Ελένη Γεωργούδη – ics23103

Όλγα Σαρίδου – ics23078

Βαλεντίνος Γούδας – ics23170

Θρασύβουλος Βασδαβάνος – ics23168

Bankruptcy Prediction using ML Classification

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ Applied Informatics UOM 2025-26

TABLE OF CONTENTS

1. Θεωρητικό Υπόβαθρο 3
2. Εισαγωγή 4
3. Μεθοδολογία 6
   3.1. Επισκόπηση και Προεπεξεργασία Δεδομένων 6
   3.2. Μοντέλα Ταξινόμησης 12
      3.2.1. K-Nearest Neighbors (Knn) 13
      3.2.2. Naïve Bayes (Nb) 13
      3.2.3. Decision Trees (Dts) 13
      3.2.4. Random Forest (RF) 13
      3.2.5. Linear Discriminant Analysis (LDA) 13
      3.2.6. Support Vector Machines (SVM) 13
      3.2.7. Logistic Regression (LR) 13
      3.2.8. XGBoost 13
   3.3 Χρονική Πολυπλοκότητα των Μοντέλων 14
   3.4. Προσδιορισμός Βέλτιστου Μοντέλου 16
   3.5. Συγκριτική Αξιολόγηση και Τελικά Συμπεράσματα 16
4. Συζήτηση 16
   4.1. Προκλήσεις και Περιορισμοί Μοντέλων 16
   4.2. Μελλοντική Έρευνα 16
5. Συμπεράσματα 16
6. Σχετική Βιβλιογραφία 17

1. Θεωρητικό Υπόβαθρο

Στο πεδίο της διαχείρισης χρηματοοικονομικών κινδύνων, η πρόβλεψη πτώχευσης αποτελεί θεμελιώδη διαδικασία. Η εξέλιξη των μοντέλων πρόβλεψης καθορίστηκε σε μεγάλο βαθμό από την πολυπλοκότητα των οικονομικών δεδομένων, οδηγώντας σε μια μετάβαση από τις κλασικές στατιστικές μεθόδους προς τις σύγχρονες τεχνικές Μηχανικής Μάθησης.

Ιστορικά, η χρήση μεθόδων όπως η λογιστική παλινδρόμηση και η διακριτική ανάλυση, αν και θεμελιώδης, αποδείχθηκε ανεπαρκής στην αποτύπωση των μη γραμμικών σχέσεων των δεδομένων. Ως απάντηση, υιοθετήθηκαν αλγόριθμοι όπως τα Support Vector Machines (SVMs), τα Random Forests, τα Νευρωνικά Δίκτυα και οι μηχανές Gradient Boosting, οι οποίοι διακρίνονται για την ικανότητά τους να διαχειρίζονται μεγάλα σύνολα δεδομένων και να εντοπίζουν σύνθετα μοτίβα.

Ωστόσο, η διαδικασία αυτή αντιμετωπίζει σημαντικές προκλήσεις. Η ανισορροπία των κλάσεων (class imbalance), δεδομένου ότι οι πτωχεύσεις είναι σπάνια γεγονότα, αντιμετωπίζεται με μεθόδους όπως το SMOTE και εξειδικευμένες συναρτήσεις απώλειας. Παράλληλα, η επιλογή μεταβλητών (feature selection) μέσω τεχνικών όπως η παλινδρόμηση Lasso, ενισχύει την ερμηνευσιμότητα και την ακρίβεια, απομονώνοντας τους πιο σημαντικούς παράγοντες κινδύνου.

Για την αξιολόγηση των μοντέλων, χρησιμοποιούνται συγκεκριμένοι δείκτες:

- Η Ακρίβεια (Accuracy) παρέχει μια γενική εικόνα της ορθότητας του μοντέλου.
- Η Ακρίβεια Θετικών Προβλέψεων (Precision) διασφαλίζει ότι δεν χαρακτηρίζονται εσφαλμένα ως πτωχευμένες οι υγιείς εταιρείες.
- Η Ευαισθησία (Recall) είναι κρίσιμη για τον έγκαιρο εντοπισμό των εταιρειών που βρίσκονται πραγματικά σε κίνδυνο.
- Το F1 Score, ως ο αρμονικός μέσος των δύο προηγούμενων, εξασφαλίζει την ισορροπία ώστε να αποφεύγονται τόσο οι άσκοπες παρεμβάσεις όσο και η αγνόηση πραγματικών κινδύνων.
- Τέλος, το AUC-ROC αξιολογεί τη διαχωριστική ικανότητα του μοντέλου σε διαφορετικά κατώφλια απόφασης, προσφέροντας μια συνολική εικόνα της προβλεπτικής του ισχύος.

2. Εισαγωγή

Μεθοδολογική Προσέγγιση και Αλγόριθμοι Πρόβλεψης Πτώχευσης

Στην παρούσα διατριβή, το ερευνητικό ενδιαφέρον εστιάζεται στην εφαρμογή της Μηχανικής Μάθησης για την πρόβλεψη της εταιρικής πτώχευσης, ένα πεδίο που συγκεντρώνει αυξανόμενη προσοχή στη χρηματοοικονομική ανάλυση και τη διαχείριση κινδύνων. Κεντρική παραδοχή της μελέτης αποτελεί το γεγονός ότι η αποτελεσματικότητα της πρόβλεψης δεν εξαρτάται αποκλειστικά από τους αλγορίθμους που χρησιμοποιούνται, αλλά εξίσου από την ποιότητα και τη συνάφεια των χρηματοοικονομικών δεδομένων.

Εξεταζόμενα Μοντέλα Μηχανικής Μάθησης

Η μελέτη αξιολογεί την προβλεπτική ικανότητα μιας ευρείας γκάμας μοντέλων ταξινόμησης, καθένα από τα οποία παρουσιάζει ξεχωριστά πλεονεκτήματα:

- Λογιστική Παλινδρόμηση (Logistic Regression): Χρησιμοποιείται ως μοντέλο αναφοράς (baseline), εκτιμώντας την πιθανότητα πτώχευσης μέσω της λογιστικής συνάρτησης.
- Γραμμική Διακριτική Ανάλυση (LDA): Ο αλγόριθμος αυτός βασίζεται στο θεώρημα του Bayes και υποθέτει ότι τα δεδομένα κάθε κλάσης προέρχονται από κατανομή Gauss με κοινό πίνακα συνδιακύμανσης, επιδιώκοντας τον βέλτιστο γραμμικό διαχωρισμό των κλάσεων.
- Naive Bayes (GaussianNB): Ένας πιθανοτικός ταξινομητής που επίσης βασίζεται στο θεώρημα του Bayes. Διακρίνεται για την υπολογιστική του ταχύτητα, με την ισχυρή (απλοποιητική) παραδοχή ότι τα χαρακτηριστικά είναι ανεξάρτητα μεταξύ τους.
- Κ-Πλησιέστεροι Γείτονες (KNN): Μια μέθοδος που ταξινομεί τα δεδομένα βάσει της πλειοψηφίας των πλησιέστερων γειτόνων τους. Αν και διακρίνεται για την απλότητά της και την απουσία υποθέσεων για την κατανομή των δεδομένων, παρουσιάζει ευαισθησία στην επιλογή του αριθμού των γειτόνων (k).
- Μηχανές Διανυσμάτων Υποστήριξης (SVM): Χρησιμοποιούνται για τον διαχωρισμό των κλάσεων μέσω υπερεπιπέδων. Η συμπερίληψή τους οφείλεται στην ισχυρή τους ικανότητα ταξινόμησης, ιδιαίτερα σε χώρους υψηλών διαστάσεων.

- Δέντρα Απόφασης (Decision Trees): Μοντέλα που δημιουργούν κανόνες απόφασης βασισμένους στα χαρακτηριστικά των δεδομένων, προσφέροντας υψηλή ερμηνευσιμότητα.
- Τυχαία Δάση (Random Forests): Ως μέθοδος συνδυαστικής μάθησης (ensemble learning), συνθέτει πολλαπλά δέντρα απόφασης για τη βελτίωση της ακρίβειας και την αποφυγή της υπερεκπαίδευσης.
- XGBoost (Extreme Gradient Boosting): Μια προηγμένη υλοποίηση αλγορίθμων ενίσχυσης κλίσης (gradient boosting). Το μοντέλο αυτό επιλέχθηκε για την υψηλή του απόδοση και την ικανότητά του να διαχειρίζεται αποτελεσματικά πολύπλοκα δεδομένα, αποτελώντας συχνά την κορυφαία επιλογή σε προβλήματα δομημένων δεδομένων.

Διαδικασία Επικύρωσης (Validation)

Ένα κρίσιμο μεθοδολογικό στοιχείο είναι η χρήση της στρωματοποιημένης διασταυρούμενης επικύρωσης (stratified cross-validation). Η τεχνική αυτή διασφαλίζει ότι κάθε υποσύνολο εκπαίδευσης και ελέγχου αντικατοπτρίζει με ακρίβεια την αναλογία πτωχευμένων και υγιών εταιρειών του συνολικού δείγματος, ενισχύοντας την αξιοπιστία της γενίκευσης του μοντέλου.

Δείκτες Αξιολόγησης (Metrics)

Για την ολοκληρωμένη εκτίμηση της απόδοσης, η μελέτη εξετάζει τους εξής δείκτες:

- Ακρίβεια (Accuracy): Η γενική ορθότητα του μοντέλου.
- Precision (Ακρίβεια Θετικών Προβλέψεων): Το ποσοστό των προβλέψεων πτώχευσης που ήταν ορθές, ελαχιστοποιώντας τα ψευδώς θετικά αποτελέσματα.
- Recall (Ανάκληση): Το ποσοστό των πραγματικών πτωχεύσεων που εντοπίστηκαν σωστά.
- F1 Score: Ο αρμονικός μέσος των Precision και Recall, που αποτελεί βασικό μέτρο ακρίβειας σε ανισοβαρή δεδομένα.
- Καμπύλη ROC και AUC: Η μελέτη της διακριτικής ικανότητας του μοντέλου σε διάφορα κατώφλια απόφασης.

Στόχος και Συνεισφορά

Μέσα από τη συγκριτική ανάλυση των παραπάνω τεχνικών, η έρευνα στοχεύει στον εντοπισμό των βέλτιστων μεθόδων πρόβλεψης πτώχευσης, προσφέροντας πολύτιμα εργαλεία για τη λήψη τεκμηριωμένων αποφάσεων.

3. Μεθοδολογία

Ανάλυση Δεδομένων και Μεθοδολογία Υλοποίησης

Η ενότητα αυτή ξεκινά με την ανάλυση του συνόλου δεδομένων που χρησιμοποιήθηκε, συμπεριλαμβανομένων των πηγών και των βασικών χαρακτηριστικών του (δείκτες απόδοσης, δυαδικοί δείκτες δραστηριότητας, κατάσταση εταιρείας και έτος αναφοράς). Ακολουθεί η περιγραφή των βημάτων προεπεξεργασίας των δεδομένων (data pre-processing), όπως η διαχείριση των ελλειπουσών τιμών και η κανονικοποίηση των δεδομένων.

Στη συνέχεια, υλοποιούνται, αξιολογούνται και βελτιστοποιούνται τα ακόλουθα μοντέλα ταξινόμησης: K-Nearest Neighbors (KNN), Δέντρα Απόφασης (Decision Trees), Τυχαία Δάση (Random Forest), Naive Bayes, Λογιστική Παλινδρόμηση (Logistic Regression), SVM, LDA και XGBoost.

Η αξιολόγηση των παραπάνω μοντέλων πραγματοποιείται μέσω των μετρικών: Ακρίβεια (Accuracy), Precision, Recall, F1 score και Εμβαδόν κάτω από την καμπύλη ROC (AUC-ROC).

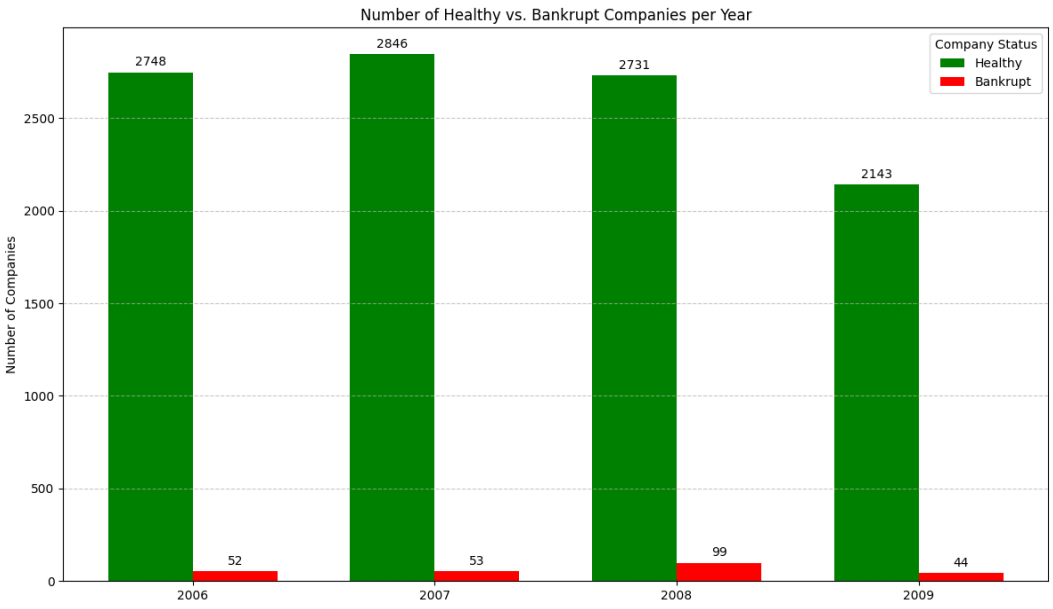
3.1. Επισκόπηση και Προεπεξεργασία Δεδομένων

Η διαδικασία της ανάλυσης ξεκινά με τη φόρτωση του συνόλου δεδομένων (dataset) στο περιβάλλον της Python, χρησιμοποιώντας τη βιβλιοθήκη Pandas. Η δημιουργία του DataFrame (με την ονομασία df) αποτελεί το θεμελιώδες βήμα για την περαιτέρω διαχείριση των πληροφοριών. Για την απόκτηση μιας πρώτης εικόνας της δομής και του περιεχομένου των δεδομένων, χρησιμοποιείται η εντολή display(df.head()), η οποία μας επιτρέπει να εξετάσουμε ένα αντιπροσωπευτικό δείγμα των εγγραφών.

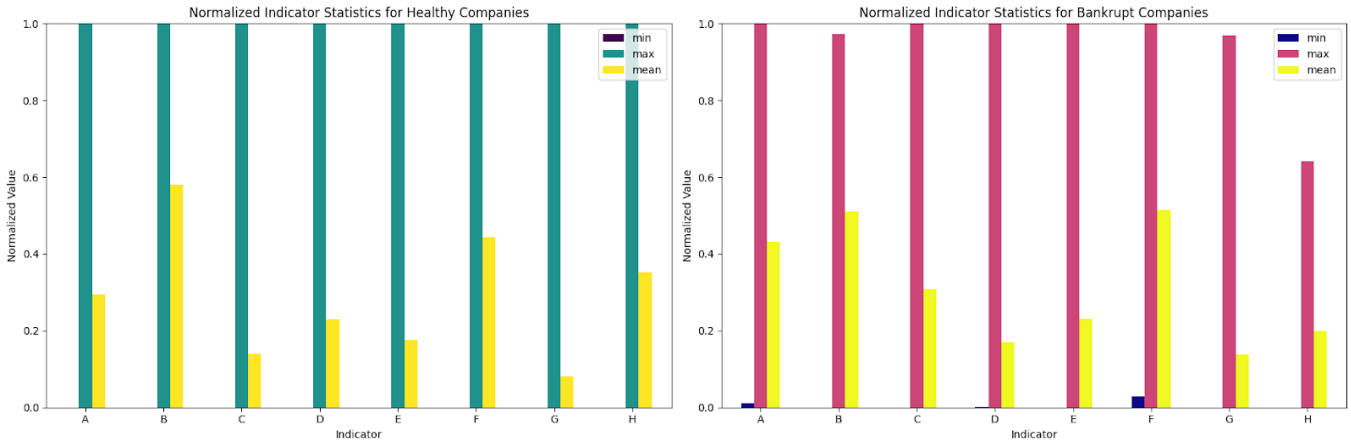
Στο στάδιο αυτό, κρίθηκε απαραίτητη η μετονομασία των στηλών του αρχείου (π.χ. σε 'Status', 'Year', 'A', 'B' κ.λπ.) προκειμένου να απλοποιηθεί η αναφορά σε αυτές κατά τη συγγραφή του κώδικα. Παράλληλα, διενεργείται ενδελεχής έλεγχος για την ύπαρξη ελλιπών τιμών (missing values), διασφαλίζοντας την ακεραιότητα και την ποιότητα των δεδομένων πριν από οποιαδήποτε στατιστική επεξεργασία.

Στη συνέχεια, εφαρμόζεται η τεχνική της κανονικοποίησης (normalization) με τη χρήση του αλγορίθμου **MinMaxScaler** από τη βιβλιοθήκη Scikit-Learn. Η διαδικασία αυτή κλιμακώνει τις τιμές των χαρακτηριστικών στο διάστημα [0, 1], εξαλείφοντας τις διαφορές μεγέθους μεταξύ των μεταβλητών που θα μπορούσαν να επηρεάσουν αρνητικά την απόδοση των μοντέλων μηχανικής μάθησης. Από την κανονικοποίηση εξαιρούνται σκόπιμα συγκεκριμένες στήλες, όπως το έτος ('Year') και η κατηγορική μεταβλητή της κατάστασης ('Status'), καθώς η φύση τους δεν απαιτεί τέτοιου είδους μετασχηματισμό.

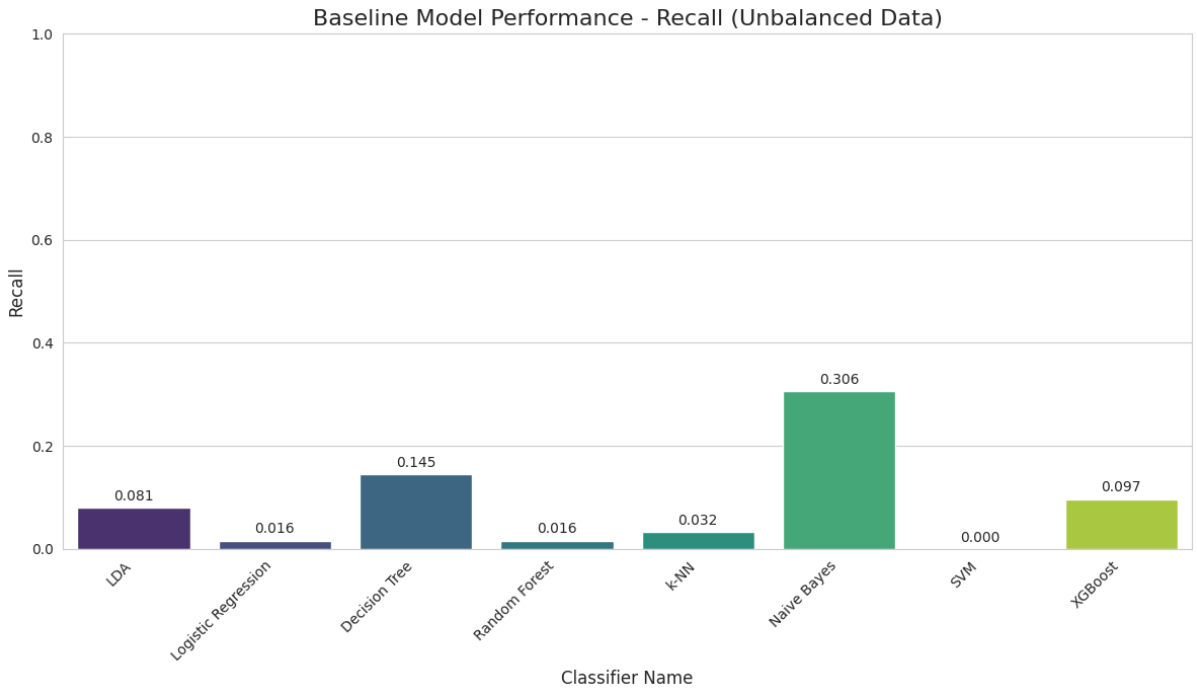
Ένα κρίσιμο κομμάτι της προεπεξεργασίας είναι η κατανόηση της κατανομής των δεδομένων. Για τον σκοπό αυτό, δημιουργείται ένα **ραβδόγραμμα (bar plot)** που απεικονίζει το πλήθος των υγιών (Healthy) έναντι των πτωχευμένων (Bankrupt) εταιρειών ανά έτος. Η οπτικοποίηση αυτή, όπως παρουσιάζεται στο **Figure 1**, αναδεικνύει ξεκάθαρα την έντονη ανισορροπία των κλάσεων (class imbalance), καθώς οι υγιείς εταιρείες υπερτερούν αριθμητικά σε σημαντικό βαθμό.

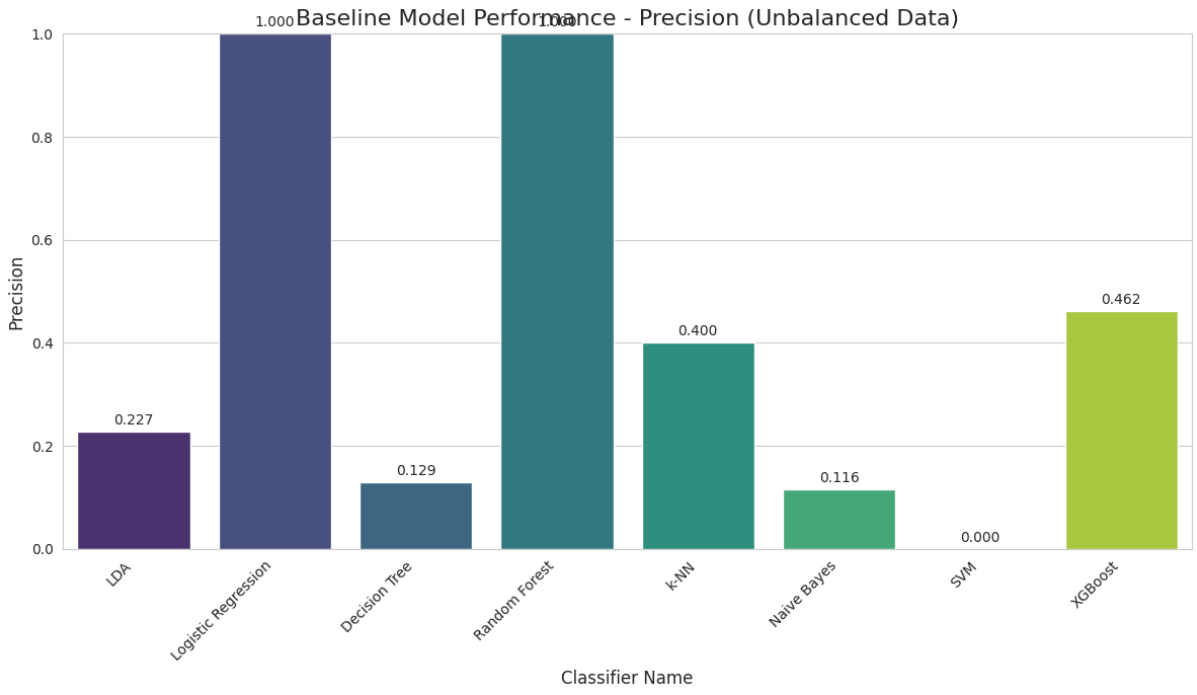


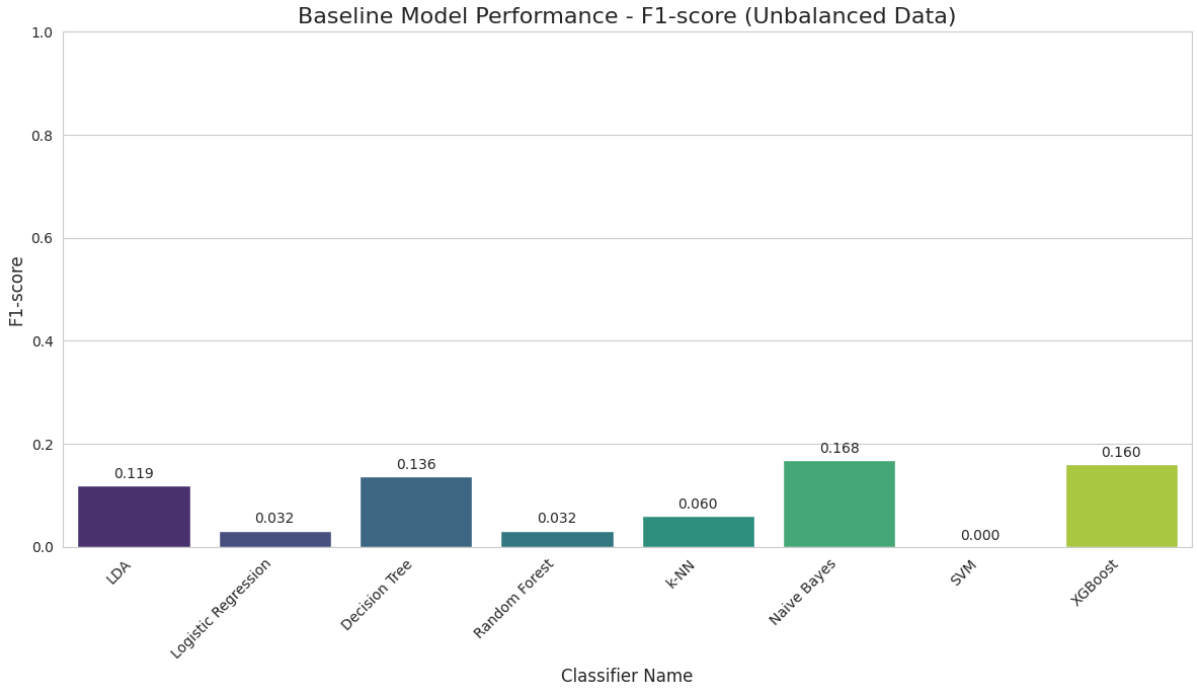
Τέλος, για τη βαθύτερη διερεύνηση των χαρακτηριστικών, ο κώδικας παράγει συγκριτικά διαγράμματα (**Figure 2**) που αντιπαραβάλλουν τα στατιστικά στοιχεία (Ελάχιστο, Μέγιστο, Μέσος Όρος) των κανονικοποιημένων δεικτών για τις δύο κατηγορίες εταιρειών. Μέσω αυτών των γραφημάτων, καθίσταται δυνατός ο εντοπισμός μοτίβων και διαφορών στη χρηματοοικονομική συμπεριφορά των υγιών και των πτωχευμένων επιχειρήσεων, παρέχοντας πολύτιμες πληροφορίες για την εκπαίδευση των αλγορίθμων πρόβλεψης.

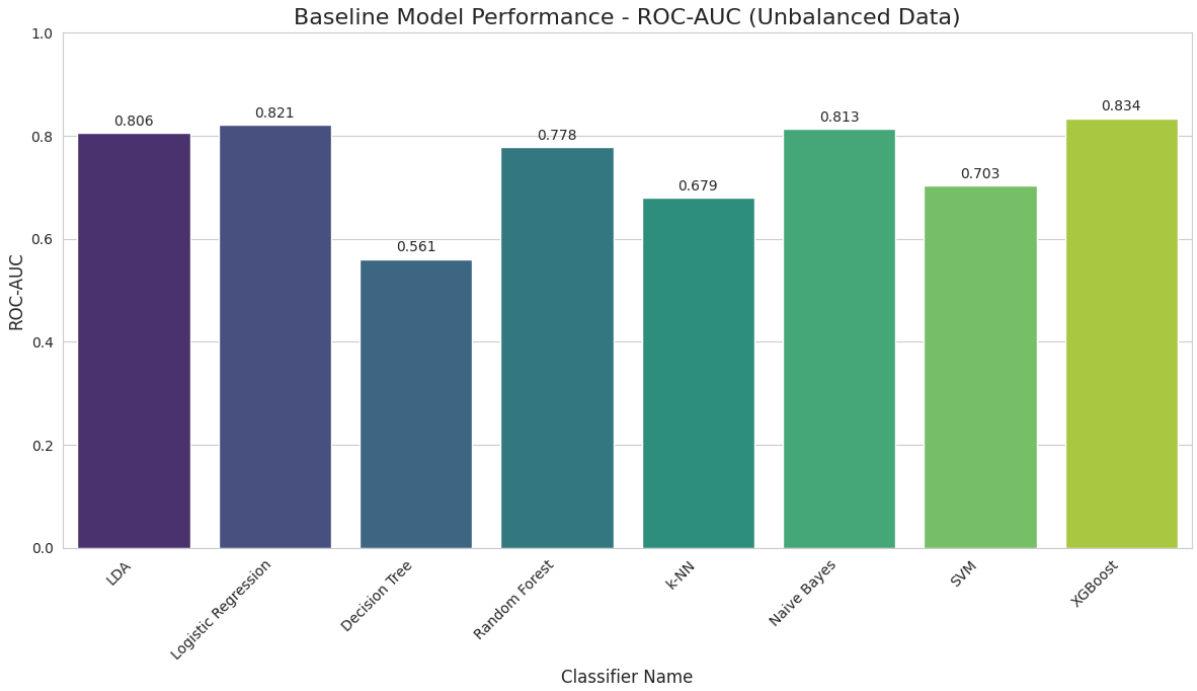


Το επόμενο μέρος της ανάλυσης περιλαμβάνει μία αρχική αξιολόγηση της ποιότητας των διαφόρων μοντέλων και δεικτών ταξινόμησης. Η ανάλυση αυτή θα διεξαχθεί σε μη επεξεργασμένα δεδομένα (unbalanced data), δηλαδή στα αρχικά μας δεδομένα προτού εφαρμοστεί η οποιαδήποτε μεταρρύθμιση ή βελτιστοποίηση. Αυτή η προκαταρκτική αξιολόγηση θα προσφέρει μία βασική κατανόηση της απόδοσης των μοντέλων στις προεπιλεγμένες ρυθμίσεις.


Baseline Model Performance - Recall (Unbalanced Data)
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.081
0.016
0.145
0.016
0.032
0.306
0.000
0.097
LDA
Logistic Regression
Decision Tree
Random Forest
k-NN
Naive Bayes
SVM
XGBoost
Classifier Name


Baseline Model Performance - Precision (Unbalanced Data)
Precision
1.0
0.8
0.6
0.4
0.2
0.0
0.227
1.000
0.129
1.000
0.400
0.116
0.000
0.462
LDA
Logistic Regression
Decision Tree
Random Forest
k-NN
Naive Bayes
SVM
XGBoost
Classifier Name

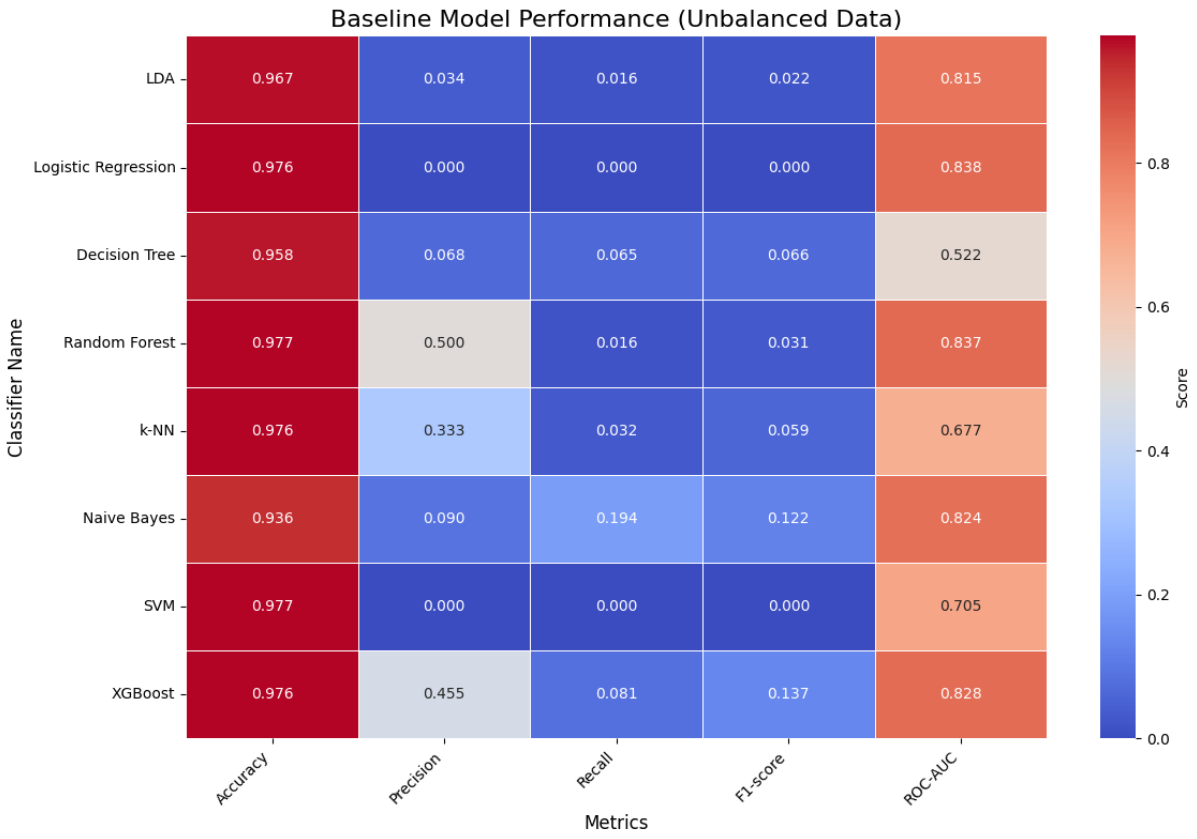




Η αρχική αξιολόγηση των ανεπεξεργασμένων τεχνικών μηχανικής μάθησης αποκαλύπτει την περιορισμένη χρησιμότητά τους στην αρχική τους μορφή, ιδιαίτερα όταν πρόκειται για σύνθετα κριτήρια. Αυτό γίνεται εμφανές κατά τη

σύγκριση των διαφορετικών μετρικών απόδοσης. Για παράδειγμα, ενώ η ακρίβεια (accuracy) φαίνεται παρόμοια σε όλα τα μοντέλα, η μετρική της ανάκλησης (recall) παρουσιάζει σημαντική μεταβλητότητα. Αυτή η αποτυχία στην ανάκληση (recall) οφείλεται στην κυριαρχία της κλάσης των υγιών εταιρειών, γεγονός που καθιστά επιβεβλημένη την εφαρμογή τεχνικών εξισορρόπησης (balancing), προκειμένου το μοντέλο να μάθει να αναγνωρίζει τις σπάνιες αλλά κρίσιμες περιπτώσεις χρεοκοπίας.

Η ενσωμάτωση των προηγούμενων γραφημάτων σε ένα θερμικό χάρτη (heatmap) παρέχει έναν ολοκληρωμένο χάρτη απόδοσης των μοντέλων. Αυτή η οπτικοποίηση διευκολύνει την καθαρότερη κατανόηση του τρόπου με τον οποίο διάφορα μοντέλα μηχανικής μάθησης αποδίδουν σε διαφορετικές μετρικές εντός του συνόλου δεδομένων.



Ο θερμικός χάρτης (heatmap) υποδεικνύει ότι τα μοντέλα Support Vector Machine (SVM) και Random Forest παρουσιάζουν την πιο ελπιδοφόρα δυναμική, παρά τις αρχικές χαμηλές τιμές τους. Αντιθέτως το μοντέλο Naive Bayes εμφανίζει την ασθενότερη απόδοση.

Οι μηδενικές τιμές στην ανάκληση (recall) για το SVM και τη Logistic Regression επιβεβαιώνουν ότι η ανισορροπία των δεδομένων αναγκάζει τους αλγορίθμους να αγνοούν πλήρως τις χρεοκοπίες για να διατηρήσουν υψηλή ακρίβεια (accuracy).

Στο πλαίσιο της πρόβλεψης της χρεοκοπίας, η υψηλή ανάκληση είναι ζωτικής σημασίας καθώς διασφαλίζει ότι οι εταιρείες που διατρέχουν κίνδυνο εντοπίζονται σωστά για πιθανή παρέμβαση. Μετρικές όπως το F1 score και το AUC ROC προσφέρουν μια πιο ισορροπημένη εικόνα απόδοσης, καθώς λαμβάνουν υπόψη το σοβαρό οικονομικό κόστος μιας εσφαλμένης ταξινόμησης.

Συνεπώς, είναι πρόωρο να επιλεγεί ένα βέλτιστο μοντέλο αποκλειστικά από αυτά τα δεδομένα. Τα αποτελέσματα αναδεικνύουν τον κομβικό ρόλο της επεξεργασίας δεδομένων (data balancing) που ακολουθεί. Αυτή είναι απαραίτητη ώστε να αποκτήσουν τα μοντέλα πραγματική διακριτική ικανότητα μεταξύ υγιών και υπό χρεοκοπία επιχειρήσεων.

3.2. Μοντέλα Ταξινόμησης

Η ανάλυση των μοντέλων υπογραμμίζει την ανάγκη εστίασης όχι μόνο στο σύνολο εκπαίδευσης (training set) αλλά και στο σύνολο ελέγχου (test set), επιδιώκοντας τη βελτιστοποίησή τους για την ενίσχυση των μετρικών απόδοσης. Αυτή η ολοκληρωμένη προσέγγιση είναι απαραίτητη για την ανάπτυξη εύρωστων μοντέλων μηχανικής μάθησης, ικανών για αξιόπιστες προβλέψεις.

Κατά την αξιολόγηση κάθε μοντέλου, κρίνεται απαραίτητο να ληφθεί υπόψη η ισορροπία μεταξύ σφαλμάτων τύπου I (ψευδώς θετικά) και των σφαλμάτων τύπου II (ψευδώς αρνητικά). Αυτλη η ισορροπία επηρεάζει άμεσα τις εξής επιχειρηματικές αποφάσεις:

- Ψευδώς Θετικά (False Positives): Ο εσφαλμένος προσδιορισμός υγιών επιχειρήσεων ως "υπό κίνδυνο" μπορεί να οδηγήσει σε αδικαιολόγητες παρεμβάσεις, διαταράσσοντας ενδεχομένως τη σταθερή λειτουργία τους.
- Ψευδώς Αρνητικά (False Negatives): Η αποτυχία εντοπισμού επιχειρήσεων που βρίσκονται πραγματικά στα πρόθυρα χρεοκοπίας θα μπορούσε να σημαίνει την απώλεια κρίσιμων ευκαιριών για έγκαιρη υποστήριξη ή παρέμβαση.

3.2.1. K-Nearest Neighbors (Knn)

3.2.2. Naïve Bayes (Nb)

3.2.3. Decision Trees (Dts)

3.2.4. Random Forest (RF)

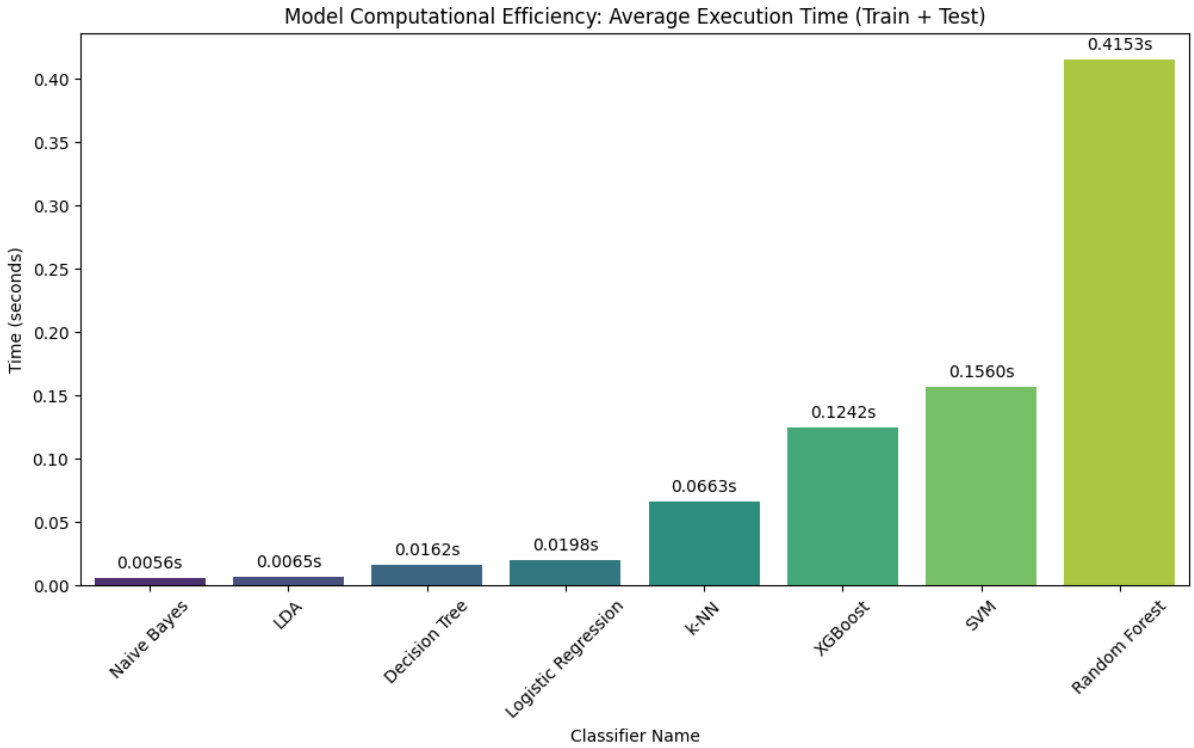
3.2.5. Linear Discriminant Analysis (LDA)

3.2.6. Support Vector Machines (SVM)

3.2.7. Logistic Regression (LR)

3.2.8. XGBoost

3.3 Χρονική Πολυπλοκότητα των Μοντέλων



Η ανάλυση της υπολογιστικής αποδοτικότητας των μοντέλων είναι κρίσιμη, καθώς η επιλογή του κατάλληλου αλγορίθμου πρέπει να ισορροπεί μεταξύ ακρίβειας και ταχύτητας εκτέλεσης. Το διάγραμμα απεικονίζει τον μέσο χρόνο που απαιτήθηκε για την εκπαίδευση και τον έλεγχο κάθε μοντέλου:

- Naive Bayes:
    - Χρόνος: Ο ταχύτερος όλων με 0.0056s.
    - Αιτιολογία: Η ταχύτητά του οφείλεται στην απλότητα της πιθανοτικής του προσέγγισης και στην υπόθεση ανεξαρτησίας μεταξύ των μεταβλητών.
- Linear Discriminant Analysis (LDA):
    - Χρόνος: Εξαιρετικά γρήγορος στα 0.0065s.

    - Αιτιολογία: Η ταχύτητα αυτή αποδίδεται στη γραμμική φύση του μοντέλου και την απλότητα των μαθηματικών υπολογισμών για τον διαχωρισμό των κλάσεων.

- Decision Tree:
    - Χρόνος: Απαιτεί 0.0162s.
    - Αιτιολογία: Παραμένει αποδοτικό, αν και ελαφρώς πιο αργό από τα προηγούμενα λόγω της διαδικασίας διακλάδωσης του δέντρου απόφασης.
- Logistic Regression:
    - Χρόνος: Ολοκληρώνεται σε 0.0198s.
    - Αποδοτικότητα: Θεωρείται ιδιαίτερα αποτελεσματική και γρήγορη μέθοδος για προβλήματα δυαδικής ταξινόμησης, όπως η πρόβλεψη χρεοκοπίας.
- k-Nearest Neighbors (k-NN):
    - Χρόνος: Στα 0.0663s.
    - Αιτιολογία: Η καθυστέρηση αυτή είναι αναμενόμενη, καθώς ο k-NN υπολογίζει τις αποστάσεις μεταξύ των δειγμάτων, μια διαδικασία που γίνεται πιο εντατική όσο αυξάνεται ο όγκος των δεδομένων.
- XGBoost:
    - Χρόνος: Χρειάζεται 0.1242s.
    - Αιτιολογία: Ως ένας αλγόριθμος ενίσχυσης (boosting), η πολυπλοκότητά του είναι μεγαλύτερη λόγω της σειριακής κατασκευής πολλαπλών δέντρων απόφασης.
- Support Vector Machine (SVM):
    - Χρόνος: Σημαντικά πιο αργός στα 0.1560s.
    - Αιτιολογία: Η εύρεση του βέλτιστου υπερεπιπέδου διαχωρισμού είναι υπολογιστικά ακριβή διαδικασία, ειδικά όταν χρησιμοποιούνται σύνθετοι πυρήνες (kernels) ή μεγάλες διαστάσεις δεδομένων.
- Random Forest:
    - Χρόνος: Ο πιο χρονοβόρος αλγόριθμος με 0.4153s.

- Αιτιολογία: Ως τεχνική συνόλου (ensemble), κατασκευάζει ένα μεγάλο πλήθος ανεξάρτητων δέντρων απόφασης, γεγονός που καθιστά τη διαδικασία εγγενώς πιο απαιτητική σε πόρους.

Συμπέρασμα Επιλογής: Για πρακτικές εφαρμογές, μοντέλα όπως το Naive Bayes και το LDA προσφέρουν άμεσα αποτελέσματα και είναι ιδανικά για συχνή επανεκπαίδευση. Αντίθετα, πιο σύνθετα μοντέλα όπως το Random Forest ή το SVM, αν και απαιτούν περισσότερο χρόνο, επιλέγονται όταν η μεγιστοποίηση της προβλεπτικής ικανότητας υπερτερεί της ανάγκης για ταχύτητα.

3.4. Προσδιορισμός Βέλτιστου Μοντέλου

3.5. Συγκριτική Αξιολόγηση και Τελικά Συμπεράσματα

4. Συζήτηση

4.1. Προκλήσεις και Περιορισμοί Μοντέλων

4.2. Μελλοντική Έρευνα

5. Συμπεράσματα

6. Σχετική Βιβλιογραφία