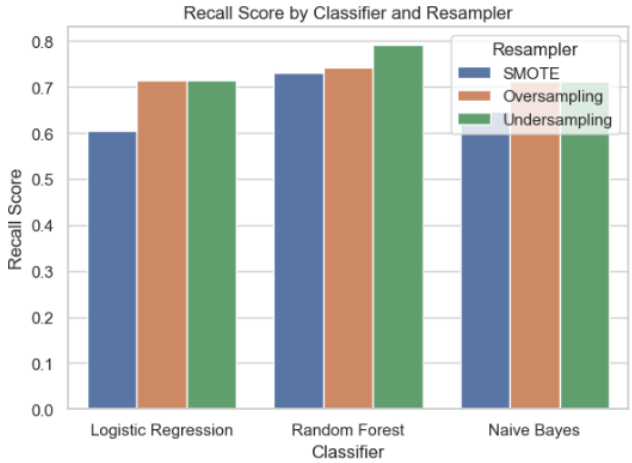
Παρακάτω ακολουθούν γραφήματα οπτικοποίησης των βασικών μετρικών που χρησιμοποιήσαμε για να αξιολογήσουμε τα μοντέλα μας:



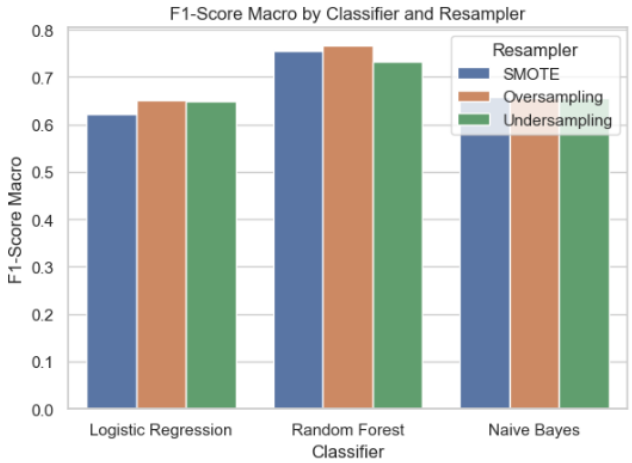
Γράφημα 3: Σύγκριση Ακρίβειας ανά Μοντέλο και Τεχνική Επαναδειγματοληψίας



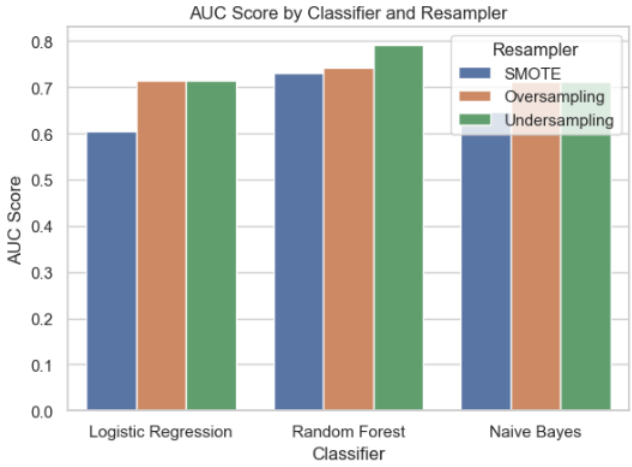

Γράφημα 4: Σύγκριση Precision ανά Μοντέλο και Τεχνική Επαναδειγματοληψίας




Γράφημα 5: Σύγκριση Recall ανά Μοντέλο και Τεχνική Επαναδειγματοληψίας




Γράφημα 6: Σύγκριση F1-Score ανά Μοντέλο και Τεχνική Επαναδειγματοληψίας




Γράφημα 7: Σύγκριση AUC Score ανά Μοντέλο και Τεχνική Επαναδειγματοληψίας

## IV. Συμπεράσματα

Από τα παραπάνω αποτελέσματα, παρατηρούμε ότι η εφαρμογή διαφορετικών τεχνικών επαναδειγματοληψίας, όπως η Random Over Sampling (ROS), Random Under Sampling (RUS), και Synthetic Minority Over-sampling Technique (SMOTE), επηρεάζει σημαντικά την απόδοση των μοντέλων μηχανικής μάθησης στο πρόβλημα ταξινόμησης. Συγκεκριμένα, για τον Random Forest, παρατηρείται ότι η τεχνική ROS προσφέρει την καλύτερη απόδοση σε όλες τις μετρικές σύγκρισης, ενώ η RUS φαίνεται να βελτιώνει την απόδοση του μοντέλου στον δείκτη AUC. Αντίστοιχα, για την Logistic Regression, η τεχνική SMOTE οδηγεί σε μεγαλύτερη ακρίβεια σε σχέση με τις άλλες δύο τεχνικές, δείχνοντας τη σημασία της επιλογής της κατάλληλης τεχνικής επαναδειγματοληψίας ανάλογα με το μοντέλο και το πρόβλημα. Τέλος, ο Naive Bayes δείχνει σχετικά ομοιόμορφη απόδοση μεταξύ των τριών τεχνικών, υπογραμμίζοντας τη σταθερότητα αυτού του μοντέλου σε διαφορετικές στρατηγικές επαναδειγματοληψίας. Συνολικά, τα αποτελέσματα αυτά υπογραμμίζουν την ανάγκη για προσεκτική επιλογή των τεχνικών επαναδειγματοληψίας και των μοντέλων μηχανικής μάθησης, προκειμένου να βελτιστοποιηθεί η απόδοση.

## V. ΒΙΒΛΙΟΓΡΑΦΙΑ


[1] ATHANASSOPOULOS,A.D.(2000).Customersatisfactioncuestosupportmarketsegmentationandexplainswitch- ing behavior. *Journal of business research*, *47*(3), 191-207.

[2] BLATTBERG,R.C.,KIM,B.D.,&NESLIN,S.A.(2008).ChurnManagement.In*DatabaseMarketing*(pp.607-633). Springer, New York, NY.

[3] COLGATE, M., STEWART, K., & KINSELLA, R. (1996). Customer defection: a study of the student market in Ire- land. International journal of bank marketing.

[4] H. Zhang, "The optimality of naive Bayes," 2004.

[5] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[6] J. R. Quinlan, "Induction of decision trees," Mach. Learn., vol. 1, no. 1, pp. 81–106, 1986.

[7] L. Breiman, "Bagging predictors," Mach. Learn., vol. 24, no. 2, pp. 123–140, 1996.

[8] E. Bauer and R. Kohavi, "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants," Mach. Learn., vol. 36, no. 1, pp. 105–139, 1999.

[9] A. Géron, Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, 2019.

[10] CHAWLA, N.V. (2009). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*, 875-886.

[11] GANESH,J.,ARNOLD,M.J.,&REYNOLDS,K.E.(2000).Understandingthecustomerbaseofserviceproviders:an examination of the differences between switchers and stayers. *Journal of marketing*, *64*(3), 65-87.