HELLENIC REPUBLIC
UNIVERSITY OF MACEDONIA

CLASSIFICATION

1st exercise of the machine learning cource

Chamalidis Vasileios Sotirios
University of Macedonia
dep. of Applied Informatics

Thessaloniki 2024

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

ΠΙΝΑΚΑΣ ΓΡΑΦΙΜΑΤΩΝ

ΠΙΝΑΚΕΣ

ΕΙΣΑΓΩΓΗ

Περιγραφή του Προβλήματος

Η χρεοκοπία αποτελεί ένα ανατρεπτικό γεγονός που μπορεί να προκληθεί από πολύπλευρους παράγοντες και έχει εκτεταμένες συνέπειες. Το αν μια επιχείρηση θα καταφέρει να παραμείνει εντός επιχειρησιακού ανταγωνισμού ή θα οδηγηθεί σε πτώχευση, αποτελεί σημαντικό ζήτημα για πιστωτές, μετόχους, διευθυντές και άλλους ενδιαφερόμενους. Οι συνέπειες της χρεοκοπίας επηρεάζουν σε μεγάλο βαθμό, τις θέσεις εργασίας μειώνοντας τες αλλά και, την οικονομική σταθερότητα. Ως εκ τούτου, η ακριβής πρόβλεψη της χρεοκοπίας αποτελεί υψίστης σημασίας τόσο, για την διαχείριση των επενδύσεων όσο και για την σωστή λήψη μέτρων.

Παραδοσιακά, η πρόβλεψη χρεοκοπίας βασίζονταν σε χρηματοοικονομικού δείκτες, οι οποίοι, λόγω της φύση τους, δυσκολεύονται να διακρίνουν, μη γραμμικές συσχετίσεις σε πολύπλοκα χρηματοοικονομικά δεδομένα. Ωστόσο, πρόοδοι στην τεχνητή νοημοσύνη, και ειδικά στον κλάδο της μηχανικής μάθησης, άνοιξαν νέους ορίζοντες στον τρόπο ανάλυσης αυτών των δεδομένων (Shi and Li, 2019).

Το παρόν έγγραφο έχει ως σκοπό, την συγκριτική ανάλυση μοντέλων ταξινόμησης που μπορούν να προβλέψουν τη χρεοκοπία. Χρησιμοποιήθηκαν δεδομένα που αφορούν την περίοδο 2006 μέχρι το 2009. Τα δεδομένα αυτά, καταμερίστηκαν σε 4 folds, όπου κάθε ένα περιέχει το δικό του train και test set και πραγματοποιήθηκε η κατάλληλη προεπεξεργασία τους. Τέλος, επιλέχτηκαν τα μοντέλα που ανταποκρίνονται σε ορισμένους περιορισμούς και αναδείχθηκε το καλύτερο.

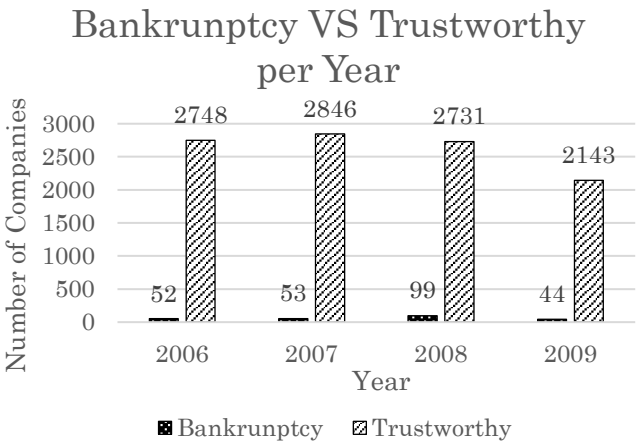
Δομή της Αναφοράς

Αρχικά, παρουσιάζεται μια προετοιμασία των δεδομένων του προβλήματος και σχολιάζονται διάφορες μέθοδοι επίλυσής ζητημάτων. Στην ενότητα ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ παρουσιάζονται τα αποτελέσματα απόδοσης των μοντέλων στο σύνολο τις βάσης και μετά ακολουθεί η σύγκρισή τους. Τέλος, γίνεται αναφορά στο καλύτερο μοντέλο και στα μοντέλα που ανταποκρίνονται σε συγκεκριμένες απαιτήσεις που μας έχουν τεθεί.

ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

Προεπεξεργασία Δεδομένων

Τα δεδομένα που χρησιμοποιήθηκαν περιλαμβάνουν: οκτώ δείκτες απόδοσης των εταιρειών, τρεις δυαδικούς δείκτες δραστηριοτήτων και μια ένδειξη ασυνέπειας. Στην ένδειξη ασυνέπειας ο αριθμός δύο υποδηλώνει, την χρεοκοπίας και ο αριθμός ένα, την μη-χρεοκοπία. Επιπλέον, υπάρχει και χρονική κατανομή στα δεδομένα. Είναι σημαντικό να σημειωθεί ότι οι τιμές της ένδειξης ασυνέπειας τροποποιήθηκαν κατά ν-1, ώστε να αναπαριστά το ένα την κλάση μειοψηφίας.



Εικόνα 1. Χρονική κατανομή των χρεοκοπημένων επιχειρήσεων και μη-χρεοκοπημένων.

Με μία πρώτη ματιά στα δεδομένα (Εικόνα 1), παρατηρούμε ότι η κλάση πλειοψηφίας αποτελείται από τις μη χρεοκοπημένες ή αλλιώς τις έμπιστες (*Trustworthy)* επιχειρήσεις. Ο αριθμός των έμπιστων επιχειρήσεων κυμαίνεται μεταξύ 2143 με 2846 ανά χρονιά. Από την άλλη πλευρά, η κλάση μειοψηφίας περιλαμβάνει τις χρεοκοπημένες επιχειρήσεις , οι οποίες αντιπροσωπεύουν ένα πολύ μικρό ποσοστό των συνολικών δεδομένων.

Αυτή η ανισορροπία ενέχει τον κίνδυνο, τα μοντέλα ταξινόμησης να επικεντρωθούν υπερβολικά στην κλάση πλειοψηφίας. Συγκεκριμένα, τα μοντέλα μπορούν να παρουσιάσουν υψηλά True Negatives και πολύ χαμηλά True Positives καθώς, θα προβλέπουν μόνο την πλειοψηφία (Wei and Jr, 2013a).

Υπάρχουν αρκετοί μέθοδοι για την διαχείριση της κλάσης πλειοψηφίας όπως το Random Undersampling, που είναι η τυχαία επιλογή υποσυνόλου των δεδομένων. Η τεχνική αυτή, μειώνει το πλήθος των δεδομένων της κλάσης πλειοψηφίας και τον υπολογιστικό χρόνο αλλά, έχει μειονέκτημα της απώλειας πολύτιμων πληροφορίων (Wei and Jr, 2013b).

Άλλες μέθοδοι είναι το Random Oversampling (Chawla et al., 2002, pp. 321–357), Informed Undersampling (Liu et al., 2009, pp. 965–969), Cluster-Based Sampling (Jo and Japkowicz, 2004, pp. 40–49). Στην παρούσα μελέτη επιλέχθηκε η τεχνική της τυχαίας επιλογής λόγω της απλότητάς της. Η αναλογία που δημιουργήθηκε είναι 1 χρεοκοπημένη προς 3 μη-χρεοκοπημένες επιχειρήσεις.

Cross Validation

Πριν την διαδικασία διάσπασης των δεδομένων σε folds πραγματοποιήθηκε έλεγχος null τιμών στους χρηματοοικονομικούς δείκτες. Από τα αποτελέσματα του ελέγχου δεν βρέθηκαν κελιά δεδομένων που περιέχουν null τιμές.

Η τεχνική cross validation χρησιμοποιείται στην μηχανική μάθηση για την αξιολόγηση της απόδοσης ενός μοντέλου σε δεδομένα που δεν έχει εκπαιδευτεί. Η τεχνική αυτή διασπάει το σύνολο των δεδομένων σε πολλαπλά υποσύνολα που αποκαλούνται folds. Κάθε ένα fold περιλαμβάνει ένα σύνολο δεδομένων για την εκπαίδευση του μοντέλου και ένα σύνολο εξέτασής του. Η διαδικασία εκπαίδευσης και αξιολόγησης ενός μοντέλου επαναλαμβάνεται πολλές φορές χρησιμοποιώντας ξεχωριστό fold. Τέλος, τα αποτελέσματα κάθε fold υπολογίζονται κατά μέσο όρο για να παραχθεί μια πιο ισχυρή εκτίμηση της απόδοσης του μοντέλου ("Cross Validation in Machine Learning," 2017).

Τα δεδομένα διασπάστηκαν σε 4 folds όπου το κάθε ένα περιλαμβάνει 7851 μη χρεοκοπημένες επιχειρήσεις και 186 χρεοκοπημένες. Το υποσύνολο εξέτασης περιλαμβάνει 2617 και 62 χρεοκοπημένες αντίστοιχα. Σε αυτά τα δεδομένα εφαρμόστηκε η τεχνική της τυχαίας επιλογής όπως εξηγήθηκε παραπάνω, για την δημιουργία σωστής αναλογίας μεταξύ των δύο τάξεων.

Κανονικοποίηση δεδομένων

Η κανονικοποίηση των δεδομένων είναι χρήσιμη όταν θέλουμε τα χαρακτηριστικά να έχουν ένα συγκεκριμένο εύρος όπως, από -1 έως 1. Πολλές φορές βελτιώνει την απόδοση συγκεκριμένων μοντέλων

ταξινόμησης και αποτελεί θεμελιώδης βήμα για την ορθή λειτουργία των μοντέλων όπως τα LR, SVM, NN, NB και KNN όπως αναφέρει ο Cabello-Solorzano et al. (2023).

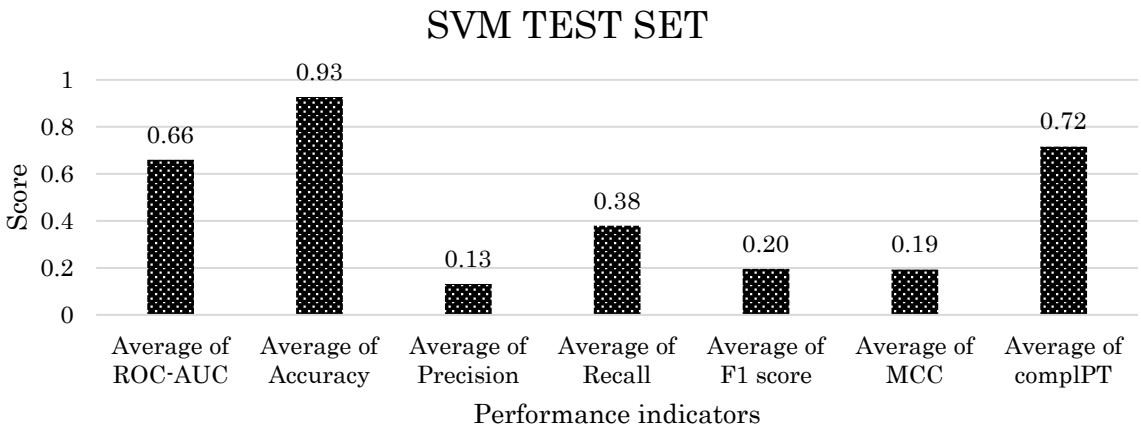
Πέραν αυτού, έχει μεγάλη σημασία ποια δεδομένα θα κανονικοποιηθούν. Άρθρα σημειώνουν ότι η κανονικοποίηση πρέπει να γίνεται ξεχωριστά για το train και το test set. Μια κανονικοποίηση στο σύνολο της βάσης θα οδηγούσε σε data leakage καθώς, μέρος του train set θα αποκτούσε πληροφορίες που κανονικά δεν θα έπρεπε να γνωρίζει (Batutin, 2024; Mucci, 2024). Για αυτόν τον λόγο, επιλέχτηκε η εφαρμογή της κανονικοποίησης ξεχωριστά για κάθε train κι test set και ξεχωριστά για κάθε fold.

Περιγραφή Τεχνικών Ταξινόμησης

Χρησιμοποιήθηκαν αλγόριθμοι ταξινόμησης από την βιβλιοθήκη scikit-learn οι οποίοι είναι: 1) Support Vector Machines, 2) Decision Trees, 3) Linear Discriminant Analysis, 4) Logistic Regression, 5) Random Forest, 6) K-Nearest Neighbors, 7) Gaussian Naïve Bayes και τέλος 8) Gradient Boosting. Για αυτούς τους αλγόριθμους δεν πραγματοποιήθηκε επιλογή παραμέτρων, παρέμειναν οι προεπιλεγμένες τιμές τους.

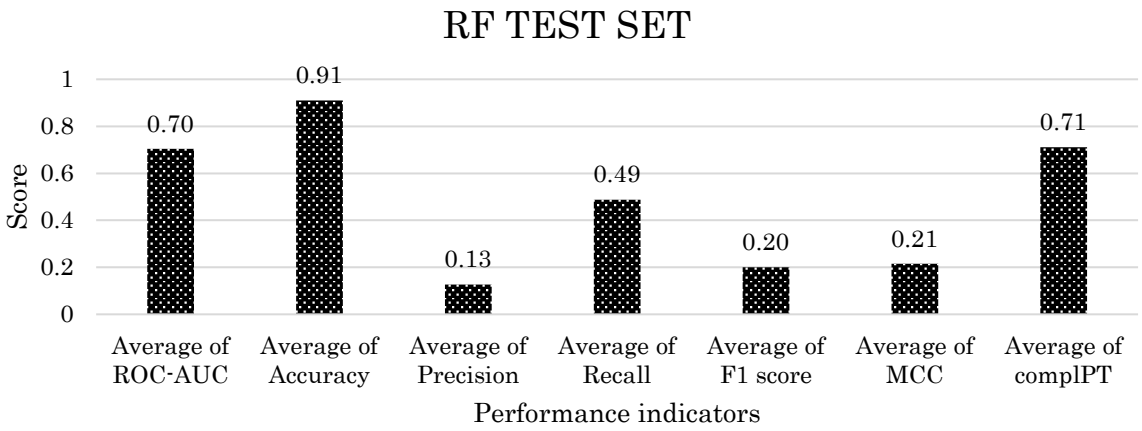
ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Αποτελέσματα Απόδοσης Μοντέλων



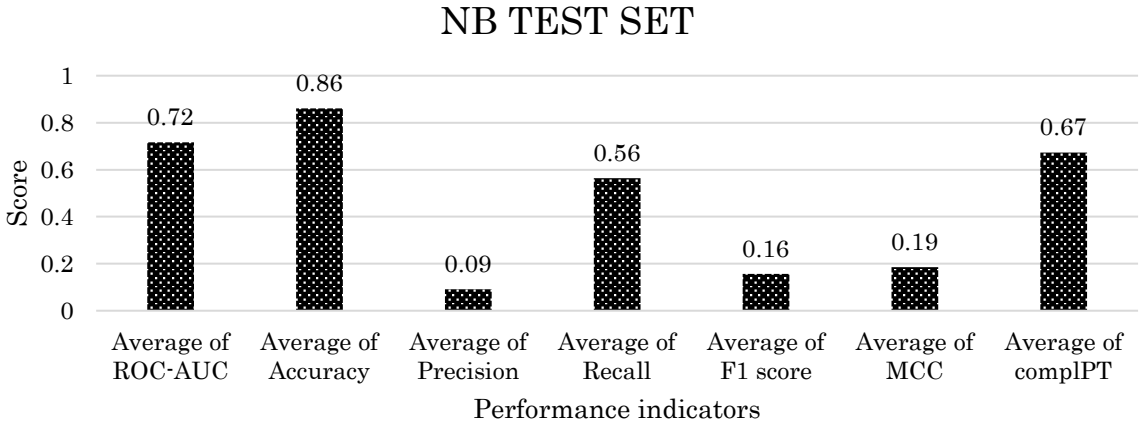
Εικόνα 2. Ραβδόγραμμα που περιγράφει την απόδοση του SVM βάσει μετρικών στο validation set

Όσον αφορά το test set, ο αλγόριθμος SVM καταγράφει τα επακόλουθα. Ο δείκτης ROC-AUC έχει τιμή 0.66, που σημαίνει ότι μοντέλο είναι καλύτερο από μια τυχαία εικασία κατά 16%. Το Accuracy καταγράφει τιμή 0.93 υποδεικνύοντας ότι το μοντέλο παράγει λίγες λανθασμένες ταξινομήσεις. Το Precision με τιμή 0.13 σημαίνει ότι μόνο το 13% του δείγματος που αναγνωρίζεται ως θετικά (χρεοκοπία) είναι πράγματι θετικά. Το Recall με τιμή κοντά στο 0.38 σημαίνει ότι το μοντέλο αναγνωρίζει μόνο το 38% των χρεοκοπημένων επιχειρήσεων. Όσο αφορά το F1, παραμένει σε τιμές κοντά στο 0.20. Με βάσει το F1, το μοντέλο δεν είναι ικανό να ταξινομεί τις κλάσεις με ακρίβεια. Για MCC με 0.19, σημαίνει ότι το μοντέλο μπερδεύει αρκετά τις σωστές ταξινομήσεις με τις λανθασμένες και δυσκολεύεται να κάνει καλό διαχωρισμό μεταξύ τους. Δείχνει πως υπάρχει συσχέτιση μεταξύ των προβλεπόμενων τάξεων, αλλά αυτή είναι αδύναμη. Τέλος, ο complPT με τιμή 0.72 σημαίνει πως, το 72% των περιπτώσεων αναμένεται να είναι αρνητικές στο σημείο ισορροπίας, όπου το μοντέλο ισορροπεί σφάλματα μεταξύ ψευδώς θετικών (FP) και ψευδώς αρνητικών (FN). Το υπόλοιπο 28% των περιπτώσεων αναμένεται να είναι θετικό για τη διατήρηση της ισορροπίας του ταξινομητή.



Εικόνα 3. Ραβδόγραμμα απόδοσης μοντέλου Random Forest ως μέσος όρος των folds του validation set.

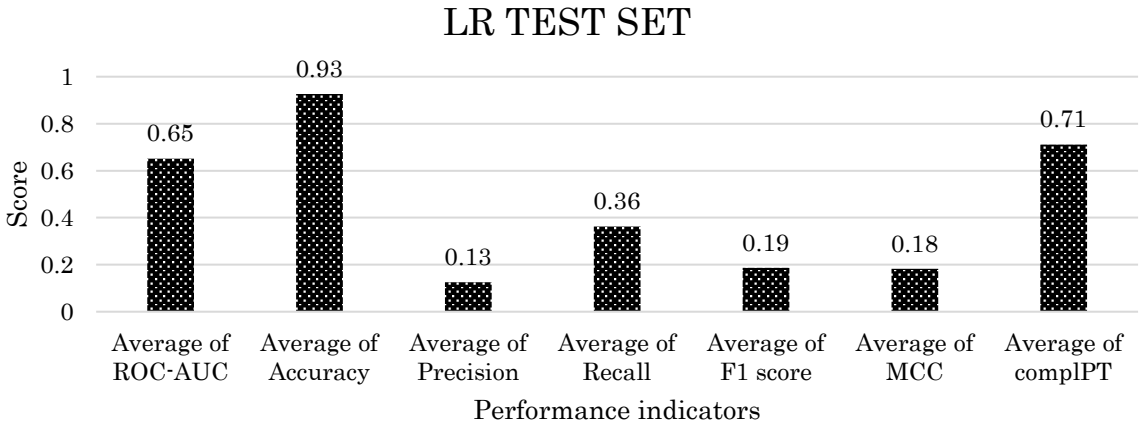
Στο παραπάνω διάγραμμα [Εικόνα 3] βλέπουμε τον μέσο όρο απόδοσης του μοντέλου Random Forest. Ξεκινώντας από τον δείκτη ROC AUC, παρατηρούμε ότι μία τιμή 0.70 δείχνει πως το μοντέλο έχει καλή ικανότητα να ξεχωρίζει τις χρεοκοπημένες εταιρείες με τις μη-χρεοκοπημένες. Το Accuracy με 0.91 δείχνει ότι το 91% των προβλέψεων ήταν σωστές. Είναι σημαντικό να επισημάνουμε ότι ο δείκτης Accuracy μόνος του δεν είναι ικανός να δείξει απόδοση του μοντέλου εξαιτίας της ανισότητας του πλήθους των δύο τάξεων. Ο δείκτης Precision δείχνει ότι μόνο το 13% των περιπτώσεων που προβλέφθηκε χρεοκοπία, όντως η εταιρεία χρεοκόπησε. Συνεχίζοντας, ο δείκτης Recall αναφέρει ότι το μοντέλο RF αναγνωρίζει 49% των πραγματικών περιπτώσεων χρεοκοπίας. Αυτό υποδηλώνει ότι το μοντέλο δυσκολεύεται στην καταγραφή θετικών περιπτώσεων. Εντούτοις, το F1 αντανακλά μία φτωχική ισορροπία μεταξύ του Precision και του Recall που δείχνει δυσκολία στην επίτευξη καλών αποτελεσμάτων και στους δύο δείκτες ταυτόχρονα. Εξάλλου, και ο δείκτης MCC δηλώνει αδυναμία στην πρόβλεψη γενικά. Τέλος, ο δείκτης Complementary Prevalence Threshold αναφέρει πως το μοντέλο είναι ρυθμισμένο για δεδομένα στα οποία το 71% περιέχουν μη χρεοκοπημένες εταιρείες, όπως είναι η αναλογία και στο σύνολο εκπαίδευσης.



Εικόνα 4. Ραβδόγραμμα απόδοσης του μοντέλου Gaussian Naive Bayes για το validation set.

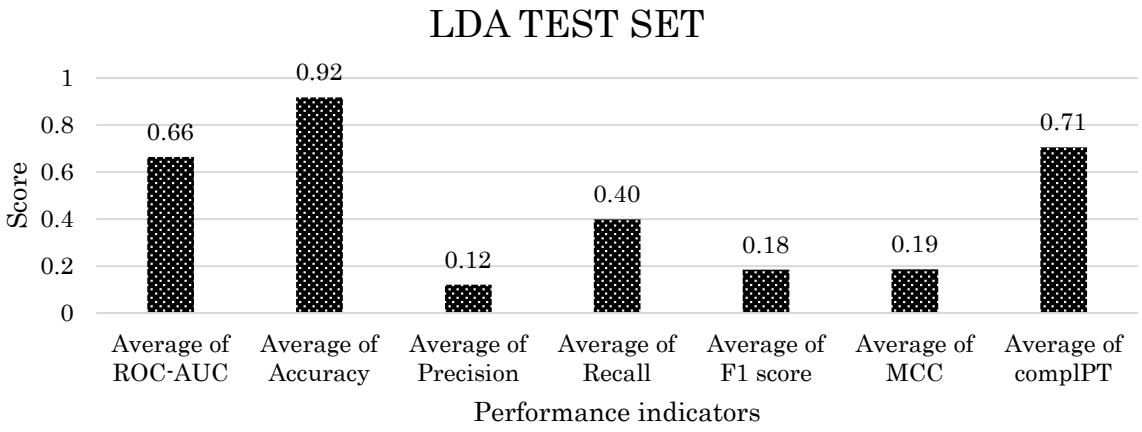
Το παραπάνω γράφημα [Εικόνα 4] δείχνει τις μετρικές απόδοσης για το μοντέλο Gaussian Naive Bayes. Ξεκινώντας από τον υψηλότερο δείκτη, το Accuracy, δείχνει ότι υπάρχουν λίγες ψευδός αρνητικές ή ψευδός αληθές ταξινομήσεις. Εξάλλου, ο δείκτης δεν είναι χρήσιμος σε άνισες κατανομές, κάτι που αποκαλύπτουν οι υπόλοιποι δείκτες. Ο δείκτης ROC-AUC, με τιμή 0.72, δείχνει ότι το μοντέλο συμπεριφέρεται κατά 22% καλύτερα από την τυχαία επιλογή. Το precision είναι πολύ χαμηλό, που σημαίνει ότι μόνο το 9% των εταιρειών που το μοντέλο πρόβλεψε ως χρεοκοπημένες, όντως χρεοκόπησαν, ενώ το Recall πετυχαίνει μόνο 56% των πραγματικών περιπτώσεων. Το F1, που είναι σε αρκετά χαμηλό επίπεδο, δείχνει τη φτωχή απόδοση τόσο του Precision όσο και του Recall. Παρόμοια, ο δείκτης MCC δείχνει αδύναμη συσχέτιση μεταξύ των προβλεπόμενων τιμών και των θετικών περιπτώσεων. Τέλος, ο δείκτης complPT με 0.67 υποδηλώνει ότι, το μοντέλο υποθέτει πως το 67% των περιπτώσεων είναι αρνητικές, το οποίο δεν

ευθυγραμμίζεται με την πραγματική κατανομή των κατηφοριών στο σύνολο της εκπαίδευσης. Αυτά τα αποτελέσματα υποδηλώνουν ότι το μοντέλο δυσκολεύεται στον εντοπισμό των θετικών και είναι πολύ προκατειλημμένο προς την τάξη της πλειοψηφίας



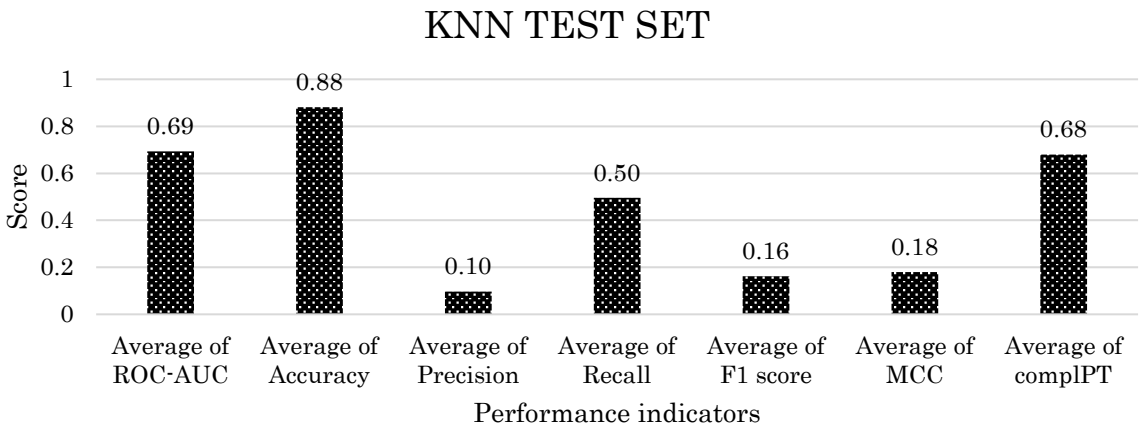
Εικόνα 5. Ραβδόγραμμα απόδοσης του μοντέλου Logistic Regression για το validation set.

Το παραπάνω διάγραμμα δείχνει τα αποτελέσματα του Logistic Regression (Εικόνα 5). Ξεκινώντας από τον ROC-AUC, δείχνει το μοντέλο έχει ικανότητα να ξεχωρίζει τις χρεοκοπημένες με τις μη χρεοκοπημένες επιχειρήσεις κατά 15% από την τυχαία επιλογή. Ο δείκτης Accuracy δείχνει ότι το 93% των ταξινομήσεων ήταν σωστές. Αντίθετα, ο δείκτης Precision φανερώνει ότι μόνο το 13% των περιπτώσεων που το μοντέλο πρόβλεψε ως χρεοκοπημένες, τελικά χρεοκόπησαν. Το Recall, δείχνει ότι το μοντέλο αναγνωρίζει το 36% των πραγματικών χρεοκοπημένων περιπτώσεων. Ο δείκτης F1 είναι χαμηλός, αντανακλώντας μια ανισορροπία μεταξύ Precision και Recall. Το MCC δείχνει ασθενή συνολική προγνωστική ισχύ και ο δείκτης complTP με τιμή 71% υποδεικνύει ότι το μοντέλο είναι ελαφριά προκατειλημμένο προς τις μη χρεοκοπημένες επιχειρήσεις.



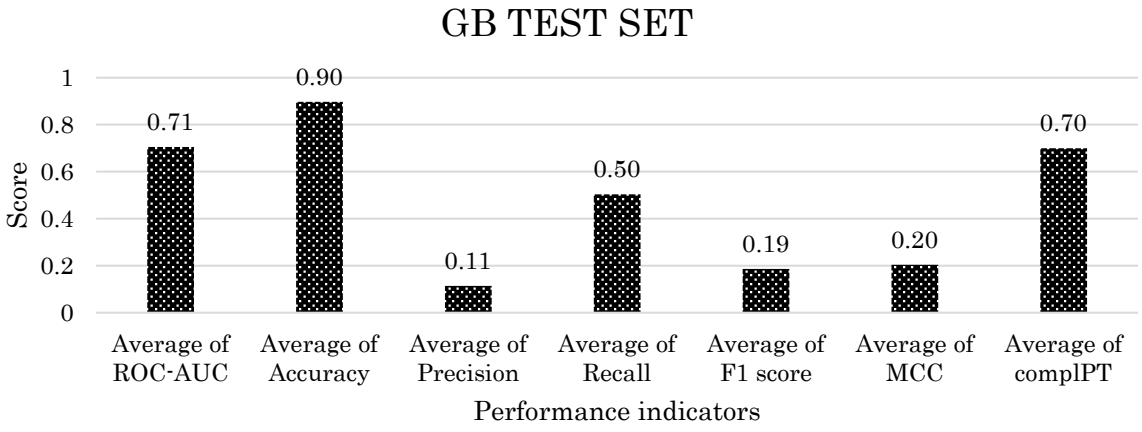
Εικόνα 6. Ραβδόγραμμα απόδοσης του μοντέλο Linear Discriminant Analysis για το validation set.

Οι μετρικές απόδοσης για το μοντέλο Linear Discriminant Analysis δείχνουν παρόμοια αποτελέσματα με του LR. Ξεκινώντας από τον ROC-AUC, με τιμή 66%, υποδεικνύεται ότι το μοντέλο είναι κατά 16% καλύτερο από την τυχαία επιλογή ταξινόμησης. Ο δείκτης Accuracy, όπως και στα προηγούμενα μοντέλα, παραμένει υψηλός λόγω της μεγάλης ανισοκατανομής των κλάσεων. Ο δείκτης Precision είναι αρκετά χαμηλός, καθώς μόνο το 12% των περιπτώσεων που το μοντέλο πρόβλεψε ως χρεοκοπία, όντως χρεοκόπησαν. Συνεχίζοντας ο δείκτης Recall που είναι εξίσου χαμηλός καθώς το μοντέλο αναγνωρίζει μόνο το 40% των πραγματικών χρεοκοπημένων περιπτώσεων. Ο δείκτης F1 είναι επίσης χαμηλός, αντανακλώντας την δυσκολία του μοντέλου στην ισόρροπη ταξινόμηση των κλάσεων. Το MCC δείχνει σημαντική αδυναμία στην ακριβή πρόβλεψη της χρεοκοπίας. Τέλος, το complPT δείχνει μια μικρή ελκυστικότητα στις μη χρεοκοπημένες.



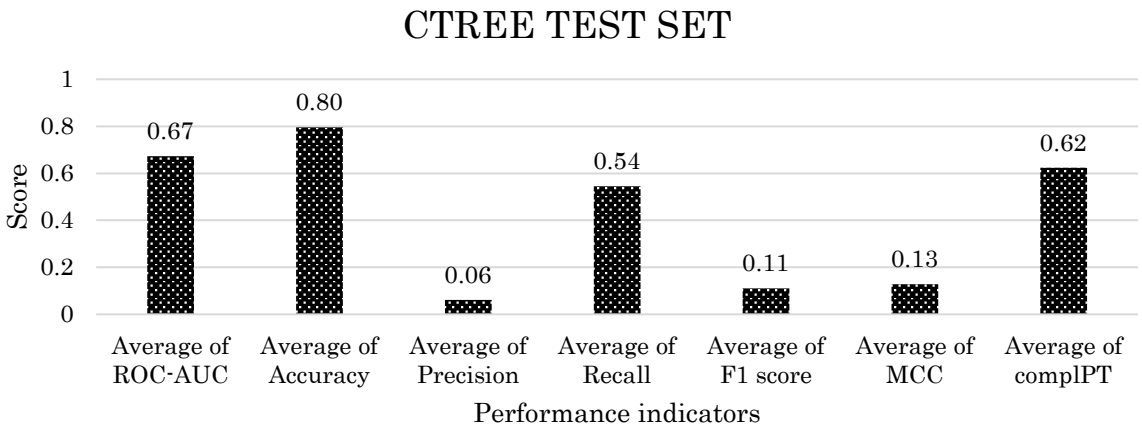
Εικόνα 7. Ραβδόγραμμα απόδοσης του μοντέλο k-Nearest Neighbors για το validation set.

Τα αποτελέσματα του μοντέλου K-Nearest Neighbors δείχνουν ότι, ο δείκτης ROC-AUC, υποδεικνύει την ικανότητα του μοντέλου να διακρίνει με 19% καλύτερα από την τυχαία επιλογή την κλάση στην οποία ανήκει κάθε εταιρεία. Ο δείκτης Accuracy είναι επίσης υψηλός, δείχνοντας ότι μοντέλο κάνει λίγα λάθη ταξινόμησης. Η μετρική Recall, δείχνει ότι το μοντέλο αναγνωρίζει το 50% των πραγματικών χρεοκοπημένων περιπτώσεων. Αντιθέτως, ο δείκτης Precision παραμένει σε χαμηλά σημεία, δείχνοντας το 10% των περιπτώσεων που το μοντέλο πρόβλεψε ως χρεοκοπημένες, όντως χρεοκόπησαν. Τα αποτελέσματα των δύο προηγούμενων μετρικών εμφανίζεται και στον δείκτη F1, που συνεχίζει να κυμαίνεται σε χαμηλά επίπεδα. Ο δείκτης MCC εμφανίζει μία χαμηλή συσχέτιση μεταξύ της πρόγνωσης και των πραγματικών αποτελεσμάτων. Τέλος, με μια τιμή 68% ο complPT αναφέρει πως, το μοντέλο είναι λίγο προκατειλημμένο προς την μη θετική τάξη.



Εικόνα 8. Ραβδόγραμμα απόδοσης του μοντέλο Gradient Boosting για το validation set.

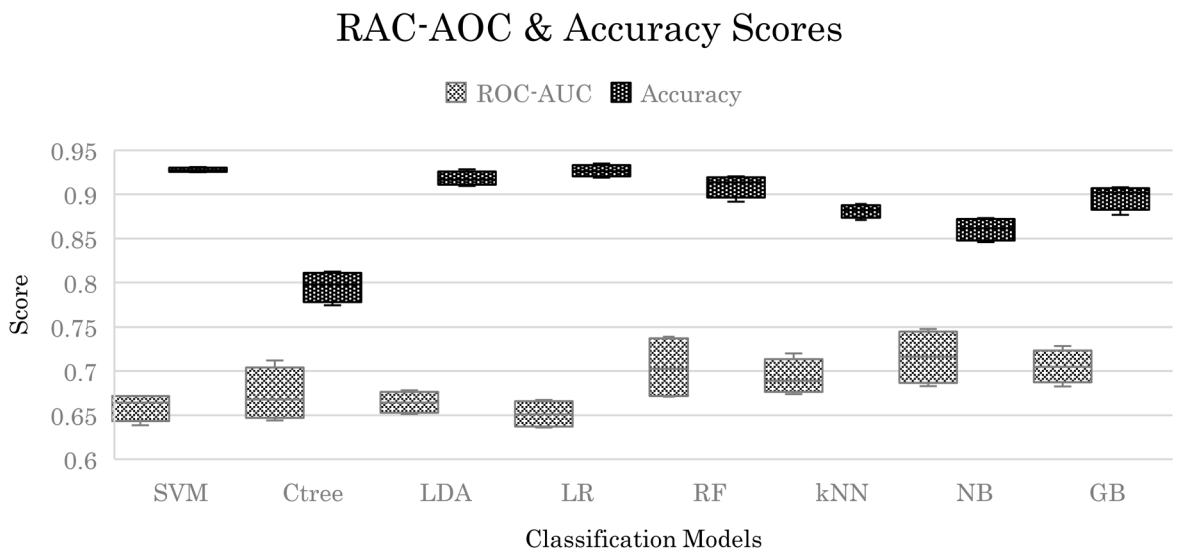
Το παραπάνω διάγραμμα δείχνει τα αποτελέσματα για το μοντέλο Gradient Boosting. Ο δείκτης ROC-AUC δείχνει ότι το μοντέλο έχει καλή ικανότητα να διακρίνει και να ταξινομεί τις τάξεις κατά 21%. Ο δείκτης Accuracy είναι υψηλός, δείχνοντας ότι υπάρχουν λίγα εσφαλμένα αποτελέσματα. Το Recall είναι μέτριο, δείχνοντας ότι το μοντέλο αναγνωρίζει το 50% των περιπτώσεων χρεοκοπίας. Ωστόσο, ο δείκτης Precision συνεχίζει να είναι χαμηλό επίπεδο δείχνοντας, μόνο το 11% της πρόγνωσης χρεοκοπίας είναι σωστό. Αυτό αντανακλάται και στον F1, που παραμένει σε χαμηλά επίπεδα. Ο δείκτης MCC δείχνει μία αδύναμη συσχέτιση με την πρόγνωση και των πραγματικών καταστάσεων. Τέλος, το complPT δείχνει μια άριστη κατανομή των δεδομένων.



Εικόνα 9. Ραβδόγραμμα απόδοσης του μοντέλο Decision Tree για το validation set.

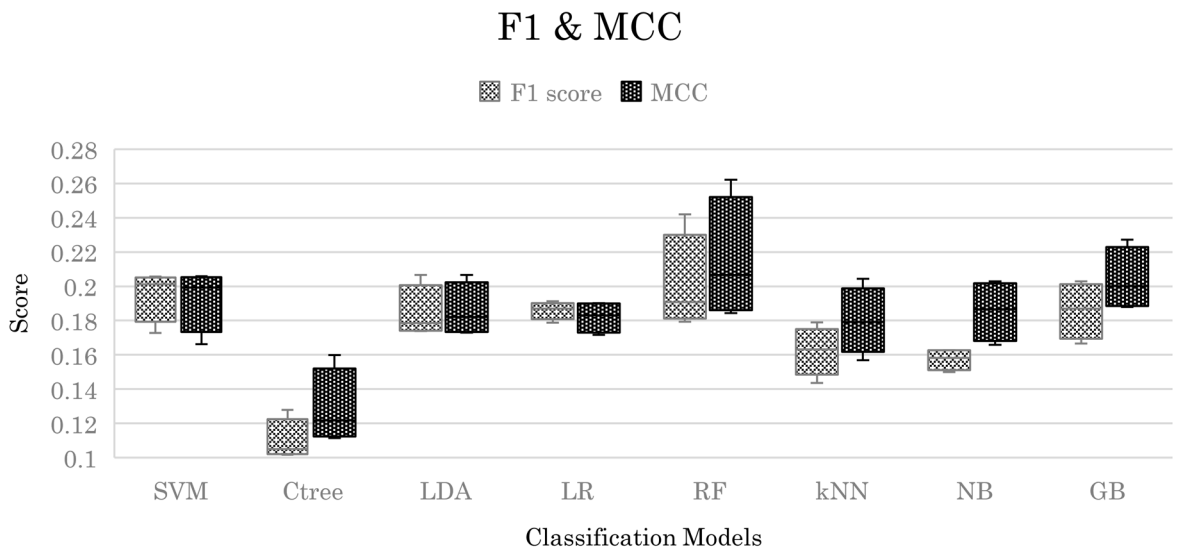
Σύμφωνα με το παραπάνω διάγραμμα απόδοσης του μοντέλου Decision Tree, παρατηρούμε τα εξής: Ο δείκτης ROC-AUC βρίσκεται σε λίγο υψηλότερα επίπεδα από την τυχαία επιλογή κατά 17%. Ο δείκτης Accuracy είναι αρκετά υψηλός, φτάνοντας το 80%. Ωστόσο, ο δείκτης Precision είναι σε πολύ χαμηλός, δείχνοντας ότι μόνο το 6% της πρόγνωσης ήταν αληθής για χρεοκοπία. Ενώ ο δείκτης Recall λέει πως το μοντέλο αναγνωρίζει μόνο το 54% των πραγματικών χρεοκοπημένων περιπτώσεων. Αυτό αντανακλάται και στον δείκτη F1, που είναι και αρκετά χαμηλός. Ο δείκτης MCC λέει ότι η συσχέτιση μεταξύ των προβλέψεων και των πραγματικών καταστάσεων είναι αδύναμη. Τέλος, ο δείκτης complPT αναφέρει το μοντέλο είναι προκατειλημμένο προς την μη θετική τάξη.

Σύγκριση Αποτελεσμάτων



Εικόνα 10. Σύγκριση των Accuracy και RAC-AUC.

Στο παραπάνω γράφημα [Εικόνα 10] παρατηρούμε πως τα μοντέλα RF κι NB έχουν τις υψηλότερες τιμές στον δείκτη ROC-AUC κατά μέσο όρο, συνοδευόμενα με αρκετά καλή απόδοση από τον δείκτη Accuracy. Τα μοντέλα αυτά είναι ικανά να διακρίνουν τις χρεοκοπημένες με τις μη χρεοκοπημένες σε καλό βαθμό. Όσον αφορά τα υπόλοιπα μοντέλα, ο δείκτης ROC-AUC πέφτει σε χαμηλά επίπεδα κοντά στο 68%. Με το χειρότερο να είναι τα μοντέλα SVM, Ctree, LDA και LR λόγω, τις μεγάλης επικάλυψης των τιμών στο διάγραμμά. Συνεχίζοντας στον δείκτη Accuracy παρατηρούμε πως, το SVM έχει την χαμηλότερη αστοχία σε miss-classification με την μικρότερη διασπορά υποδεικνύοντας, σταθερά αποτελέσματα στην μετρική αυτή. Χωρίς να μας λέει για το πως πάει με τις χρεοκοπημένες, σχετικά ο δείκτης Accuracy παραμένει σε υψηλά επίπεδα.



Εικόνα 11. Σύγκριση των F1 και Matthews correlation coefficient score.

Με μία πρώτη ματιά, το RF φαίνεται να είναι το καλύτερο βάσει του F1 σκορ από την μέγιστη τιμή του αλλά με μεγάλη διακύμανση υποδεικνύοντας μη σταθερή απόδοση μεταξύ των folds. Για τα υπόλοιπα μοντέλα, το F1 σκορ μένει σε τιμές μεταξύ 0.14 - 0.21 εκτός του Ctree που καταγράφει στατιστικά την χειρότερη απόδοση. Παρόμοια αποτελέσματα παρατηρούνται και για το σκορ MCC. Τα μοντέλα το SVM, LDA, LR, RF και το GB συνεχίζουν να βρίσκονται σε παρόμοιες τιμές. Αν και το μοντέλο RF έχει τις υψηλότερες τιμές τα προηγούμενα μοντέλα παρουσιάζουν σταθερότερα αποτελέσματα.

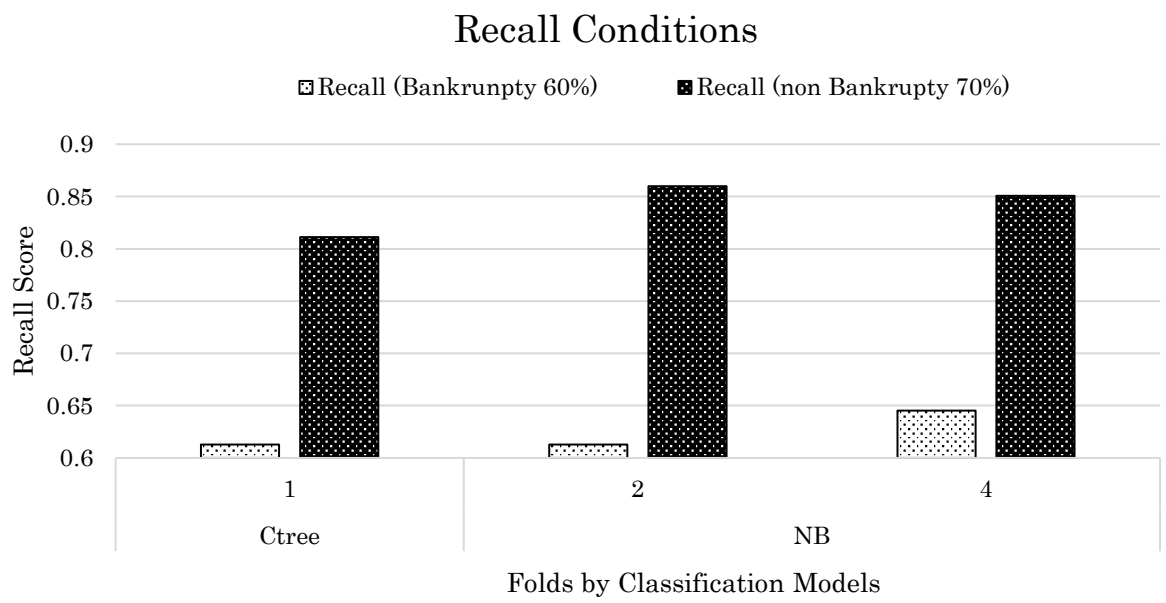
ΣΥΜΠΕΡΑΣΜΑΤΑ

Κύρια Συμπεράσματα

Συνοψίζοντας, σκοπός του παρόντος εγγράφου ήταν η ανάλυση και ο εντοπισμός του καλύτερου μοντέλου ταξινόμησης για το συγκεκριμένο πρόβλημα. Έγινε αναφορά στα δεδομένα του προβλήματος και πως αυτά επηρεάζουν την απόδοση των μοντέλων μέσω, ανισοκατανομής των τάξεων και της κανονικοποίησης. Παρουσιάστηκαν οι παράμετροι που βελτιώνουν την απόδοση και τέλος, εκπαιδεύτηκαν και απέδωσαν τα παραπάνω αποτελέσματα που καταγράφονται στις προηγούμενες ενότητες.

Για την εκτίμηση της απόδοσης κάθε μοντέλου χρησιμοποιήθηκαν διάφορες μετρικές απόδοσης, όπως οι δείκτες ROC-AUC, Accuracy, Precision, Recall, F1, MCC και τέλος τον δείκτη complPT. Οι μετρικές Accuracy και F1 αποτελούν τις πιο γνωστές μετρικές αξιολόγησης της απόδοσης ενός μοντέλου ταξινόμησης. Χάρη στην μετρική F1, το μοντέλο RF αποδείχθηκε πως ήταν το καλύτερο λόγω των υψηλότερων τιμών. Από την άλλη μεριά, τα Accuracy και F1 μπορούν να παράγουν παραπλανητικά αποτελέσματα σε άνισες τάξεις. Για αυτόν τον λόγο, προτείνεται από τους Chicco και Jurman (2020), η χρήση του δείκτη MCC καθώς, για την απόκτηση μεγάλου σκορ το μοντέλο πρέπει να κάνει, σωστές προγνώσεις τόσο για τις αρνητικές τάξεις όσο και για τις θετικές, ανεξαρτήτως την αναλογία τους. Όπου, και πάλι κερδίζει το RF καθώς, έχουν το υψηλότερο σκορ με μια μέση απόδοση παρόμοια των SVM και GB αλλά με την ικανότητα υψηλότερων τιμών.

Προϋποθέσεις Απόδοσης



Εικόνα 12. Ραβδόγραμμα Recall των εταιρειών που θα πτωχεύσουν κατά 60% και αυτών που θα επιβιώσουν κατά 70%. Επιλέχτηκαν μονάχα τα μοντέλα που εκπληρώνουν τους δύο αυτούς περιορισμούς.

Βάσει των προκαθορισμένων προϋποθέσεων για τα μοντέλα, στόχος ήταν: 1) Αναγνωρίζουν το 60% των χρεοκοπημένων περιπτώσεων και 2) Να αναγνωρίζουν το 70% των περιπτώσεων που δεν οδηγούν σε χρεοκοπία. Με βάση το διάγραμμα [Εικόνα 12] παρατηρούμε όλα τα μοντέλα που ανταποκρίνονται σε αυτές τις απαιτήσεις. Οι αριθμοί 1, 2 και 4 αντιπροσωπεύουν τα folds του test set.

Τα μοντέλα Decision Trees και Naive Bayes (NB) εκπληρώνουν τις απαιτήσεις που τέθηκαν. Αν και τα Ctree ικανοποιεί και τις δύο προϋποθέσεις, δεν προτείνονται λόγω της χαμηλής απόδοσης στις μετρικές F1 και MCC. Αντίθετα, το μοντέλο NB καταγράφει καλύτερες αποδόσεις στις μετρικές Recall και υπερέχει στατιστικά του Ctree, όπως προκύπτει από τις μετρήσεις F1 και MCC.

Table 1. Confusion Matrix για το μοντέλο NB στο fold 4.

GB	**0**	**1**
0	2226	391
1	22	40

Παρατηρώντας τον Πίνακας 2, το μοντέλο NB αναδείχθηκε ως το καλύτερο βάσει των απαιτήσεων που τέθηκαν. Αναλύοντας τον Confusion Matrix, παρατηρείται ότι το μοντέλο εντόπισε σωστά 2226 περιπτώσεις όπου δεν υπήρξε χρεοκοπία και αναγνώρισε 40 από τις περιπτώσεις χρεοκοπίας. Ωστόσο, 22 περιπτώσεις χρεοκοπίας δεν ανιχνεύτηκαν, ενώ 391 περιπτώσεις αναγνωρίστηκαν λανθασμένα ως χρεοκοπίες. Συνολικά, το NB πληροί τις προϋποθέσεις, εντοπίζοντας σωστά το 60% των χρεοκοπημένων περιπτώσεων και το 70% των περιπτώσεων που δεν θα χρεοκοπήσουν.

Συστάσεις

Για τη βελτίωση της απόδοσης των μοντέλων, θα μπορούσαν να εφαρμοστούν πιο εξελιγμένοι μέθοδοι δειγματοληψίας, όπως το informed undersampling(Liu et al., 2009, pp. 965–969) και το cluster-based sampling (Jo and Japkowicz, 2004, pp. 40–49), σε αντίθεση με την τυπική τυχαία δειγματοληψία (Random Sample). Όπως αποδείχθηκε στην ανάλυση των αποτελεσμάτων με τη μετρική complPT, μια βελτιωμένη κατανομή μεταξύ των τάξεων θα μπορούσε να βελτιώσει την απόδοση των μοντέλων Logistic Regression (LR), Naive Bayes (NB) και Classification Trees (CTree). Αυτοί οι μέθοδοι θα μπορούσαν να βοηθήσουν στην επίτευξη καλύτερης ισορροπίας στις τάξεις, επιτρέποντας στα μοντέλα να κάνουν πιο ακριβείς προβλέψεις

ΒΙΒΛΙΟΓΡΑΦΙΑ

Batutin, A., 2024. Data Leakage in Machine Learning Models [WWW Document]. Shelf. URL https://shelf.io/blog/preventing-data-leakage-in-machine-learning-models/ (accessed 11.18.24).

Cabello-Solorzano, K., Ortigosa de Araujo, I., Peña, M., Correia, L., J. Tallón-Ballesteros, A., 2023. The Impact of Data Normalization on the Accuracy of Machine Learning Algorithms: A Comparative Analysis, in: García Bringas, P., Pérez García, H., Martínez de Pisón, F.J., Martínez Álvarez, F., Troncoso Lora, A., Herrero, Á., Calvo Rolle, J.L., Quintián, H., Corchado, E. (Eds.), 18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023). Springer Nature Switzerland, Cham, pp. 344–353. https://doi.org/10.1007/978-3-031-42536-3_33

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953

Chicco, D., Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21, 6. https://doi.org/10.1186/s12864-019-6413-7

Cross Validation in Machine Learning [WWW Document], 2017. . GeeksforGeeks. URL https://www.geeksforgeeks.org/cross-validation-machine-learning/ (accessed 12.14.24).

Jo, T., Japkowicz, N., 2004. Class imbalances versus small disjuncts. SIGKDD Explor Newsl 6, 40–49. https://doi.org/10.1145/1007730.1007737

Liu, X.-Y., Wu, J., Zhou, Z.-H., 2009. Exploratory Undersampling for Class-Imbalance Learning. IEEE Trans. Syst. Man Cybern. Part B Cybern. 39, 539–550. https://doi.org/10.1109/TSMCB.2008.2007853

Mucci, T., 2024. What is Data Leakage in Machine Learning? | IBM [WWW Document]. URL https://www.ibm.com/think/topics/data-leakage-machine-learning (accessed 11.18.24).

Shi, Y., Li, X., 2019. A bibliometric study on intelligent techniques of bankruptcy prediction for corporate firms. Heliyon 5, e02997. https://doi.org/10.1016/j.heliyon.2019.e02997

Wei, Q., Jr, R.L.D., 2013a. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. PLOS ONE 8, e67863. https://doi.org/10.1371/journal.pone.0067863

Wei, Q., Jr, R.L.D., 2013b. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. PLOS ONE 8, e67863. https://doi.org/10.1371/journal.pone.0067863