Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Εκτίμηση της Ζημίας σε Περίπτωση Αθέτησης σε Χαρτοφυλάκια Καταναλωτικών Δανείων με Αλγορίθμους Μηχανικής Μάθησης

Διπλωματική Εργασία

Κυρατσούς Χρήστος

ΑΕΜ:10105

Επιβλέπων : Κωνσταντίνος Παπαλάμπρου

Καθηγητής Α.Π.Θ.

Θεσσαλονίκη, Μάρτιος 2025

Περιεχόμενα

Ευχαριστίες 6
Περίληψη 7
Abstract 8
Γλωσσάριο 9
Ευρετήριο Εικόνων 11
Ευρετήριο Πινάκων 12
Κεφάλαιο 1: Εισαγωγή 13
1.1 Πιστωτικός Κίνδυνος 14
1.2 Είδη Πιστωτικού Κινδύνου 14
1.3 Διαχείριση Πιστωτικού Κινδύνου 15
1.4 Διαδικασία Πιστωτικής Βαθμολόγησης 15
1.5 Στοιχεία που Επηρεάζουν τη Βαθμολόγηση 16
1.6 Σημασία της Πιστωτικής Βαθμολόγησης 16
1.7 Ζημία Λόγω Αθέτησης 17
1.8 Πιθανότητα Θεραπείας 18
1.9 Πιθανότητα Ανάκτησης 18
1.10 Μηχανική Μάθηση και Πιστωτικός Κίνδυνος 18
1.11 Δυνατότητες Μηχανικής Μάθησης στην Ανάλυση Πιστωτικού Κινδύνου 19
1.12 Προκλήσεις και Σημεία Προσοχής για τους Αναλυτές 20
Κεφάλαιο 2: Μετρικές στατιστικού ελέγχου και μετρικές σφάλματος 21
2.1: Μετρικές στατιστικού ελέγχου 21
2.2: Μετρικές σφάλματος 23
Κεφάλαιο 3: Επιβλεπόμενη και μη Επιβλεπόμενη Μηχανική Μάθηση 29
3.1 Επιβλεπόμενη μηχανική μάθηση 29
3.2 Μη Επιβλεπόμενη Μηχανική Μάθηση 31
3.3 Μέθοδοι Ταξινόμησης 32
3.3.1 Λογιστική Παλινδρόμηση 33
3.3.2 Τυχαίο Δάσος 34

3.3.3 Extreme Gradient Boosting 37
3.3.4 Βαθιά Νευρωνικά Δίκτυα 39
3.4 Μέθοδοι Παλινδρόμησης 41
3.4.1 Γραμμική Παλινδρόμηση 42
3.4.2 Τυχαίο Δάσος για Παλινδρόμηση 42
3.4.3 Extreme Gradient Boosting (XGBoost) για Παλινδρόμηση 43
3.4.4 Βαθιά Νευρωνικά Δίκτυα για Παλινδρόμηση 44
3.5 Μέθοδοι Εύρεσης Βέλτιστων Υπερπαραμέτρων 45
3.5.1 Αναζήτηση με Χρήση Πλέγματος 45
3.5.2 Τυχαία Αναζήτηση 46
3.5.3 Μπεϋζιανή Βελτιστοποίηση 47
3.5.4 Γενετικοί Αλγόριθμοι 48
3.6 Διασταυρουμενη Επικύρωση k-fold 49
Κεφάλαιο 4: Κατασκευή και Προεπεξεργασία Δεδομένων 51
4.1 Σύνολο Δεδομένων (Dataset) 51
4.2 Προεπεξεργασία των Δεδομένων 53
Κεφάλαιο 5: Μεθοδολογία 60
5.1 Προετοιμασία Δεδομένων και Διαχωρισμός 60
5.2. Περιγραφή των Μοντέλων Κατηγοριοποίησης 60
5.2.1 Λογιστική Παλινδρόμηση 60
5.2.2 Τυχαίο Δάσος (Random Forest) 61
5.2.3 XGBoost 62
5.2.4 Βαθύ Νευρωνικό Δίκτυο (Neural Network) 63
5.3 Περιγραφή των Μοντέλων Παλινδρόμησης 65
5.3.1 Γραμμική Παλινδρόμηση 65
5.3.2 Τυχαίο Δάσος (Random Forest) 66
5.3.3 XGBoost 67
5.3.4 Βαθύ Νευρωνικό Δίκτυο (Neural Network) 68
Κεφάλαιο 6: Αποτελέσματα Μοντέλων 73

6.1 Μοντέλα Κατηγοριοποίησης 73
6.2 Αποτελέσματα Μοντέλων Παλινδρόμησης 77
Κεφάλαιο 7: Συμπεράσματα και Μελλοντικές Προεκτάσεις 84
Βιβλιογραφία 86
Παράρτημα 1: Σύστημα και Βιβλιοθήκες 89
Παράρτημα 2 91

Ευχαριστίες

Θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον Επίκουρο Καθηγητή κ. Κωνσταντίνο Παπαλάμπρου, επιβλέποντα της διπλωματικής μου εργασίας, για την πολύτιμη καθοδήγηση, την επιστημονική του υποστήριξη και τις γόνιμες συζητήσεις καθ' όλη τη διάρκεια της ερευνητικής μου προσπάθειας. Η αφοσίωσή του, η εμπεριστατωμένη του γνώση και οι εύστοχες συμβουλές του υπήρξαν καθοριστικές για την ολοκλήρωση αυτής της μελέτης.

Επιπλέον, θα ήθελα να εκφράσω την ευγνωμοσύνη μου στην οικογένειά μου για την αδιάκοπη ηθική και συναισθηματική τους υποστήριξη. Η ενθάρρυνση, η κατανόηση και η στήριξή τους αποτέλεσαν πολύτιμο εφόδιο σε όλη τη διάρκεια των σπουδών μου, δίνοντάς μου τη δύναμη να ολοκληρώσω αυτή την ερευνητική εργασία.

Χωρίς την καθοδήγηση του επιβλέποντα καθηγητή μου και την αμέριστη συμπαράσταση της οικογένειάς μου, η ολοκλήρωση της παρούσας διπλωματικής εργασίας δεν θα ήταν εφικτή. Τους ευχαριστώ θερμά.

Περίληψη

Η παρούσα διπλωματική εργασία πραγματεύεται τη χρήση προηγμένων τεχνικών μηχανικής μάθησης με στόχο την πρόβλεψη της Ζημίας Λόγω Αθέτησης (Loss Given Default - LGD), ενός κρίσιμου παράγοντα στη διαχείριση πιστωτικού κινδύνου. Ιδιαίτερη έμφαση δίνεται στην ανάλυση και πρόβλεψη δύο σημαντικών μεγεθών, του ποσοστού θεραπείας (Cure Rate) και του ποσοστού ανάκτησης (Recovery Rate), τα οποία καθορίζουν σε μεγάλο βαθμό το ύψος της ζημίας που υφίσταται ένας χρηματοπιστωτικός οργανισμός όταν ένας δανειολήπτης αθετεί τις υποχρεώσεις του.

Αρχικά, παρουσιάζεται αναλυτικά το θεωρητικό υπόβαθρο των μοντέλων μηχανικής μάθησης και των στατιστικών μεθοδολογιών που εφαρμόζονται για την αξιολόγηση της απόδοσης των προβλέψεων. Στο πλαίσιο αυτό, εξετάζονται και εφαρμόζονται διάφορα μοντέλα, όπως η Γραμμική και Λογιστική Παλινδρόμηση, το Τυχαίο Δάσος, το Extreme Gradient Boosting και τα Νευρωνικά Δίκτυα.

Η επίδοση των μοντέλων αξιολογείται μέσω κατάλληλων στατιστικών μετρικών και δεικτών σφάλματος, με στόχο την ανάδειξη του πιο αξιόπιστου μοντέλου για την ακριβή και αποτελεσματική πρόβλεψη των παραμέτρων που επηρεάζουν την LGD. Τα αποτελέσματα της εργασίας συμβάλλουν τόσο στη θεωρητική κατανόηση του ζητήματος όσο και στην πρακτική αξιοποίηση αυτών των μοντέλων στον τραπεζικό και χρηματοοικονομικό κλάδο, ενισχύοντας τη διαδικασία λήψης αποφάσεων και μειώνοντας τον πιστωτικό κίνδυνο.

Λέξεις Κλειδιά: Ζημία Λόγω Αθέτησης, Ποσοστό Θεραπείας, Ποσοστό Ανάκτησης, Μηχανική Μάθηση, Γραμμική και Λογιστική Παλινδρόμηση, Τυχαίο Δάσος, Extreme Gradient Boosting, Νευρωνικά Δίκτυα, Πιστωτικός Κίνδυνος

Abstract

This thesis focuses on the use of advanced machine learning techniques aimed at predicting Loss Given Default (LGD), a crucial factor in credit risk management. Emphasis is placed on analyzing and predicting two key metrics: the Cure Rate and the Recovery Rate, both significantly determining the level of loss experienced by a financial institution when a borrower defaults on their obligations.

Initially, the theoretical background of machine learning models and statistical methodologies employed for evaluating prediction performance is comprehensively presented. Within this context, various models such as Linear and Logistic Regression, Random Forest, Extreme Gradient Boosting, and Neural Networks are explored and applied.

Model performance is assessed using appropriate statistical metrics and error indices, with the aim of identifying the most reliable model for accurately and effectively predicting the parameters influencing LGD. The results of this study contribute both to the theoretical understanding of the topic and the practical application of these models in the banking and financial sectors, enhancing decision-making processes and reducing credit risk.

Keywords: Loss Given Default, Cure Rate, Recovery Rate, Machine Learning, Linear and Logistic Regression, Random Forest, Extreme Gradient Boosting, Neural Networks, Credit Risk

Γλωσσάριο

Το ακόλουθο ευρετήριο περιλαμβάνει τους βασικούς όρους και έννοιες που χρησιμοποιούνται στη διπλωματική εργασία:

Αγγλικός Όρος	**Συντομογραφία**	**Ελληνικός 'Ορος**
Loss Given Default	LGD	Ζημία Λόγω Αθέτησης
Cure Rate		Ποσοστό Θεραπείας
Recovery Rate		Ποσοστό Ανάκτησης
Machine Learning	MM	Μηχανική Μάθηση
Logistic Regression	LR	Λογιστική Παλινδρόμηση
Extreme Gradient Boost	XGBoost	Ακραία Ενίσχυση Κλίσης
Random Forest	RF	Τυχαίο Δάσος
Deep Learning	DL	Βαθιά Μάθηση
Deep Neural Network	DNN	Βαθύ Νευρωνικό Δίκτυο
Supervised Learning		Επιβλεπόμενη Μάθηση
Coefficient		Συντελεστής
Intercept		Σταθερά
Bias		Σταθερά πόλωσης
Activation Function		Συνάρτηση Ενεργοποίησης
Mean		Μέσος όρος
Median		Διάμεσος
Standard Deviation	STD	Τυπική Απόκλιση
Grid Search		Αναζήτηση με χρήση πλέγματος
Bayesian Optimization		Μπεϋζιανή Βελτιστοποίηση
False Negative	FN	Εσφαλμένα αρνητικό
False Positive	FP	Εσφαλμένα Θετικό
True Negative	TN	Πραγματικά Αρνητικό
True Positive	TP	Πραγματικά Θετικό

k- fold Cross Validation		Διασταυρούμενη Επικύρωση k Πτυχών
Decision Tree	DT	Δέντρο Απόφασης
Train Data		Δεδομένα εκπαίδευσης
Test Data		Δεδομένα Ελέγχου
Hyperparameter		Υπερπαράμετρος
Maximum	max	Μέγιστη τιμή
Minimum	min	Ελάχιστη Τιμή
Threshold		Κατώφλι
Variance	Var	Διακύμανση
Covariance	Cov	Συνδιακύμανση
Confusion Matrix		Πίνακας Σύγχυσης
Accuracy		Ακρίβεια
Dataset		Σύνολο Δεδομένων
Mean Square Error	MSE	Μέσο Τετραγωνικό Σφάλμα
Mean Absolute Error	MAE	Μέσο Απόλυτο Σφάλμα
Classification		Ταξινόμηση
Regression		Παλινδρόμηση
Interquartile Range	IQR	Διαμεσολαδικό Εύρος

Πίνακας 1: Γλωσσάριο

Ευρετήριο Εικόνων

Εικόνα 1: Πίνακας Σύγχυσης 25
Εικόνα 2: Επιβλεπόμενη Μάθηση 30
Εικόνα 3: Λογιστική Συνάρτηση 33
Εικόνα 4: Παράδειγμα Δέντρου Απόφασης 35
Εικόνα 5: Τυχαίο Δάσος 36
Εικόνα 6: Αλγόριθμος XGBoost 38
Εικόνα 7: Παράδειγμα Βαθύ Νευρωνικού Δικτύου 40
Εικόνα 8: Σύγκριση Μεθόδων Βελτιστοποίησης Υπερπαραμέτρων 47
Εικόνα 9: Διασταυρούμενη Επικύρωση 5-fold 50
Εικόνα 10: Διάγραμμα αναζήτησης πλέγματος για Τυχαίο Δάσος Κατηγοριοποίησης 70
Εικόνα 11: Διάγραμμα αναζήτησης πλέγματος για XGBoost Κατηγοριοποίησης 70
Εικόνα 12: Διάγραμμα Αναζήτησης Πλέγματος για Τυχαίο Δάσος Παλινδρόμησης 71
Εικόνα 13: Διάγραμμα Αναζήτησης Πλέγματος για XGBoost Παλινδρόμησης 71
Εικόνα 14: Διάγραμμα Bayesian Βελτιστοποίησης για Νευρωνικό Δίκτυο Παλινδρόμησης 72
Εικόνα 15: Πίνακες Σύγχυσης Μοντέλων Ταξινόμησης 75
Εικόνα 16: Διαγράμματα Ακρίβειας και Loss για Νευρωνικό Δίκτυο Ταξινόμησης 76
Εικόνα 17: Διαγράμματα Υπολειμμάτων Μοντέλων Παλινδρόμησης 79
Εικόνα 18: Διαγράμματα Πραγματικών vs Προβλεπόμενων Τιμών Μοντέλων Παλινδρόμησης 81
Εικόνα 19: Διαγράμματα Εκπαίδευσης Νευρωνικού Δικτύου Παλινδρόμησης 83

Ευρετήριο Πινάκων

Πίνακας 1: Γλωσσάριο 10
Πίνακας 2: Τα χαρακτηριστικά του Dataset και η Περιγραφή τους 52
Πίνακας 3: Αριθμητικές και Κατηγορικές Μεταβλητές 54
Πίνακας 4: Τελικές Κατασκευασμένες Μεταβλητές 56
Πίνακας 5: Πίνακας Συσχέτισης 58
Πίνακας 6: Τελικά Χαρακτηριστικά Συνόλου 59
Πίνακας 7: Υπερπαράμετροι τελικού μοντέλου Τυχαίου Δάσους 61
Πίνακας 8: Υπερπαράμετροι τελικού μοντέλου XGBoost 63
Πίνακας 9: Υπερπαράμετροι Τελικού Μοντέλου Νευρωνικού Δικτύου 64
Πίνακας 10: Υπερπαράμετροι Τελικού Μοντέλου Τυχαίου Δάσους Παλινδρόμησης 66
Πίνακας 11: Υπερπαράμετροι Τελικού Μοντέλου XGBoost Παλινδρόμησης 68
Πίνακας 12: Υπερπαράμετροι Τελικού Νευρωνικού Δικτύου Παλινδρόμησης 69
Πίνακας 13: Αποτελέσματα Μοντέλων Ταξινόμησης 73
Πίνακας 14: Αποτελέσματα Μοντέλων Παλινδρόμησης 77

Κεφάλαιο 1: Εισαγωγή

Η ανάλυση του πιστωτικού κινδύνου αποτελεί κρίσιμο κομμάτι της λειτουργίας των χρηματοπιστωτικών ιδρυμάτων, καθώς βοηθά στην αξιολόγηση της πιθανότητας αθέτησης πληρωμών από τους δανειολήπτες και στην εκτίμηση των σχετικών επιπτώσεων. Μία από τις σημαντικές μετρικές που χρησιμοποιούνται είναι η Ζημία Λόγω Αθέτησης (Loss Given Default - LGD), δηλαδή το ποσοστό απωλειών που μπορεί να υποστεί ένας δανειστής σε περίπτωση που ο δανειολήπτης αθετήσει το δάνειο. Η ακριβής πρόβλεψη του LGD είναι κρίσιμη για τον σωστό υπολογισμό των αποθεματικών που πρέπει να διαθέσει το ίδρυμα, για την αποφυγή περαιτέρω οικονομικών απωλειών. [1]

Ωστόσο, η εκτίμηση του LGD είναι περίπλοκη, καθώς επηρεάζεται από πληθώρα παραγόντων, συμπεριλαμβανομένων των επιτοκίων, της οικονομικής κατάστασης του δανειολήπτη και της αγοράς. Η μελέτη των δεικτών «Cure Rate» (ο ρυθμός επαναφοράς στην κανονικότητα των δανείων) και «Recovery Rate» (το ποσοστό αποπληρωμής μετά από αθέτηση) προσφέρει κρίσιμες πληροφορίες για τον υπολογισμό του LGD, καθώς δείχνουν πώς αντιδρούν τα δάνεια μετά την αθέτηση.

Η χρήση της Μηχανικής Μάθησης (ΜΜ) έρχεται να ενισχύσει αυτή τη διαδικασία, δίνοντας τη δυνατότητα ανάλυσης μεγάλων όγκων δεδομένων και αναγνώρισης μοτίβων που δεν είναι εύκολα ανιχνεύσιμα με παραδοσιακές μεθόδους. Οι αλγόριθμοι ΜΜ μπορούν να συμβάλουν στην ακριβέστερη πρόβλεψη του LGD μέσα από την ανάλυση των Cure και Recovery Rates, προσφέροντας έτσι καλύτερη κατανόηση των κινδύνων και πιο στοχευμένες στρατηγικές για τη διαχείριση του πιστωτικού κινδύνου.

1.1 Πιστωτικός Κίνδυνος

Ο **πιστωτικός κίνδυνος** αναφέρεται στην πιθανότητα να μην εκπληρώσει ο δανειολήπτης τις υποχρεώσεις του προς τον δανειστή, προκαλώντας οικονομικές απώλειες στον τελευταίο. Είναι ένας από τους κυριότερους κινδύνους για τα χρηματοπιστωτικά ιδρύματα, καθώς οφείλεται στην αδυναμία του δανειολήπτη να αποπληρώσει το κεφάλαιο και τους τόκους του δανείου, γεγονός που επηρεάζει αρνητικά τα έσοδα και τη ρευστότητα του ιδρύματος. [2]

1.2 Είδη Πιστωτικού Κινδύνου

Τα κύρια είδη πιστωτικού κινδύνου περιλαμβάνουν:

Κίνδυνος Αθέτησης: Η πιθανότητα ο δανειολήπτης να μην μπορέσει να αποπληρώσει το χρέος του. Ο κίνδυνος αυτός διακρίνεται σε δύο βασικές κατηγορίες [3] :

- **Επενδυτικής κατηγορίας**, όπου η πιθανότητα αθέτησης είναι σχετικά χαμηλή, συνήθως λόγω της ισχυρής χρηματοοικονομικής θέσης του δανειολήπτη.
- **Μη επενδυτικής κατηγορίας,** όπου ο κίνδυνος αθέτησης είναι υψηλότερος και συνδέεται με δανειολήπτες χαμηλότερης πιστοληπτικής ικανότητας.

Κίνδυνος Συγκέντρωσης: Προκύπτει όταν ένα χρηματοπιστωτικό ίδρυμα έχει μεγάλη έκθεση σε έναν μεμονωμένο δανειολήπτη, κλάδο, ή γεωγραφική περιοχή. Αυτό αυξάνει τον κίνδυνο απώλειας αν οι συγκεκριμένες οντότητες ή περιοχές αντιμετωπίσουν οικονομικές δυσκολίες.

Κίνδυνος Χρηματοδότησης: Προκύπτει όταν το ίδρυμα δεν έχει αρκετή ρευστότητα για να καλύψει τις ανάγκες του σε δάνεια ή χρηματοδότηση. Αυτό μπορεί να συμβεί λόγω καθυστερημένων πληρωμών ή ανεπαρκούς εισροής κεφαλαίων. [4]

Κίνδυνος Ανάκτησης: Αναφέρεται στην αβεβαιότητα ως προς το ποσό που θα ανακτηθεί εάν ο δανειολήπτης αθετήσει την υποχρέωσή του, δηλαδή το μέρος της αρχικής έκθεσης που θα επιστραφεί.

1.3 Διαχείριση Πιστωτικού Κινδύνου

Η διαχείριση του πιστωτικού κινδύνου περιλαμβάνει την αξιολόγηση της πιστοληπτικής ικανότητας των δανειοληπτών μέσω της χρήσης εξωτερικών αξιολογήσεων από πιστοληπτικούς οργανισμούς, οι οποίοι παρέχουν πιστοληπτικές βαθμολογίες (credit scores). Αυτές οι βαθμολογίες βοηθούν τα χρηματοπιστωτικά ιδρύματα να εκτιμήσουν τον κίνδυνο αθέτησης και να προσαρμόσουν τις στρατηγικές δανεισμού τους. [5]

Πιστωτική Βαθμολόγηση

Η πιστωτική βαθμολόγηση αναφέρεται στη διαδικασία αξιολόγησης της πιστοληπτικής ικανότητας ενός ατόμου ή μιας εταιρείας και αποτελεί βασικό εργαλείο για την εκτίμηση του πιστωτικού κινδύνου. Η βαθμολόγηση αυτή χρησιμοποιείται από τα χρηματοπιστωτικά ιδρύματα για να καθορίσουν την πιθανότητα αθέτησης των υποχρεώσεων ενός δανειολήπτη και να λάβουν αποφάσεις σχετικά με τη χορήγηση δανείων ή πιστώσεων. Οι βαθμολογίες αυτές αντανακλούν την ικανότητα του δανειολήπτη να αποπληρώσει το χρέος του βάσει της οικονομικής του κατάστασης, της ιστορίας του και των άλλων σχετικών παραμέτρων. [6]

1.4 Διαδικασία Πιστωτικής Βαθμολόγησης

Η διαδικασία της πιστωτικής βαθμολόγησης περιλαμβάνει τη συλλογή και ανάλυση δεδομένων από διάφορες πηγές, όπως οι τράπεζες, οι εταιρείες πιστωτικών καρτών και άλλοι χρηματοπιστωτικοί οργανισμοί. Στη συνέχεια, οι οργανισμοί βαθμολόγησης (όπως η S&P Global, η Moody’s και η Fitch Ratings) χρησιμοποιούν στατιστικά μοντέλα και αλγορίθμους για να αποδώσουν μια βαθμολογία που εκφράζει τον κίνδυνο αθέτησης του δανειολήπτη. [7]

Οι βαθμολογίες κυμαίνονται από υψηλές (που δείχνουν μικρότερο κίνδυνο αθέτησης) έως χαμηλές (που δείχνουν υψηλότερο κίνδυνο αθέτησης). Οι υψηλές βαθμολογίες, όπως η κατηγορία Investment Grade, αντιπροσωπεύουν ασφαλή και αξιόπιστα άτομα ή εταιρείες, ενώ οι χαμηλές, όπως οι κατηγορίες Junk ή Speculative Grade, αναφέρουν υψηλότερο κίνδυνο για τους δανειστές. [8]

1.5 Στοιχεία που Επηρεάζουν τη Βαθμολόγηση

Η πιστωτική βαθμολογία επηρεάζεται από διάφορους παράγοντες, όπως [9]:

Ιστορικό Πληρωμών: Αν ο δανειολήπτης έχει καθυστερήσει ή αθετήσει πληρωμές στο παρελθόν.
Ποσό Χρέους: Το συνολικό ποσό του χρέους που έχει ο δανειολήπτης σε σχέση με το εισόδημα ή τα περιουσιακά του στοιχεία.
Διάρκεια Ιστορίας Πιστώσεων: Η διάρκεια κατά την οποία ο δανειολήπτης χρησιμοποιεί πιστωτικά προϊόντα.
Νέα Πιστωτικά Προϊόντα: Αριθμός νέων πιστωτικών γραμμών ή δανείων που έχει ανοίξει ο δανειολήπτης.
Τύπος Πιστωτικών Προϊόντων: Ο τύπος των δανείων που έχει χρησιμοποιήσει ο δανειολήπτης, π.χ. προσωπικά δάνεια, στεγαστικά δάνεια ή πιστωτικές κάρτες.

1.6 Σημασία της Πιστωτικής Βαθμολόγησης

Η πιστωτική βαθμολογία έχει σημαντική επίδραση στις οικονομικές αποφάσεις των ατόμων και των οργανισμών. Για τους δανειολήπτες, η καλή βαθμολογία μπορεί να οδηγήσει σε χαμηλότερα επιτόκια και πιο ευνοϊκούς όρους δανεισμού, ενώ η κακή βαθμολογία μπορεί να οδηγήσει σε υψηλότερους τόκους ή ακόμη και στην απόρριψη αίτησης για δάνειο. Για τα χρηματοπιστωτικά ιδρύματα, η πιστωτική βαθμολόγηση είναι κρίσιμη για την αξιολόγηση του πιστωτικού κινδύνου και την ασφαλή διαχείριση των χαρτοφυλακίων τους. [11]

1.7 Ζημία Λόγω Αθέτησης

Η Ζημία Λόγω Αθέτησης (LGD) [10] είναι μια σημαντική παράμετρος στην ανάλυση πιστωτικού κινδύνου και αναφέρεται στο ποσοστό των απωλειών που θα υποστεί ένας δανειστής σε περίπτωση που ο δανειολήπτης αθετήσει την υποχρέωσή του. Ο υπολογισμός του LGD περιλαμβάνει την εκτίμηση του ποσού που δεν θα καταστεί δυνατόν να ανακτηθεί από το συνολικό οφειλόμενο ποσό, μετά από την αθέτηση του δανείου. Το LGD είναι καθοριστικός παράγοντας για την αποτίμηση του κινδύνου και τη διαχείριση του κεφαλαίου που απαιτείται για την κάλυψη των πιθανών απωλειών. Ένα υψηλότερο LGD σημαίνει μεγαλύτερο κίνδυνο απώλειας για τον δανειστή και ενδεχομένως την ανάγκη για αυξημένα αποθεματικά. [13]

Το LGD επηρεάζεται από παράγοντες όπως:

- Ο τύπος του χρέους (π.χ. στεγαστικά δάνεια, καταναλωτικά δάνεια).
- Η ασφάλεια ή τα ενέχυρα που έχουν παρασχεθεί.
- Οι συνθήκες της αγοράς κατά την περίοδο αθέτησης.
- Η διαδικασία εκτέλεσης των δικαιωμάτων του δανειστή για την ανάκτηση των οφειλών.
- Η ακριβής εκτίμηση του LGD είναι απαραίτητη για τη διαχείριση του πιστωτικού κινδύνου και την αποτελεσματική λήψη αποφάσεων στον τομέα του δανεισμού.

Στα πλαίσια της εργασίας αυτής, χρησιμοποιείται η φόρμουλα:

$$LGD = \frac{(1-\text{cure rate})\times(1-\text{recovery rate})}{EAD} + \frac{costs}{EAD} \quad [53] \qquad (1)$$

Η φόρμουλα αυτή χρησιμοποιείται από την ING και συνδυάζει τους όρους του LGD, Cure Rate και Recovery Rate, με τον βέλτιστο δυνατό τρόπο, ώστε να επιτυγχάνονται οι καλύτερες δυνατές προβλέψεις και να περιορίζονται στο ελάχιστο οι απώλειες των δανειοδοτών.

1.8 Πιθανότητα Θεραπείας

Η Πιθανότητα Θεραπείας [16] (Cure Rate) αναφέρεται στο ποσοστό των δανείων που, παρά την αθέτησή τους, επανέρχονται σε κανονική κατάσταση. Στην πράξη, αυτό σημαίνει ότι ο δανειολήπτης καθίσταται ικανός να αποπληρώσει το χρέος του μετά από μια περίοδο αθέτησης ή καθυστέρησης. Η υψηλή Πιθανότητα Θεραπείας υποδεικνύει ότι οι δανειολήπτες που αντιμετωπίζουν προσωρινές δυσκολίες οικονομικής φύσης, έχουν την ικανότητα να αποκαταστήσουν τη φερεγγυότητά τους και να αποφύγουν τη συνολική αθέτηση. Είναι μια σημαντική παράμετρος για τον υπολογισμό του LGD, καθώς επηρεάζει τον τελικό όγκο των απολεσθέντων κεφαλαίων.

1.9 Πιθανότητα Ανάκτησης

Η Πιθανότητα Ανάκτησης (Recovery Rate) αναφέρεται στο ποσοστό της αρχικής οφειλής που ο δανειστής καταφέρνει να ανακτήσει μετά την αθέτηση του δανείου. Αντιπροσωπεύει την αποπληρωμή του χρέους από τον δανειολήπτη ή μέσω της υλοποίησης εγγυήσεων ή ενέχυρων. Η υψηλή Πιθανότητα Ανάκτησης υποδηλώνει ότι ο δανειστής έχει τη δυνατότητα να ανακτήσει σημαντικό μέρος της απώλειας του χρέους, ενώ η χαμηλή σημαίνει μεγαλύτερες απώλειες για τον δανειστή. [17]

1.10 Μηχανική Μάθηση και Πιστωτικός Κίνδυνος

Η Μηχανική Μάθηση (ΜΜ) παρέχει ισχυρά εργαλεία για την ανάλυση του πιστωτικού κινδύνου, βελτιώνοντας την ακρίβεια της πρόβλεψης της αθέτησης πληρωμών και άλλων κρίσιμων παραμέτρων όπως το LGD, το Cure Rate και το Recovery Rate. Μέσω της εφαρμογής αλγορίθμων ΜΜ, οι αναλυτές μπορούν να αξιοποιήσουν μεγάλα σύνολα δεδομένων για την ανίχνευση μοτίβων και σχέσεων που δεν είναι εύκολα ορατά με παραδοσιακές μεθόδους. [19]

1.11 Δυνατότητες Μηχανικής Μάθησης στην Ανάλυση Πιστωτικού Κινδύνου

- **Αναγνώριση Σχέσεων και Μοτίβων**: Η ΜΜ μπορεί να επεξεργαστεί μεγάλους όγκους δεδομένων, όπως ιστορικά δεδομένα πληρωμών, οικονομικά μεγέθη, και συμπεριφορές δανειοληπτών, για να ανιχνεύσει σχέσεις που μπορεί να μην είναι εμφανείς με παραδοσιακές αναλύσεις.
- **Ακρίβεια Πρόβλεψης**: Μέσω αλγορίθμων όπως τα δέντρα αποφάσεων και τα νευρωνικά δίκτυα, οι αναλυτές μπορούν να προβαίνουν σε πιο ακριβείς προβλέψεις σχετικά με την πιθανότητα αθέτησης ενός δανειολήπτη ή τον υπολογισμό του LGD, βασιζόμενοι σε παράγοντες όπως η συμπεριφορά πληρωμών και τα χαρακτηριστικά του δανειολήπτη.
- **Διαχείριση Χαμένων Δεδομένων**: Η ΜΜ έχει τη δυνατότητα να χειρίζεται και να επεξεργάζεται δεδομένα που μπορεί να είναι ελλιπή ή θορυβώδη, καθιστώντας τη πολύτιμη στην ανάλυση πιστωτικού κινδύνου όπου συχνά υπάρχουν ατελή ή ανομοιογενή δεδομένα. [20]
- **Εξατομικευμένες Εκτιμήσεις Κινδύνου**: Οι αλγόριθμοι μπορούν να προσαρμοστούν ώστε να δημιουργούν εξατομικευμένες εκτιμήσεις πιστωτικού κινδύνου, λαμβάνοντας υπόψη τις συγκεκριμένες συνθήκες και χαρακτηριστικά κάθε δανειολήπτη, παρέχοντας πιο ακριβείς και στοχευμένες αποφάσεις.
- **Στρατηγικές Διαχείρισης Κινδύνου**: Η ΜΜ μπορεί να ενισχύσει τη διαχείριση κινδύνου παρέχοντας τακτικές προειδοποιήσεις για πιθανές αλλαγές στην πιστωτική κατάσταση των δανειοληπτών, επιτρέποντας στα χρηματοπιστωτικά ιδρύματα να προσαρμόσουν τις στρατηγικές τους και να μειώσουν τις απώλειες. [21]

1.12 Προκλήσεις και Σημεία Προσοχής για τους Αναλυτές

Ωστόσο, η χρήση της Μηχανικής Μάθησης στην ανάλυση πιστωτικού κινδύνου δεν είναι χωρίς προκλήσεις. Ορισμένα σημεία που πρέπει να προσέχουν οι αναλυτές περιλαμβάνουν [22]:

1. **Αναγνώριση Υποκείμενων Σχεδίων**: Η ΜΜ μπορεί να αναπαράγει ή ακόμα και να ενισχύσει υπάρχουσες προκαταλήψεις στα δεδομένα, εάν τα δεδομένα εκπαίδευσης περιλαμβάνουν αθέλητα προκατειλημμένες πληροφορίες. Ο αναλυτής πρέπει να είναι προσεκτικός στην επιλογή και προετοιμασία των δεδομένων, ώστε να αποφευχθεί η δημιουργία μοντέλων που θα μπορούσαν να οδηγήσουν σε αθέμιτες διακρίσεις ή εσφαλμένες προβλέψεις.
2. **Διαφάνεια και Ερμηνευσιμότητα**: Ορισμένοι αλγόριθμοι ΜΜ, όπως τα νευρωνικά δίκτυα, είναι πολύπλοκοι και δύσκολα εξηγήσιμοι (black box models). Αυτό σημαίνει ότι μπορεί να είναι δύσκολο να κατανοήσουμε τις αιτίες πίσω από τις προβλέψεις, γεγονός που καθιστά τη διαδικασία λήψης αποφάσεων πιο αδιαφανή. Η εξασφάλιση ερμηνευσιμότητας είναι ζωτικής σημασίας για την εμπιστοσύνη και τη νομιμότητα των προβλέψεων.
3. **Ποιότητα και Καθαρότητα Δεδομένων**: Η ποιότητα των δεδομένων είναι κρίσιμη για την επιτυχία των μοντέλων ΜΜ. Λανθασμένα, ελλιπή ή θορυβώδη δεδομένα μπορεί να οδηγήσουν σε λανθασμένες προβλέψεις και αυξημένο πιστωτικό κίνδυνο. Η καθαρότητα και η ομοιομορφία των δεδομένων πρέπει να διασφαλίζεται, και οι αναλυτές πρέπει να είναι προσεκτικοί κατά τη διάρκεια της διαδικασίας προετοιμασίας των δεδομένων.
4. **Συνεχής Παρακολούθηση και Βελτίωση Μοντέλων**: Τα μοντέλα ΜΜ δεν είναι στατικά και χρειάζονται συνεχιζόμενη παρακολούθηση και αναβάθμιση για να παραμείνουν ακριβή καθώς τα δεδομένα και οι συνθήκες της αγοράς εξελίσσονται. Η τακτική επανεκπαίδευση των μοντέλων με νέα δεδομένα είναι απαραίτητη για την ενίσχυση της ακρίβειας και της αξιοπιστίας τους.

Η σωστή εφαρμογή και χρήση των μεθόδων ΜΜ στην ανάλυση πιστωτικού κινδύνου μπορεί να προσφέρει σημαντικά πλεονεκτήματα, αλλά απαιτεί προσεκτική διαχείριση για να αποφευχθούν τα προβλήματα που σχετίζονται με την ποιότητα των δεδομένων, την ερμηνευσιμότητα των μοντέλων και τους κινδύνους από τις προκαταλήψεις. [23]

Κεφάλαιο 2: Μετρικές στατιστικού ελέγχου και μετρικές σφάλματος

2.1: Μετρικές στατιστικού ελέγχου

Στο παρόν κεφάλαιο γίνεται αναφορά στις μετρικές στατιστικού ελέγχου που χρησιμοποιούνται στη συνέχεια της εργασίας για την προεπεξαργία των δεδομένων καθώς και στις μετρικές σφάλματος που χρησιμοποιούνται για την αξιολόγηση των τεχνικών μηχανικής μάθησης. [24]

- **Μέση Τιμή (Mean)**: Η μέση τιμή ενός συνόλου δεδομένων είναι το άθροισμα όλων των τιμών, διαιρούμενο με το πλήθος των τιμών.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

όπου X_i είναι οι επιμέρους τιμές των δεδομένων και n είναι το πλήθος των παρατηρήσεων.

- **Διάμεσος (Median)**: Η διάμεσος είναι η κεντρική τιμή ενός διατεταγμένου συνόλου δεδομένων. Αν το πλήθος των τιμών είναι περιττό, είναι η μεσαία τιμή. Αν το πλήθος είναι άρτιο, η διάμεσος είναι ο μέσος όρος των δύο κεντρικών τιμών.

- **Διασπορά (Variance)**: Η διασπορά εκφράζει το πόσο απομακρυσμένες είναι οι τιμές από τη μέση τιμή. Για δεδομένα $\mathrm{X_i}$ με μέση τιμή X:

$$\sigma^2 = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}} (\mathrm{X_i} - \overline{\mathrm{X}})^2$$

όπου σ^2 είναι η διασπορά.

- **Τυπική Απόκλιση (Standard Deviation)**: Η τυπική απόκλιση είναι η ρίζα της διασποράς και δείχνει την απόσταση των τιμών από τη μέση τιμή, σε μονάδες της αρχικής κλίμακας.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- **Τυπική Τιμή Z (Z-score)**: Η τυπική τιμή Z μετρά πόσες τυπικές αποκλίσεις απέχει μια παρατήρηση από τη μέση τιμή. Για δεδομένο X_i:

$$Z = \frac{X_i - \bar{X}}{\sigma}$$

όπου Z είναι η τυπική τιμή της παρατήρησης X_i.

- **Πίνακας Συνδιακύμανσης (Covariance Matrix)**: Ο πίνακας συνδιακύμανσης δείχνει την συνδιακύμανση ανάμεσα σε δύο ή περισσότερες τυχαίες μεταβλητές. Για δύο μεταβλητές X και Y, η συνδιακύμανση δίνεται από:

$$\text{cov}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

Ο πίνακας συνδιακύμανσης Σ έχει στη θέση (i, j) τη συνδιακύμανση μεταξύ των μεταβλητών X_i και X_j.

Οι μετρικές αυτές βοηθούν στην κατανόηση της διασποράς και της συσχέτισης των δεδομένων, παρέχοντας το θεωρητικό υπόβαθρο για περαιτέρω ανάλυση. [25]

2.2: Μετρικές σφάλματος

Στη συνέχεια του κεφαλαίου για τις μετρικές σφάλματος, μπορούμε να ορίσουμε τις βασικές μετρικές που χρησιμοποιούνται για την αξιολόγηση της απόδοσης μοντέλων, ιδιαίτερα στην πρόβλεψη συνεχών μεταβλητών. Οι μετρικές αυτές βοηθούν στη μέτρηση της απόκλισης μεταξύ των προβλεπόμενων και των πραγματικών τιμών. [26]

Mean Absolute Error (MAE) – Μέση Απόλυτη Απόκλιση

Η μέση απόλυτη απόκλιση υπολογίζει τον μέσο όρο της απόλυτης απόκλισης μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Χρησιμοποιείται για την εκτίμηση του μέσου σφάλματος χωρίς να λαμβάνεται υπόψη το πρόσημο.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Όπου y_i είναι η πραγματική τιμή και $\hat{y}_i$ είναι η προβλεπόμενη τιμή.

Mean Squared Error (MSE) – Μέσο Τετραγωνικό Σφάλμα

Το μέσο τετραγωνικό σφάλμα υπολογίζει το μέσο όρο του τετραγώνου των αποκλίσεων μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Έχει την τάση να δίνει μεγαλύτερη βαρύτητα στα μεγαλύτερα σφάλματα.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE) – Ριζική Μέση Τετραγωνική Απόκλιση

Η ριζική μέση τετραγωνική απόκλιση είναι η τετραγωνική ρίζα του MSE και δίνει μια εκτίμηση του μέσου σφάλματος σε μονάδες της αρχικής κλίμακας.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}$$

Μέσο Απόλυτο Ποσοστό Σφάλματος – (MAPE)

Η μέση απόλυτη ποσοστιαία απόκλιση εκφράζει το μέσο σφάλμα ως ποσοστό της πραγματικής τιμής, καθιστώντας το ανεξάρτητο από τη μονάδα μέτρησης.

$$\text{MAPE} = \frac{100\backslash\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y_i}}{y_i}\right|$$

Συντελεστής Προσδιορισμού – R-squared (R^2)

Ο συντελεστής προσδιορισμού R^2 δείχνει το ποσοστό της διασποράς των δεδομένων που εξηγείται από το μοντέλο. Ορίζεται ως:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

όπου $\bar{y}$ είναι η μέση τιμή των πραγματικών τιμών. Το R^2 κυμαίνεται από 0 έως 1, με τιμές κοντά στο 1 να δείχνουν υψηλή ακρίβεια του μοντέλου.

Ακρίβεια [27]

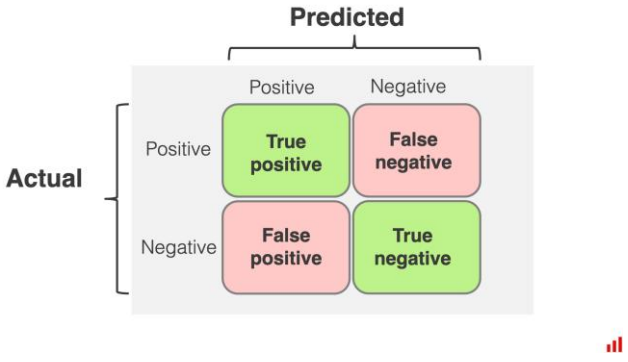
Η ακρίβεια ενός ταξινομητικού μοντέλου υπολογίζεται ως το ποσοστό των σωστών προβλέψεων στο σύνολο των προβλέψεων. Είναι η βασική μετρική απόδοσης για προβλήματα ταξινόμησης.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}$$

Πίνακας Σύγχυσης

Ο πίνακας σύγχυσης είναι ένας πίνακας που εμφανίζει τις προβλέψεις του μοντέλου σε σχέση με τις πραγματικές κλάσεις. Έχει τέσσερα βασικά στοιχεία:

- **True Positives (TP)**: Προβλέψεις θετικών τάξεων που είναι σωστές.
- **True Negatives (TN)**: Προβλέψεις αρνητικών τάξεων που είναι σωστές.
- **False Positives (FP)**: Προβλέψεις θετικών τάξεων που είναι λάθος (σφάλματα τύπου I).
- **False Negatives (FN)**: Προβλέψεις αρνητικών τάξεων που είναι λάθος (σφάλματα τύπου II).



Εικόνα 1: Πίνακας Σύγχυσης [60]

Ο πίνακας βοηθά στην κατανόηση της ακρίβειας του μοντέλου στις διάφορες κατηγορίες.

Τιμή F1

Η τιμή F1 είναι ο αρμονικός μέσος της ακρίβειας (precision) και της ευαισθησίας (recall) και χρησιμοποιείται για να ισορροπήσει μεταξύ αυτών των δύο μετρικών. Ιδιαίτερα χρήσιμη σε περιπτώσεις ανισοκατανομής δεδομένων.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

όπου η Precision υπολογίζεται ως Precision = $\frac{TP}{TP+FP}$ και η Recall ως Recall = $\frac{TP}{TP+FN}$

Ευαισθησία

Η μετρική Ευαισθησία (Recall) δείχνει πόσο καλά εντοπίζονται οι θετικές κλάσεις από το μοντέλο. Μετρά το ποσοστό των πραγματικά θετικών δειγμάτων που το μοντέλο εντόπισε σωστά. Είναι ιδιαίτερα σημαντική σε εφαρμογές όπου τα false negatives είναι πιο σημαντικά, όπως στον ιατρικό τομέα.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP) + False Negatives (FN)}}$$

Ακρίβεια

Η μετρική Ακρίβεια (Precision) δείχνει το ποσοστό των προβλέψεων θετικής κλάσης που ήταν σωστές. Μετρά, δηλαδή, πόσα από τα δείγματα που προβλέφθηκαν ως θετικά ήταν πραγματικά θετικά. Είναι χρήσιμη όταν τα false positives είναι πιο σημαντικά, όπως σε προβλήματα όπου η εσφαλμένη αναγνώριση θετικής κλάσης έχει υψηλό κόστος.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP) + False Positives (FP)}}$$

Support (Υποστήριξη)

Το Support είναι το πλήθος των πραγματικών δειγμάτων σε κάθε κλάση (δηλαδή, το πλήθος των θετικών ή αρνητικών δειγμάτων) και χρησιμεύει για να κατανοήσουμε τη σχετική σημασία των διαφορετικών κλάσεων στο σύνολο των δεδομένων. Αν για παράδειγμα ένα σύνολο δεδομένων έχει ανισοκατανομή μεταξύ κλάσεων, το support για την κάθε κλάση δείχνει τη σημασία της και επηρεάζει την ερμηνεία των άλλων μετρικών (recall, precision).

$$\text{Support} = \text{Number of true samples per class}$$

Αυτές οι μετρικές δίνουν μια σαφέστερη εικόνα για την απόδοση του μοντέλου σε κάθε κλάση, ειδικά σε προβλήματα ανισοκατανομής, και βοηθούν στη λήψη αποφάσεων σχετικά με τη βελτίωση ή προσαρμογή του μοντέλου.

Τα Υπολείμματα στη Στατιστική και στη Μηχανική Μάθηση [28]

Στη στατιστική ανάλυση και στη μηχανική μάθηση, τα υπολείμματα (residuals) αποτελούν βασικό εργαλείο για την αξιολόγηση της ποιότητας ενός μοντέλου. Ο όρος residual αναφέρεται στη διαφορά μεταξύ της πραγματικής τιμής μιας μεταβλητής-στόχου και της αντίστοιχης προβλεφθείσας τιμής που παρέχει ένα μοντέλο. Συγκεκριμένα, αν y_i είναι η πραγματική τιμή και $\widehat{y_i}$ η προβλεφθείσα τιμή για μια συγκεκριμένη παρατήρηση i, τότε το residual ορίζεται ως:

$$e_i = y_i - \widehat{y_i}$$

Χρήση των Residuals στην Ανάλυση Μοντέλων

Η ανάλυση των residuals είναι κρίσιμη για την αξιολόγηση της ποιότητας ενός μοντέλου και την ανίχνευση πιθανών προβλημάτων. Οι κύριες χρήσεις τους περιλαμβάνουν:

1. Έλεγχος της καταλληλότητας του μοντέλου: Εάν τα residuals δεν ακολουθούν συγκεκριμένες στατιστικές ιδιότητες, το μοντέλο ενδέχεται να μην είναι το κατάλληλο για τα δεδομένα. Για παράδειγμα, σε ένα γραμμικό μοντέλο, αναμένουμε τα residuals να κατανέμονται κανονικά γύρω από το μηδέν χωρίς συστηματικά μοτίβα.

2. Διάγνωση ετεροσκεδαστικότητας: Αν τα residuals εμφανίζουν μεταβαλλόμενη διακύμανση (heteroscedasticity), δηλαδή γίνονται μεγαλύτερα ή μικρότερα καθώς αυξάνονται οι προβλεφθείσες τιμές, αυτό μπορεί να σημαίνει ότι το μοντέλο δεν έχει συμπεριλάβει σωστά όλες τις σχετικές μεταβλητές ή ότι η σχέση μεταξύ των μεταβλητών δεν είναι γραμμική.

3. Ανίχνευση outliers και σφαλμάτων: Παρατηρήσεις με εξαιρετικά μεγάλα residuals (outliers) ενδέχεται να υποδεικνύουν λάθη στη συλλογή δεδομένων, την παρουσία ακραίων τιμών ή ότι το μοντέλο δεν έχει καταγράψει κάποια σημαντική μεταβλητή που επηρεάζει το αποτέλεσμα.

4. Ανίχνευση μη γραμμικών σχέσεων: Αν τα residuals δείχνουν κάποιο μη τυχαίο μοτίβο, μπορεί να σημαίνει ότι η πραγματική σχέση μεταξύ των μεταβλητών δεν είναι γραμμική. Σε αυτή την περίπτωση, μπορεί να χρειαστεί να χρησιμοποιηθεί ένα πιο σύνθετο μοντέλο, όπως ένα πολυωνυμικό ή μη γραμμικό μοντέλο.

Η σωστή ανάλυση των residuals μας επιτρέπει να κατανοήσουμε τις αδυναμίες και τα δυνατά σημεία ενός μοντέλου πρόβλεψης. Εάν τα residuals κατανέμονται ομοιόμορφα γύρω από το μηδέν, χωρίς κάποιο εμφανές μοτίβο, μπορούμε να υποθέσουμε ότι το μοντέλο περιγράφει καλά τα δεδομένα. Αντίθετα, αν παρατηρούμε μοτίβα, αλλαγές στη διακύμανση ή ασυνήθιστα υψηλά υπολείμματα, αυτό σημαίνει ότι το μοντέλο χρειάζεται βελτίωση.

Σε εφαρμογές μηχανικής μάθησης, η μείωση των residuals αποτελεί κεντρικό στόχο κατά τη βελτίωση ενός μοντέλου. Τεχνικές όπως η χρήση πιο πολύπλοκων αλγορίθμων, η βελτίωση της επιλογής χαρακτηριστικών, η μείωση της διάστασης ή η εφαρμογή τεχνικών κανονικοποίησης μπορούν να συμβάλουν στη μείωση των residuals και, συνεπώς, στη βελτίωση της ακρίβειας των προβλέψεων.

Κεφάλαιο 3: Επιβλεπόμενη και μη Επιβλεπόμενη Μηχανική Μάθηση

Η μηχανική μάθηση παίζει καθοριστικό ρόλο στην πιστωτική βαθμολόγηση και την εκτίμηση του πιστωτικού κινδύνου, επιτρέποντας στις τράπεζες και σε άλλους χρηματοπιστωτικούς οργανισμούς να αξιολογούν με ακρίβεια την πιθανότητα αθέτησης πληρωμών από πελάτες. Χρησιμοποιώντας μεγάλα σύνολα δεδομένων –όπως οικονομικά χαρακτηριστικά, ιστορικό πληρωμών, και δημογραφικές πληροφορίες– η μηχανική μάθηση μπορεί να ανιχνεύει σύνθετα μοτίβα και να δημιουργεί μοντέλα που προβλέπουν την πιστοληπτική συμπεριφορά με μεγαλύτερη ακρίβεια από τις παραδοσιακές μεθόδους. [29]

Οι δυνατότητες της μηχανικής μάθησης για την ανάλυση πιστωτικού κινδύνου είναι σημαντικές: προσφέρει μεγαλύτερη ταχύτητα και ακρίβεια στις αποφάσεις, επιτρέπει τη χρήση δεδομένων σε πραγματικό χρόνο και συμβάλλει στην καλύτερη διαχείριση των κινδύνων, μειώνοντας τα ποσοστά χρεοκοπίας και βελτιώνοντας την εμπιστοσύνη στον πιστωτικό τομέα.

Ωστόσο, η χρήση της μηχανικής μάθησης εγκυμονεί κινδύνους. Οι αλγόριθμοι είναι ευαίσθητοι στα δεδομένα που χρησιμοποιούνται για την εκπαίδευσή τους, και η μεροληψία (bias) των δεδομένων μπορεί να οδηγήσει σε αθέμιτες πρακτικές, όπως διακρίσεις σε βάρος συγκεκριμένων ομάδων. Επιπλέον, η έλλειψη διαφάνειας στους αλγορίθμους ενδέχεται να δημιουργήσει προβλήματα κατανόησης και εμπιστοσύνης από τους χρήστες και τους πελάτες. Ως εκ τούτου, η χρήση μηχανικής μάθησης στην πιστωτική αξιολόγηση απαιτεί προσεκτική διαχείριση, ώστε να διασφαλίζεται η ακρίβεια, η διαφάνεια και η αμεροληψία των προβλέψεων.

3.1 Επιβλεπόμενη μηχανική μάθηση

Η επιβλεπόμενη μηχανική μάθηση είναι μια μέθοδος κατά την οποία ένα μοντέλο εκπαιδεύεται να κάνει προβλέψεις ή να αναγνωρίζει μοτίβα βασισμένα σε δεδομένα που έχουν ήδη κατηγοριοποιηθεί ή "επισημανθεί" από πριν. Ουσιαστικά, η επιβλεπόμενη μάθηση απαιτεί ένα σύνολο δεδομένων εκπαίδευσης, το οποίο περιλαμβάνει εισόδους (χαρακτηριστικά) και τις αντίστοιχες εξόδους (ετικέτες). Με βάση αυτά τα δεδομένα, το μοντέλο μαθαίνει να αντιστοιχίζει τις εισόδους με τις επιθυμητές εξόδους, με στόχο να γενικεύει και να προβλέπει με ακρίβεια για νέα, άγνωστα δεδομένα. [30]

Στόχοι της Επιβλεπόμενης Μηχανικής Μάθησης

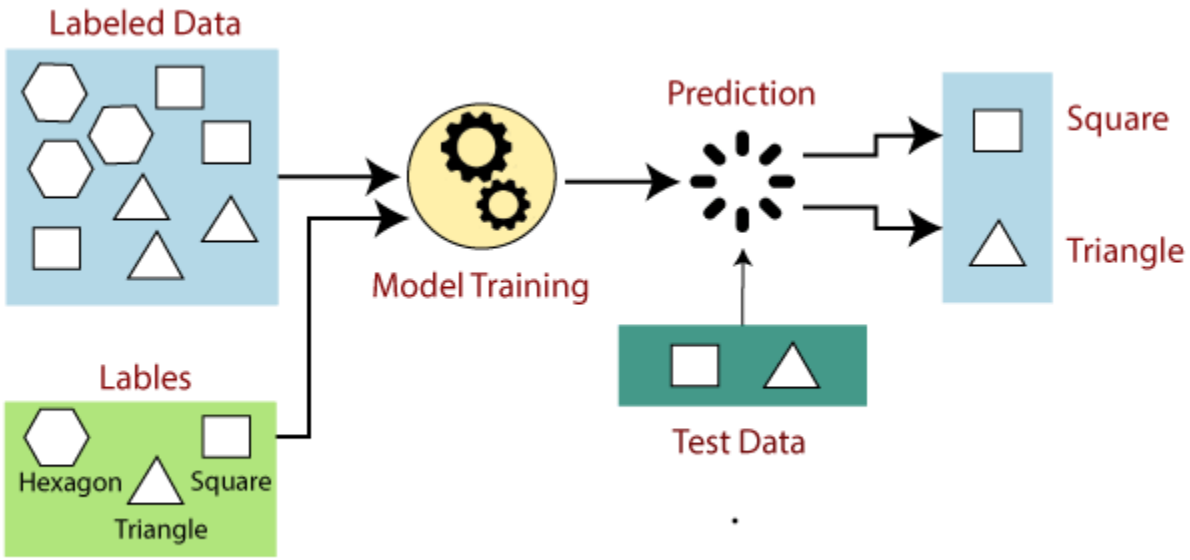
Ο βασικός στόχος της επιβλεπόμενης μηχανικής μάθησης είναι η πρόβλεψη αποτελεσμάτων. Τα μοντέλα που αναπτύσσονται μπορούν να εξυπηρετήσουν διαφορετικούς στόχους, όπως:

- Ταξινόμηση (Classification): όπου τα δεδομένα χωρίζονται σε κατηγορίες. Ένα παράδειγμα θα ήταν ένα μοντέλο που προβλέπει αν ένας πελάτης μιας τράπεζας είναι πιθανό να χρεοκοπήσει ή όχι.
- Παλινδρόμηση (Regression): όπου το μοντέλο κάνει προβλέψεις σε συνεχή δεδομένα. Για παράδειγμα, ένα μοντέλο μπορεί να προβλέπει το ύψος των αναμενόμενων ζημιών από αθέτηση πληρωμής.

Διαδικασία Εκπαίδευσης

Η διαδικασία εκπαίδευσης στην επιβλεπόμενη μάθηση περιλαμβάνει τρία βασικά στάδια [31]:

1. Συλλογή και προετοιμασία δεδομένων: Τα δεδομένα πρέπει να είναι όσο το δυνατόν πιο αντιπροσωπευτικά του προβλήματος και να περιλαμβάνουν τις πραγματικές τιμές εξόδου για κάθε παρατήρηση.
2. Εκπαίδευση του μοντέλου: Με τα δεδομένα αυτά, το μοντέλο χρησιμοποιεί αλγορίθμους για να μάθει την αντιστοιχία μεταξύ των εισόδων και των εξόδων.
3. Αξιολόγηση της απόδοσης: Με την αξιολόγηση σε νέα, άγνωστα δεδομένα (δεδομένα επαλήθευσης ή ελέγχου), το μοντέλο ελέγχεται για την ικανότητά του να γενικεύει.



Εικόνα 2: Επιβλεπόμενη Μάθηση [54]

Παράδειγμα Επιβλεπόμενης Μάθησης: Πρόβλεψη Πιστωτικού Κινδύνου

Ένα χαρακτηριστικό παράδειγμα επιβλεπόμενης μάθησης είναι η πρόβλεψη του πιστωτικού κινδύνου, το οποίο αφορά την εκτίμηση του κατά πόσο ένας πελάτης μιας τράπεζας είναι πιθανό να αθετήσει την αποπληρωμή ενός δανείου. Σε αυτήν την περίπτωση, η τράπεζα διαθέτει ιστορικά δεδομένα πελατών, τα οποία περιλαμβάνουν πληροφορίες όπως το εισόδημα του πελάτη, την ηλικία, το πιστωτικό ιστορικό και το εάν ο πελάτης έχει αθετήσει πληρωμές στο παρελθόν.

Με αυτά τα δεδομένα, δημιουργείται ένα μοντέλο επιβλεπόμενης μάθησης που μπορεί να προβλέψει, με βάση τα χαρακτηριστικά ενός νέου πελάτη, αν είναι πιθανό να αθετήσει την πληρωμή ή όχι. Ο στόχος εδώ είναι η ταξινόμηση: το μοντέλο πρέπει να κατατάξει τον πελάτη σε μία από δύο κατηγορίες (καλή πιστοληπτική ικανότητα ή πιθανή αθέτηση). Τα δεδομένα εκπαίδευσης περιλαμβάνουν τις εισόδους (χαρακτηριστικά πελάτη) και την αντίστοιχη έξοδο (αποτέλεσμα πιστοληπτικής συμπεριφοράς), επιτρέποντας στο μοντέλο να μάθει την αντιστοιχία.

3.2 Μη Επιβλεπόμενη Μηχανική Μάθηση

Η μη επιβλεπόμενη μάθηση είναι μια κατηγορία αλγορίθμων της μηχανικής μάθησης που εκπαιδεύεται χωρίς να διαθέτει προκαθορισμένες "ετικέτες" στα δεδομένα εκπαίδευσης. Αντί να μαθαίνει μέσω συσχετίσεων ανάμεσα σε εισόδους και προκαθορισμένες εξόδους, όπως στην επιβλεπόμενη μάθηση, η μη επιβλεπόμενη μάθηση στοχεύει στον εντοπισμό κρυφών μοτίβων και σχέσεων μέσα στα δεδομένα. Οι αλγόριθμοι αυτής της κατηγορίας οργανώνουν τις εισόδους, είτε ομαδοποιώντας παρόμοια δεδομένα (clustering) είτε ανακαλύπτοντας τη δομή των δεδομένων. [32]

Οι αλγόριθμοι μη επιβλεπόμενης μάθησης παρέχουν πληροφορίες που μπορεί να είναι πολύτιμες σε αναλύσεις, ιδιαίτερα όπου οι ετικέτες των δεδομένων είναι περιορισμένες ή άγνωστες. Για παράδειγμα, οι αλγόριθμοι ομαδοποίησης (όπως το K-means) χωρίζουν τα δεδομένα σε ομάδες με βάση την ομοιότητά τους, επιτρέποντας την κατηγοριοποίηση πελατών ή την ανίχνευση μη φυσιολογικών μοτίβων σε συναλλαγές, που θα μπορούσαν να υποδηλώνουν ενδεχόμενο απάτης.

Στον τομέα της διαχείρισης πιστωτικού κινδύνου, η μη επιβλεπόμενη μάθηση χρησιμοποιείται κυρίως σε προβλήματα όπως η ανίχνευση απάτης και η τμηματοποίηση πελατών. Μέσω της ομαδοποίησης πελατών με παρόμοια χαρακτηριστικά ή συμπεριφορές, οι τράπεζες μπορούν να

κατανοήσουν καλύτερα τα χαρακτηριστικά διαφόρων ομάδων πελατών και να προσαρμόσουν τα προϊόντα και τις στρατηγικές κινδύνου τους. [33]

Στην παρούσα εργασία, η οποία εστιάζει στην πρόβλεψη της ζημίας λόγω αθέτησης, δεν χρησιμοποιούνται μέθοδοι μη επιβλεπόμενης μηχανικής μάθησης, καθώς η φύση του προβλήματος απαιτεί ακριβή εκτίμηση των ετικετών εξόδου (όπως το αν ο πελάτης θα αποπληρώσει ή όχι). Παρόλα αυτά, η αναφορά της μη επιβλεπόμενης μάθησης γίνεται για λόγους πληρότητας και καλύτερης κατανόησης του πεδίου της μηχανικής μάθησης.

3.3 Μέθοδοι Ταξινόμησης

Οι μέθοδοι ταξινόμησης αποτελούν κεντρικό κομμάτι της επιβλεπόμενης μηχανικής μάθησης, με στόχο να κατατάσσουν δεδομένα σε προκαθορισμένες κατηγορίες. Η βασική ιδέα πίσω από αυτούς τους αλγορίθμους είναι να δημιουργηθεί ένα μοντέλο που, βασισμένο σε ένα σύνολο δεδομένων εκπαίδευσης, μπορεί να μάθει και να γενικεύσει μοτίβα στα δεδομένα, επιτρέποντας την ακριβή ταξινόμηση νέων δεδομένων σε συγκεκριμένες κατηγορίες. [34]

Οι αλγόριθμοι ταξινόμησης χρησιμοποιούν ως εισόδους ένα σύνολο χαρακτηριστικών (features), τα οποία περιγράφουν κάθε παρατήρηση. Αυτά τα χαρακτηριστικά μπορεί να είναι δημογραφικά στοιχεία, οικονομικά δεδομένα ή οποιαδήποτε άλλη πληροφορία σχετίζεται με το πρόβλημα. Το σύνολο των παρατηρήσεων έχει επίσης μια ετικέτα κατηγορίας (class label), που προσδιορίζει την κατηγορία στην οποία ανήκει κάθε παρατήρηση στο σύνολο εκπαίδευσης.

Η έξοδος ενός μοντέλου ταξινόμησης είναι μια κατηγορία ή μια πιθανότητα που αντιστοιχεί στην κατηγορία στην οποία ανήκει μια νέα παρατήρηση. Με άλλα λόγια, το μοντέλο καλείται να "μαντέψει" σε ποια κατηγορία θα τοποθετηθεί μια νέα είσοδος, βάσει των χαρακτηριστικών της.

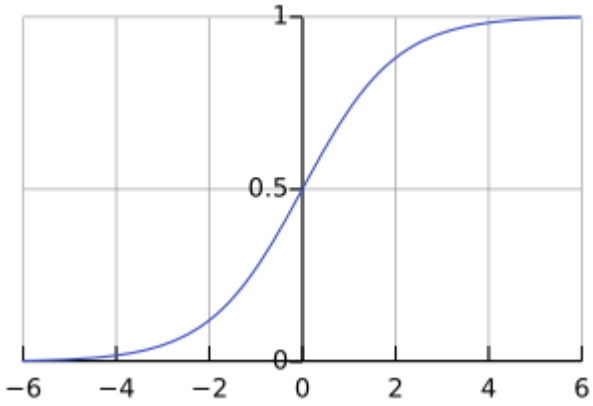
Ο στόχος της ταξινόμησης είναι να επιτυγχάνεται υψηλή ακρίβεια σε νέες, άγνωστες παρατηρήσεις. Οι αλγόριθμοι εκπαιδεύονται ώστε να μειώνουν τα σφάλματα πρόβλεψης και να εξασφαλίζουν ότι η ταξινόμηση είναι αξιόπιστη και συνεπής, γεγονός που καθιστά τους αλγόριθμους ταξινόμησης ιδανικούς για προβλήματα όπου απαιτείται σαφής διαχωρισμός, όπως η πρόβλεψη της αθέτησης πληρωμών ή η ανίχνευση απάτης.

3.3.1 Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση είναι μια δημοφιλής μέθοδος ταξινόμησης που χρησιμοποιείται ευρέως σε προβλήματα όπου η έξοδος πρέπει να ανήκει σε μία από δύο κατηγορίες, όπως η εκτίμηση πιστωτικού κινδύνου. Στη λογιστική παλινδρόμηση, ο στόχος είναι να μοντελοποιηθεί η πιθανότητα η παρατήρηση να ανήκει σε μια συγκεκριμένη κατηγορία, π.χ., αν ένας πελάτης μιας τράπεζας θα αποπληρώσει ή όχι ένα δάνειο. [35]

Πώς δουλεύει η Μέθοδος της Λογιστικής Παλινδρόμησης;

Η λογιστική παλινδρόμηση χρησιμοποιεί μια συνάρτηση, γνωστή ως συνάρτηση λογιστικής ή σιγμοειδούς συνάρτησης (sigmoid function), για να μετατρέψει τις προβλέψεις του μοντέλου σε πιθανότητες. Η σιγμοειδής συνάρτηση έχει ως στόχο να μετατρέπει οποιαδήποτε πραγματική τιμή σε μια τιμή εντός του διαστήματος (0, 1), επιτρέποντας την εκτίμηση της πιθανότητας ενός δυαδικού αποτελέσματος.



Εικόνα 3: Λογιστική Συνάρτηση [55]

Μαθηματικό Υπόβαθρο

Η λογιστική παλινδρόμηση [36] εκτιμά την πιθανότητα μιας παρατήρησης να ανήκει στην κατηγορία "1" (π.χ., ο πελάτης θα αθετήσει το δάνειο), με βάση ένα σύνολο εισόδων (χαρακτηριστικά), που ονομάζονται $x_1, x_2, \dots, x_n$. Το μοντέλο εκτιμά αυτήν την πιθανότητα ως εξής:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}}$$

όπου:

- Y είναι η δυαδική μεταβλητή εξόδου που παίρνει τις τιμές 0 ή 1.
- X είναι το σύνολο των χαρακτηριστικών εισόδου.
- $\beta_0, \beta_1, \ldots, \beta_n$ είναι οι συντελεστές του μοντέλου που εκτιμώνται κατά τη διαδικασία εκπαίδευσης.

Η λογιστική παλινδρόμηση υπολογίζει το λογάριθμο των πιθανοτήτων (log-odds) ως γραμμική συνάρτηση των εισόδων:

$$\text{log-odds} = \ln\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Αυτό σημαίνει ότι οι πιθανότητες της κατηγορίας "1" αυξάνονται εκθετικά με βάση τις τιμές των χαρακτηριστικών.

3.3.2 Τυχαίο Δάσος

Η μέθοδος του τυχαίου δάσους (Random Forest) είναι ένας ισχυρός αλγόριθμος μηχανικής μάθησης, ιδιαίτερα δημοφιλής για προβλήματα ταξινόμησης και παλινδρόμησης. Η κεντρική ιδέα της μεθόδου βασίζεται στη δημιουργία ενός συνόλου από δέντρα απόφασης και στη συνδυασμένη πρόβλεψη αυτών των δέντρων. Ας αναλύσουμε πρώτα τι είναι ένα δέντρο απόφασης και πώς λειτουργεί πριν δούμε την προσέγγιση του τυχαίου δάσους. [37]

<u>Τι είναι το Δέντρο Απόφασης;</u>

Το δέντρο απόφασης είναι ένας αλγόριθμος που χωρίζει τα δεδομένα σε επαναλαμβανόμενα υποσύνολα για να φτάσει σε μια απόφαση σχετικά με την ταξινόμηση ή την πρόβλεψη ενός αποτελέσματος. Το δέντρο αποτελείται από κόμβους [38]:

- Ριζικός κόμβος: Ο κόμβος που περιλαμβάνει όλα τα δεδομένα και από τον οποίο ξεκινά ο διαχωρισμός.
- Εσωτερικοί κόμβοι: Οι κόμβοι όπου γίνονται διαχωρισμοί βάσει χαρακτηριστικών.
- Φύλλα: Οι τερματικοί κόμβοι, στους οποίους καταλήγουν οι παρατηρήσεις και αντιπροσωπεύουν την τελική κατηγορία ή την πρόβλεψη.

Εκπαίδευση ενός Δέντρου Απόφασης

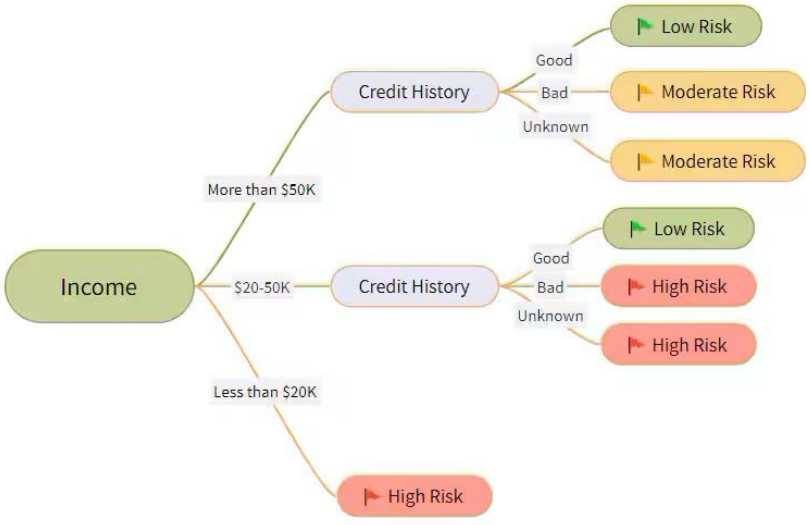
Η εκπαίδευση ενός δέντρου απόφασης περιλαμβάνει τη διαίρεση των δεδομένων σε διαφορετικούς κόμβους με βάση τα χαρακτηριστικά που προσφέρουν τη μέγιστη πληροφόρηση. Το κριτήριο επιλογής βασίζεται συχνά σε μετρήσεις όπως:

- Εντροπία και κέρδος πληροφορίας (Information Gain), που μετρά την ομοιογένεια των δεδομένων μετά από έναν διαχωρισμό.
- Gini index, που μετρά την καθαρότητα των υποσυνόλων μετά από κάθε διαχωρισμό.

Η διαδικασία συνεχίζεται επαναλαμβανόμενα έως ότου οι κόμβοι φτάσουν σε ένα προκαθορισμένο βάθος ή οι διαχωρισμοί δεν προσφέρουν επιπλέον πληροφορία. Ωστόσο, ένα μόνο δέντρο μπορεί να είναι ευαίσθητο στην υπερπροσαρμογή (overfitting), αν ταιριάζει υπερβολικά στα δεδομένα εκπαίδευσης.

Τι είναι ο Αλγόριθμος Τυχαίου Δάσους;

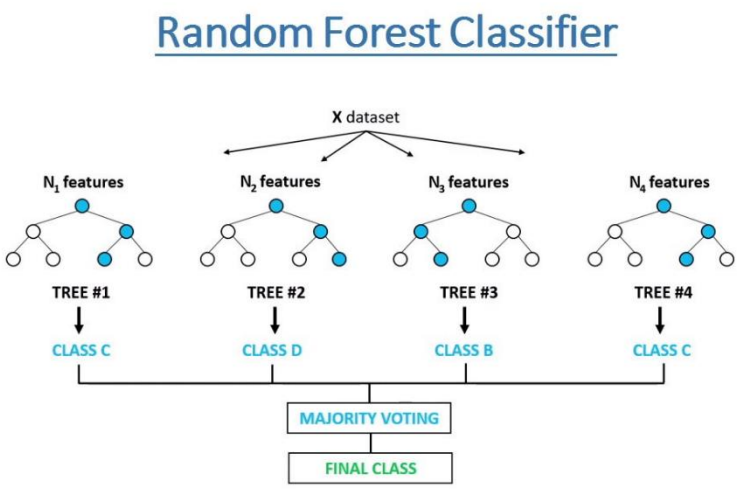
Το τυχαίο δάσος επιλύει το πρόβλημα της υπερπροσαρμογής ενός μεμονωμένου δέντρου απόφασης συνδυάζοντας πολλά δέντρα απόφασης σε ένα σύνολο (ensemble). Βασίζεται στην τεχνική του bagging (bootstrap aggregating) και στη χρήση τυχαίων δειγμάτων δεδομένων και χαρακτηριστικών για την εκπαίδευση κάθε δέντρου. Το αποτέλεσμα του συνόλου των δέντρων είναι η τελική πρόβλεψη, η οποία προκύπτει από τον μέσο όρο (σε προβλήματα παλινδρόμησης) ή την πλειοψηφική ψήφο (σε προβλήματα ταξινόμησης) όλων των δέντρων.



Εικόνα 4: Παράδειγμα Δέντρου Απόφασης [56]

Πώς Λειτουργεί το Τυχαίο Δάσος

1. Δημιουργία Δειγμάτων Εκπαίδευσης: Από το σύνολο των δεδομένων δημιουργούνται πολλαπλά δείγματα μέσω επαναληπτικής επιλογής (bootstrap sampling). Κάθε δείγμα εκπαιδεύει ένα διαφορετικό δέντρο.
2. Επιλογή Τυχαίων Χαρακτηριστικών: Σε κάθε κόμβο ενός δέντρου, επιλέγεται ένα υποσύνολο χαρακτηριστικών από το πλήρες σύνολο. Ο αλγόριθμος αποφασίζει τον βέλτιστο διαχωρισμό βάσει των επιλεγμένων χαρακτηριστικών.
3. Συνδυασμός των Αποτελεσμάτων: Αφού εκπαιδευτούν όλα τα δέντρα, η τελική πρόβλεψη του τυχαίου δάσους προκύπτει από την ψήφο όλων των δέντρων (σε ταξινόμηση) ή τον μέσο όρο των προβλέψεων (σε παλινδρόμηση)



Εικόνα 5: Τυχαίο Δάσος [57]

3.3.3 Extreme Gradient Boosting

Ο αλγόριθμος Extreme Gradient Boosting (XGBoost) είναι μια ισχυρή τεχνική ενίσχυσης (boosting) που χρησιμοποιείται κυρίως σε προβλήματα ταξινόμησης και παλινδρόμησης. Η κύρια φιλοσοφία του είναι η βελτίωση των προβλέψεων μέσω της κατασκευής ενός συνόλου αδύναμων μοντέλων (συνήθως δέντρων απόφασης) με τρόπο που εστιάζει συνεχώς στα σφάλματα των προηγούμενων προβλέψεων. [39]

Πώς Λειτουργεί ο Αλγόριθμος

1. Ο αλγόριθμος XGBoost ανήκει στην οικογένεια των αλγορίθμων gradient boosting και χρησιμοποιεί δέντρα απόφασης ως αδύναμα μοντέλα. Η βασική διαδικασία είναι η εξής:
2. Αρχικό Μοντέλο: Ξεκινάει με ένα αρχικό μοντέλο που μπορεί να είναι είτε ένα σταθερό μοντέλο (όπως η μέση τιμή για παλινδρόμηση) είτε ένα απλό δέντρο.
3. Διαδοχική Κατασκευή Νέων Δέντρων: Κάθε νέο δέντρο προστίθεται για να διορθώσει τα σφάλματα του συνόλου των προηγούμενων δέντρων. Η προσέγγιση αυτή βασίζεται στην ελαχιστοποίηση της συνάρτησης απώλειας μέσω της μέθοδος της κλίσης (gradient descent).
4. Εκπαίδευση σε Σφάλματα: Το κάθε δέντρο εκπαιδεύεται για να ελαχιστοποιήσει την απώλεια, με έμφαση στα παραδείγματα που το προηγούμενο μοντέλο δεν κατάφερε να ταξινομήσει σωστά.
5. Συνολική Πρόβλεψη: Η τελική πρόβλεψη προκύπτει ως το άθροισμα των προβλέψεων όλων των δέντρων, επιτυγχάνοντας σταδιακή βελτίωση.

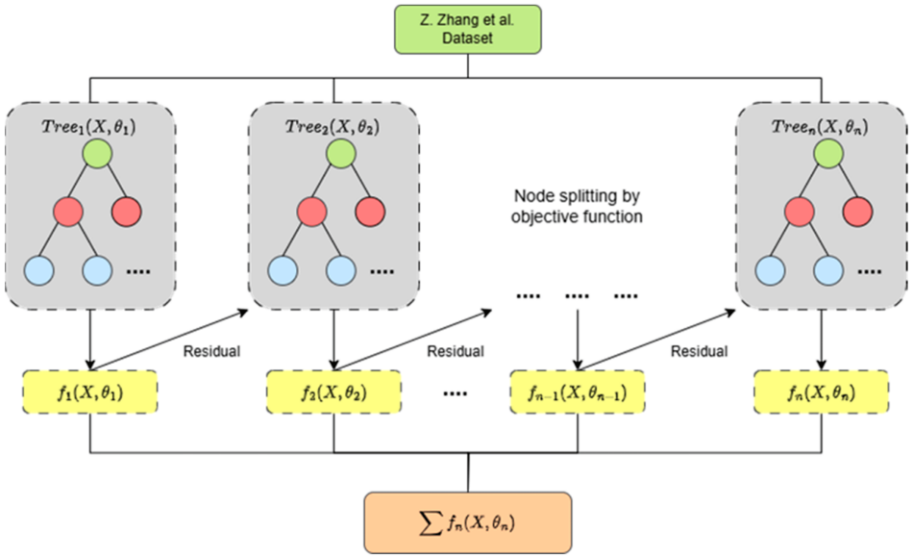
Ο αλγόριθμος επιδιώκει επίσης την εύρεση της βέλτιστης δομής για κάθε νέο δέντρο, λαμβάνοντας υπόψη παράγοντες όπως το βάθος του δέντρου και την πολυπλοκότητά του.

Κύριες Παράμετροι του XGBoost [40]

- Η επίδοση του XGBoost εξαρτάται από ορισμένες σημαντικές παραμέτρους:
- Μέγιστο βάθος δέντρων (max_depth): Καθορίζει πόσο πολύπλοκα μπορούν να είναι τα δέντρα. Τα βαθύτερα δέντρα ενδέχεται να ταιριάξουν καλύτερα τα δεδομένα εκπαίδευσης αλλά αυξάνουν τον κίνδυνο υπερπροσαρμογής.
- Ρυθμός εκμάθησης (learning rate): Ορίζει πόσο μεγάλη είναι η συνεισφορά κάθε νέου δέντρου στο τελικό μοντέλο. Μικρότερες τιμές του learning rate βελτιώνουν την απόδοση, αλλά απαιτούν περισσότερα δέντρα.
- Αριθμός δέντρων (n_estimators): Το πλήθος των δέντρων που θα εκπαιδευτούν. Πιο πολλά δέντρα συνήθως βελτιώνουν την απόδοση αλλά αυξάνουν τον χρόνο εκπαίδευσης.
- Subsample: Η τυχαία δειγματοληψία του συνόλου των δεδομένων για κάθε δέντρο. Χρησιμοποιείται για την αποφυγή της υπερπροσαρμογής.
- Γενική παράμετρος κανονικοποίησης (regularization): Περιλαμβάνει παραμέτρους όπως λ και α, οι οποίες περιορίζουν την πολυπλοκότητα του μοντέλου, μειώνοντας την τάση για υπερπροσαρμογή.

<u>Δυνατότητες και Πλεονεκτήματα</u>

Το XGBoost είναι γνωστό για την ταχύτητά του και την ικανότητά του να χειρίζεται μεγάλες ποσότητες δεδομένων λόγω της βελτιστοποίησης στη χρήση μνήμης και της υποστήριξης για παράλληλη επεξεργασία. Παράλληλα, είναι ανθεκτικό στην υπερπροσαρμογή, κυρίως λόγω των τεχνικών κανονικοποίησης που ενσωματώνει.



Εικόνα 6: Αλγόριθμος XGBoost [58]

3.3.4 Βαθιά Νευρωνικά Δίκτυα

Τα βαθιά νευρωνικά δίκτυα (Deep Neural Networks, DNNs) είναι μια κατηγορία αλγορίθμων μηχανικής μάθησης εμπνευσμένων από τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου. Εφαρμόζονται ευρέως σε προβλήματα κατηγοριοποίησης, όπου το ζητούμενο είναι η ταξινόμηση παρατηρήσεων σε προκαθορισμένες κατηγορίες. Η δομή των βαθιών νευρωνικών δικτύων βασίζεται σε διαδοχικά επίπεδα, καθένα από τα οποία αποτελείται από πολλούς τεχνητούς νευρώνες, που συνεργάζονται για την επεξεργασία σύνθετων σχέσεων και μοτίβων στα δεδομένα. [41]

Πώς Δουλεύουν τα Νευρωνικά Δίκτυα

Τα νευρωνικά δίκτυα αποτελούνται από εισόδους, κρυφά επίπεδα και έξοδο. Τα δεδομένα εισάγονται στο δίκτυο μέσω των νευρώνων του εισόδου, επεξεργάζονται στα κρυφά επίπεδα και αποδίδουν ένα αποτέλεσμα στον νευρώνα εξόδου. Τα βαθιά νευρωνικά δίκτυα έχουν πολλά κρυφά επίπεδα, γεγονός που τους επιτρέπει να αναγνωρίζουν σύνθετα μοτίβα.

Δομή και Λειτουργία ενός Νευρώνα

Κάθε νευρώνας [42] σε ένα δίκτυο είναι μια μονάδα επεξεργασίας που δέχεται πολλαπλές εισόδους, τις οποίες ζυγίζει και στη συνέχεια τις συνδυάζει σε μία έξοδο, μέσω μιας συνάρτησης ενεργοποίησης. Για να περιγράψουμε τη λειτουργία ενός νευρώνα, μπορούμε να εξετάσουμε τα εξής:

- Έστω ότι ένας νευρώνας έχει είσοδο x_i από κάθε συνδεδεμένο νευρώνα του προηγούμενου επιπέδου.
- Κάθε είσοδος x_i έχει έναν συντελεστή βάρους w_i, ο οποίος καθορίζει τη σημασία της συγκεκριμένης εισόδου.
- Ο νευρώνας υπολογίζει το ζυγισμένο άθροισμα των εισόδων ως:

$$z = \sum_{\{i=1\}}^{n} w_i x_i + b$$

όπου b είναι ο όρος μεροληψίας που βοηθά στην προσαρμογή του αποτελέσματος της συνάρτησης.

- Το z υποβάλλεται σε μια μη γραμμική συνάρτηση ενεργοποίησης $\sigma(z)$ η οποία επιτρέπει στο νευρωνικό δίκτυο να μάθει πολύπλοκες σχέσεις. Συνηθισμένες συναρτήσεις

ενεργοποίησης είναι η ReLU (Rectified Linear Unit), η Sigmoid, και η Softmax (για προβλήματα κατηγοριοποίησης).

Μαθηματικό Υπόβαθρο της Συνάρτησης Ενεργοποίησης

Για ένα πρόβλημα κατηγοριοποίησης, χρησιμοποιούμε συχνά τη Softmax στον τελευταίο επίπεδο. Η Softmax δίνει την πιθανότητα να ανήκει η είσοδος σε μια συγκεκριμένη κατηγορία

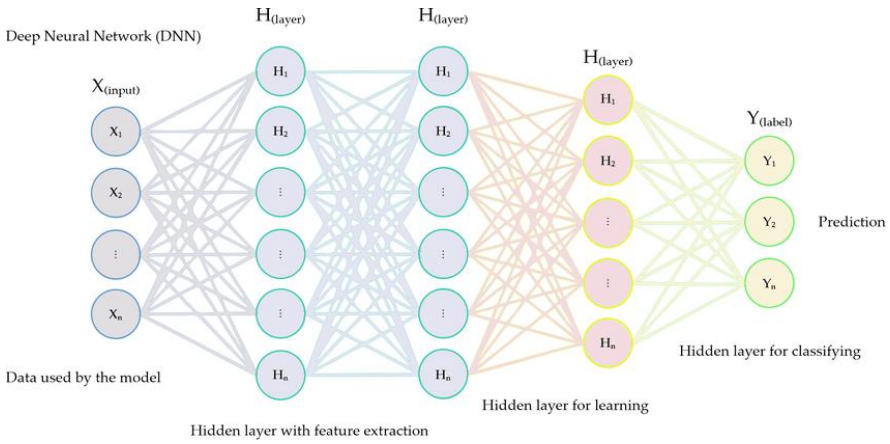
$$\text{Softmax}(z_k) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}}$$

όπου K είναι το πλήθος των κατηγοριών και z_k το ζυγισμένο άθροισμα των εισόδων για την κατηγορία K. Αυτή η εξίσωση διασφαλίζει ότι οι εξόδους είναι θετικές και αθροίζουν στη μονάδα.

Εκπαίδευση Νευρωνικών Δικτύων

Η εκπαίδευση ενός νευρωνικού δικτύου πραγματοποιείται μέσω της προς τα πίσω διάδοσης (backpropagation) και της βελτιστοποίησης του βαθμιαίου κατηφορικού αλγορίθμου (gradient descent). Ο στόχος είναι η ελαχιστοποίηση μιας συνάρτησης απώλειας, η οποία μετρά την απόκλιση των προβλέψεων του μοντέλου από τις πραγματικές τιμές. Κατά την εκπαίδευση:

1. Πρόβλεψη (Forward Pass): Τα δεδομένα εισόδου περνούν από το δίκτυο για την παραγωγή μιας πρόβλεψης.
2. Υπολογισμός Σφάλματος: Η συνάρτηση απώλειας μετρά την απόκλιση της πρόβλεψης.
3. Προς τα Πίσω Διάδοση: Υπολογίζεται ο ρυθμός μεταβολής του σφάλματος (gradient) ως προς κάθε βάρος.
4. Ενημέρωση Βαρών: Το μοντέλο προσαρμόζει τα βάρη και τους όρους μεροληψίας του χρησιμοποιώντας το gradient, βελτιώνοντας τις προβλέψεις.



Εικόνα 7: Παράδειγμα Βαθύ Νευρωνικού Δικτύου [62]

3.4 Μέθοδοι Παλινδρόμησης

Η πάλινδρόμηση είναι μια τεχνική μηχανικής μάθησης που χρησιμοποιείται για την πρόβλεψη μιας συνεχούς τιμής με βάση δεδομένα εισόδου. Στο πλαίσιο των προβλημάτων παλινδρόμησης, η στόχευση είναι να μοντελοποιηθεί η σχέση μεταξύ των εισαγωγικών χαρακτηριστικών (ή ανεξάρτητων μεταβλητών) και της εξαρτημένης μεταβλητής, η οποία είναι μια συνεχής τιμή που προσπαθούμε να προβλέψουμε.

Είσοδοι και Έξοδοι

- Είσοδοι: Στα προβλήματα παλινδρόμησης, οι είσοδοι συνήθως αποτελούνται από ένα ή περισσότερα χαρακτηριστικά που περιγράφουν το πρόβλημα. Αυτά τα χαρακτηριστικά μπορεί να είναι οποιαδήποτε δεδομένα που είναι γνωστά και με τα οποία θέλουμε να κάνουμε πρόβλεψη. Για παράδειγμα, σε ένα πρόβλημα πρόβλεψης της αξίας ενός ακινήτου, τα χαρακτηριστικά μπορεί να είναι η επιφάνεια του ακινήτου, η περιοχή, η ηλικία του κτηρίου, κ.λπ.
- Έξοδοι: Η έξοδος σε ένα πρόβλημα παλινδρόμησης είναι μια συνεχής τιμή, η οποία εξαρτάται από τις εισόδους. Συνεπώς, ο στόχος είναι να εκπαιδεύσουμε ένα μοντέλο που μπορεί να δώσει μια ακριβή εκτίμηση αυτής της τιμής. Για παράδειγμα, στην πρόβλεψη του πιστωτικού κινδύνου, η έξοδος θα μπορούσε να είναι η ζημία λόγω αθέτησης για έναν πελάτη, ή η αξία που αναμένεται να ανακτηθεί από μια εταιρεία σε περίπτωση αθέτησης.

Οι μέθοδοι παλινδρόμησης εκπαιδεύονται ώστε να μάθουν την κατάλληλη συνάρτηση που συσχετίζει τις εισόδους με τις εξόδους, ώστε να κάνουν ακριβείς προβλέψεις για νέα, άγνωστα δεδομένα.

Οι μέθοδοι που αναφέρθηκαν, όπως η γραμμική παλινδρόμηση, το τυχαίο δάσος, το Extreme Gradient Boosting και τα βαθιά νευρωνικά δίκτυα, έχουν εφαρμογές και στα προβλήματα παλινδρόμησης, τα οποία αποσκοπούν στην πρόβλεψη μιας συνεχούς τιμής, όπως για παράδειγμα η πρόβλεψη του πιστωτικού κινδύνου ή της ζημίας λόγω αθέτησης. Κάθε μέθοδος προσεγγίζει το πρόβλημα από διαφορετική σκοπιά και έχει τα δικά της πλεονεκτήματα και περιορισμούς.

3.4.1 Γραμμική Παλινδρόμηση

Η γραμμική παλινδρόμηση [43] είναι μία από τις απλούστερες και πιο διαδεδομένες μεθόδους για προβλήματα παλινδρόμησης. Η βασική ιδέα πίσω από τη γραμμική παλινδρόμηση είναι ότι η σχέση μεταξύ των χαρακτηριστικών εισόδου και της εξόδου μπορεί να μοντελοποιηθεί μέσω μιας γραμμικής εξίσωσης:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

Όπου:

- y είναι η εξαρτημένη συνεχής μεταβλητή (π.χ. η ζημία λόγω αθέτησης).
- $x_1, x_2, \ldots, x_p$ είναι οι ανεξάρτητες μεταβλητές (χαρακτηριστικά).
- $\beta_0, \beta_1, \ldots, \beta_p$ είναι οι παράμετροι που πρέπει να εκτιμηθούν.
- ϵ είναι το σφάλμα της μοντέλου.

Η μέθοδος προσδιορίζει τις παραμέτρους μέσω της ελαχιστοποίησης του σφάλματος (συνήθως του τετραγώνου του σφάλματος). Παρά την απλότητά της, η γραμμική παλινδρόμηση μπορεί να έχει περιορισμένη ακρίβεια όταν η σχέση μεταξύ των χαρακτηριστικών και της εξαρτημένης μεταβλητής δεν είναι γραμμική.

3.4.2 Τυχαίο Δάσος για Παλινδρόμηση

Η μέθοδος τυχαίου δάσους (Random Forest) επεκτείνει την ιδέα των δέντρων απόφασης και χρησιμοποιείται σε προβλήματα παλινδρόμησης με εντυπωσιακά αποτελέσματα. Αντί να βασίζεται σε ένα μόνο δέντρο απόφασης, το τυχαίο δάσος δημιουργεί ένα σύνολο από δέντρα (φόρμα ψηφιακής συνδυαστικής μηχανής), όπου κάθε δέντρο εκπαιδεύεται σε τυχαία επιλεγμένο υποσύνολο των δεδομένων και με τυχαία υποσύνολα χαρακτηριστικών για κάθε διαίρεση του δέντρου.

Η έξοδος του τυχαίου δάσους για παλινδρόμηση υπολογίζεται ως ο μέσος όρος των εξόδων όλων των δέντρων:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} f_i(x)$$

όπου $f_i(x)$ είναι η πρόβλεψη του i-ου δέντρου για την είσοδο x, και N είναι ο αριθμός των δέντρων στο δάσος.

Τα τυχαία δάση είναι ιδιαίτερα ισχυρά όταν τα δεδομένα περιλαμβάνουν πολύπλοκες, μη γραμμικές σχέσεις. Προσφέρουν εξαιρετική αντοχή σε υπερπροσαρμογή (overfitting), επειδή συνδυάζουν πολλαπλές προβλέψεις από ανεξάρτητους δέντρους.

3.4.3 Extreme Gradient Boosting (XGBoost) για Παλινδρόμηση

Ο αλγόριθμος Extreme Gradient Boosting (XGBoost) είναι μία εξαιρετικά αποδοτική τεχνική για παλινδρόμηση, η οποία βασίζεται στην ενίσχυση (boosting) των δέντρων απόφασης. Στην ενίσχυση, τα δέντρα εκπαιδεύονται σειριακά, με κάθε νέο δέντρο να προσπαθεί να διορθώσει τα σφάλματα του προηγούμενου.

Η κεντρική φιλοσοφία πίσω από το XGBoost είναι η βελτιστοποίηση της συνάρτησης απώλειας μέσω του αλγορίθμου gradient descent. Η κύρια εξίσωση που καθορίζει τον τρόπο με τον οποίο το XGBoost εκπαιδεύεται είναι η εξίσωση της συνολικής συνάρτησης απώλειας:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

όπου:

- $l(y_i, \hat{y}_i)$ είναι η συνάρτηση απώλειας μεταξύ της πραγματικής τιμής y_i και της προβλεπόμενης $\hat{y}_i$
- $\Omega(f_k)$ είναι ο κανονικοποιητικός όρος για το δέντρο f_k, που ενισχύει την απλότητα του μοντέλου.
- θ είναι οι παράμετροι που πρέπει να βρεθούν.

Η προσαρμογή των βαρών για κάθε νέο δέντρο γίνεται με βάση το gradient boosting, το οποίο επικεντρώνεται στα σημεία που έχουν κάνει μεγαλύτερο σφάλμα.

3.4.4 Βαθιά Νευρωνικά Δίκτυα για Παλινδρόμηση

Τα βαθιά νευρωνικά δίκτυα μπορούν επίσης να χρησιμοποιηθούν για παλινδρόμηση, ιδίως όταν τα δεδομένα είναι πολύπλοκα και περιλαμβάνουν μη γραμμικές σχέσεις. Ένα νευρωνικό δίκτυο για παλινδρόμηση αποτελείται από πολλαπλά επίπεδα νευρώνων, με κάθε επίπεδο να αναλύει διαφορετικές πτυχές των δεδομένων.

Για το πρόβλημα της παλινδρόμησης, η έξοδος του δικτύου είναι μια συνεχής τιμή, η οποία υπολογίζεται με τη χρήση μιας συνάρτησης ενεργοποίησης κατάλληλης για συνεχή προβλέψεις, όπως η γραμμική συνάρτηση ενεργοποίησης. Στην εκπαίδευση, το δίκτυο προσπαθεί να ελαχιστοποιήσει τη συνάρτηση απώλειας (συνήθως το μέσο τετραγωνικό σφάλμα) για να βρει τα βέλτιστα βάρη που προσεγγίζουν καλύτερα την πραγματική τιμή.

Τα βαθιά νευρωνικά δίκτυα είναι ιδιαίτερα ισχυρά για πολύπλοκα προβλήματα, καθώς μπορούν να μάθουν πολύπλοκες σχέσεις μεταξύ των χαρακτηριστικών, και ενδείκνυνται για δεδομένα υψηλής διάστασης.

Συνολική Επισκόπηση

Η επιλογή της κατάλληλης μεθόδου για ένα πρόβλημα παλινδρόμησης εξαρτάται από τη φύση των δεδομένων και την πολυπλοκότητά τους. Η γραμμική παλινδρόμηση είναι ιδανική για απλές, γραμμικές σχέσεις, ενώ μέθοδοι όπως το τυχαίο δάσος, το XGBoost και τα βαθιά νευρωνικά δίκτυα είναι πιο ισχυρές και προτιμούνται σε περιπτώσεις όπου τα δεδομένα περιλαμβάνουν πολύπλοκες, μη γραμμικές σχέσεις και μεγάλες ποσότητες χαρακτηριστικών.

3.5 Μέθοδοι Εύρεσης Βέλτιστων Υπερπαραμέτρων

Η βελτιστοποίηση υπερπαραμέτρων ενός μοντέλου είναι η διαδικασία αναζήτησης για την καλύτερη συνδυαστική επιλογή των υπερπαραμέτρων που ελέγχουν τη συμπεριφορά και απόδοση ενός μοντέλου μηχανικής μάθησης. Οι υπερπαράμετροι είναι παράμετροι που καθορίζονται πριν από την εκπαίδευση του μοντέλου και δεν προκύπτουν άμεσα από τα δεδομένα. Αντίθετα, οι παράμετροι του μοντέλου (όπως τα βάρη σε ένα νευρωνικό δίκτυο) μαθαίνονται κατά τη διάρκεια της εκπαίδευσης. [44]

Σκοπός της Βελτιστοποίησης

Ο στόχος της βελτιστοποίησης των υπερπαραμέτρων είναι να βρει τη βέλτιστη ρύθμιση για να επιτευχθεί η καλύτερη απόδοση του μοντέλου, συνήθως σε σχέση με κάποιο μέτρο απόδοσης, όπως η ακρίβεια, η ευαισθησία, η F1-score ή το R^2 (ανάλογα με την περίπτωση ταξινόμησης ή παλινδρόμησης). [45]

3.5.1 Αναζήτηση με Χρήση Πλέγματος

Η μέθοδος Αναζήτησης με Χρήση Πλέγματος (Grid Search) εξαντλεί τον χώρο των υπερπαραμέτρων, δοκιμάζοντας όλους τους δυνατούς συνδυασμούς από τις καθορισμένες τιμές κάθε υπερπαραμέτρου. Για παράδειγμα, αν έχουμε δύο υπερπαραμέτρους με δύο πιθανές τιμές για την κάθε μία, θα δοκιμάσει 4 διαφορετικούς συνδυασμούς. [46]

Πλεονεκτήματα:

- Απλότητα και πλήρης κάλυψη: Είναι εύκολο στην εφαρμογή και εξασφαλίζει ότι όλες οι συνδυασμένες τιμές των υπερπαραμέτρων θα δοκιμαστούν, έτσι δεν υπάρχει κίνδυνος να παραληφθεί κάποια σημαντική τιμή.
- Ιδανικό όταν ο χώρος υπερπαραμέτρων είναι σχετικά μικρός και οι υπολογιστικοί πόροι το επιτρέπουν.

Μειονεκτήματα:

- Υψηλή υπολογιστική πολυπλοκότητα: Όταν οι υπερπαράμετροι είναι πολλές και οι πιθανές τιμές μεγάλες, ο αριθμός των συνδυασμών μπορεί να γίνει εκθετικός και να απαιτεί πολύ χρόνο και υπολογιστική ισχύ.

3.5.2 Τυχαία Αναζήτηση

Η μέθοδος Τυχαίας Αναζήτησης (Random Search) επιλέγει τυχαία παραμέτρους από τον χώρο υπερπαραμέτρων και εκτελεί τη διαδικασία εκπαίδευσης με αυτές τις τιμές. Αντί να εξετάζει όλους τους δυνατούς συνδυασμούς όπως το Grid Search, επιλέγει τυχαία συγκεκριμένο αριθμό συνδυασμών και μετράει την απόδοση κάθε συνδυασμού. [47]

Πλεονεκτήματα:

- Αποτελεσματικότητα στον χρόνο: Αν το διάστημα των υπερπαραμέτρων είναι μεγάλο και η αναζήτηση είναι πολύπλοκη, η random search μπορεί να βρει καλές παραμέτρους με λιγότερους υπολογισμούς από το Grid Search. [48]
- Συχνά βρίσκει εξίσου καλές ή και καλύτερες παραμέτρους από το Grid Search σε λιγότερο χρόνο.

Μειονεκτήματα:

- Μπορεί να μην καλύπτει όλο το χώρο: Αν και μπορεί να είναι αποτελεσματική, ενδέχεται να παραλείψει σημαντικούς συνδυασμούς αν οι επιλογές είναι πολύ τυχαίες.

3.5.3 Μπεϋζιανή Βελτιστοποίηση

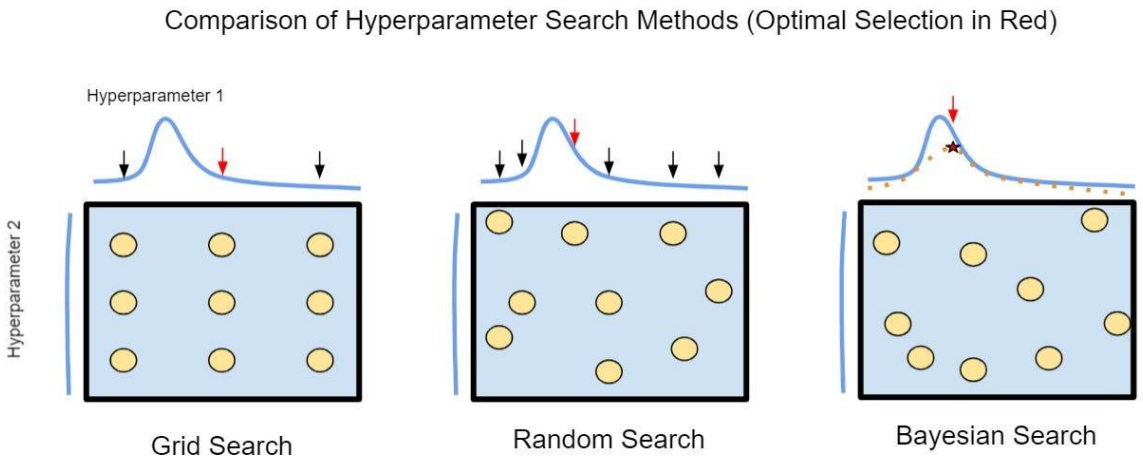
Η Μπεϋζιανή Βελτιστοποίηση (Bayesian Optimization) χρησιμοποιεί στατιστικά μοντέλα (συνήθως Γκαντικό μοντέλο ή Gaussian Process) για να προβλέψει ποιοι συνδυασμοί υπερπαραμέτρων είναι πιο πιθανό να οδηγήσουν σε καλή απόδοση. Χρησιμοποιεί τα αποτελέσματα από προηγούμενες αναζητήσεις για να "καθοδηγήσει" την αναζήτηση υπερπαραμέτρων, επιλέγοντας τις πιο υποσχόμενες παραμέτρους για δοκιμή. [49]

Πλεονεκτήματα:

- Αποτελεσματικότητα με λιγότερες δοκιμές: Αντί να δοκιμάζει όλους τους συνδυασμούς, επιλέγει εκείνους που έχουν τις περισσότερες πιθανότητες να βελτιώσουν την απόδοση του μοντέλου. [50]
- Μπορεί να μειώσει σημαντικά τον αριθμό των αναγκαστικών δοκιμών σε σχέση με άλλες μεθόδους, ιδιαίτερα σε προβλήματα με μεγάλο χώρο υπερπαραμέτρων.

Μειονεκτήματα:

- **Πολυπλοκότητα στην εφαρμογή**: Η εφαρμογή της Bayesian Optimization απαιτεί την κατανόηση των στατιστικών μοντέλων και μπορεί να είναι πιο περίπλοκη στην υλοποίηση.
- Μπορεί να απαιτεί μεγαλύτερη υπολογιστική ισχύ για τη διαχείριση του μοντέλου και την ενημέρωση της κατανομής πιθανοτήτων.



Εικόνα 8: Σύγκριση Μεθόδων Βελτιστοποίησης Υπερπαραμέτρων [61]

3.5.4 Γενετικοί Αλγόριθμοι

Οι γενετικοί αλγόριθμοι βασίζονται στην εξελικτική θεωρία και χρησιμοποιούν διαδικασίες όπως η αναπαραγωγή, η μετάλλαξη και η επιλογή για να εξελίξουν τον πληθυσμό των υπερπαραμέτρων. Ξεκινά με έναν "πληθυσμό" τυχαίων υπερπαραμέτρων και, μέσω διαδικασιών όπως η "διασταύρωση" και η "μετάλλαξη", παράγονται νέες παραμέτροι που αξιολογούνται για την απόδοσή τους. [51]

Πλεονεκτήματα:

- Δυνατότητα να ξεπεράσει τοπικά ελάχιστα: Η χρήση εξελικτικών διαδικασιών μπορεί να βοηθήσει στη διαφυγή από τοπικά ελάχιστα που μπορεί να προσπελαστεί με πιο παραδοσιακές μεθόδους.
- Πολύ χρήσιμοι σε προβλήματα με μεγάλους χώρους παραμέτρων και στην αναζήτηση μη γραμμικών σχέσεων μεταξύ των υπερπαραμέτρων.

Μειονεκτήματα:

- Υψηλή υπολογιστική πολυπλοκότητα: Η εκτέλεση των γενετικών αλγορίθμων μπορεί να είναι πολύ χρονοβόρα, ειδικά όταν ο πληθυσμός είναι μεγάλος και ο χώρος υπερπαραμέτρων είναι εκτεταμένος.
- Πολύπλοκο να ρυθμιστεί σωστά, απαιτώντας μεγάλη εμπειρία και υπολογιστικούς πόρους.

3.6 Διασταυρουμενη Επικύρωση k-fold

Η διασταυρούμενη επικύρωση k-fold [52] είναι μια δημοφιλής τεχνική αξιολόγησης της απόδοσης ενός μοντέλου μηχανικής μάθησης, που χρησιμοποιείται για να εκτιμήσει την ικανότητα γενικοποίησης ενός μοντέλου σε δεδομένα που δεν έχει ξαναδεί. Η βασική ιδέα πίσω από αυτήν την τεχνική είναι να χωρίσουμε το σύνολο των δεδομένων σε **k** υποσύνολα (γνωστά ως "folds") και στη συνέχεια να εκπαιδεύσουμε και να αξιολογήσουμε το μοντέλο πολλές φορές, με κάθε φορά διαφορετικό υποσύνολο ως δεδομένα αξιολόγησης (test set) και τα υπόλοιπα δεδομένα ως δεδομένα εκπαίδευσης (training set). Η διασταυρούμενη επικύρωση k-fold είναι μια εξαιρετική μέθοδος για την αξιολόγηση της απόδοσης ενός μοντέλου, ιδίως όταν έχουμε περιορισμένα δεδομένα. Παρά τα αυξημένα υπολογιστικά κόστη, τα πλεονεκτήματα της ακριβούς εκτίμησης της απόδοσης και της μείωσης του κινδύνου υπερεκπαίδευσης την καθιστούν μια από τις πιο αξιόπιστες τεχνικές αξιολόγησης στον τομέα της μηχανικής μάθησης.

Διαδικασία k-fold Cross-Validation:

1. Διαίρεση Δεδομένων: Το σύνολο των δεδομένων χωρίζεται σε k ίσα (ή σχεδόν ίσα) υποσύνολα, τα οποία ονομάζονται folds.
2. Εκπαίδευση και Αξιολόγηση: Ο αλγόριθμος εκπαιδεύεται **k φορές**, κάθε φορά χρησιμοποιώντας k-1 folds για την εκπαίδευση και το υπόλοιπο fold ως δεδομένα αξιολόγησης.
3. **Υπολογισμός Απόδοσης**: Κάθε φορά που το μοντέλο εκπαιδεύεται και αξιολογείται, υπολογίζεται μια μέτρηση απόδοσης (π.χ., ακρίβεια, F1-score, κλπ.). Οι μετρήσεις απόδοσης από τα k folds συνδυάζονται για να δώσουν μια συνολική εκτίμηση της απόδοσης του μοντέλου.

Πλεονεκτήματα:

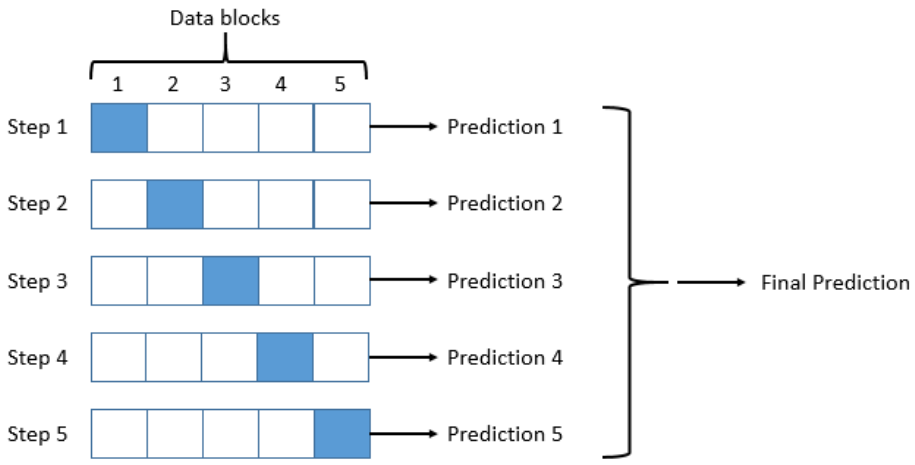
- Αξιολόγηση σε όλα τα δεδομένα: Κάθε δείγμα δεδομένων χρησιμοποιείται για εκπαίδευση και αξιολόγηση, κάτι που προσφέρει πιο αξιόπιστη εκτίμηση της απόδοσης του μοντέλου.
- Μείωση κινδύνου υπερεκπαίδευσης: Η διασταυρούμενη επικύρωση βοηθά στην αποφυγή της υπερεκπαίδευσης, καθώς το μοντέλο αξιολογείται σε διαφορετικά σύνολα δεδομένων και δεν εκπαιδεύεται μόνο σε ένα σταθερό σύνολο εκπαίδευσης.
- Στατιστικά αξιόπιστα αποτελέσματα: Η χρήση του μέσου όρου των αποτελεσμάτων από όλες τις δοκιμές προσφέρει πιο αξιόπιστα αποτελέσματα και μετρά τη μεταβλητότητα στην απόδοση του μοντέλου.

Μειονεκτήματα:

- Υπολογιστικό κόστος: Η διαδικασία απαιτεί περισσότερους υπολογιστικούς πόρους, καθώς ο αλγόριθμος πρέπει να εκπαιδευτεί και να αξιολογηθεί k φορές, κάτι που μπορεί να είναι χρονοβόρο, ειδικά για μεγάλα σύνολα δεδομένων ή σύνθετα μοντέλα.
- Ακατάλληλο για πολύ μικρά σύνολα δεδομένων: Αν τα δεδομένα είναι λίγα, η διασταυρούμενη επικύρωση μπορεί να μην είναι τόσο χρήσιμη, καθώς μπορεί να μην είναι αρκετά για να εκπαιδεύσουν αξιόπιστα το μοντέλο σε κάθε fold.

Κοινές Παραλλαγές:

- Stratified k-fold: Χρησιμοποιείται σε προβλήματα ταξινόμησης, διασφαλίζοντας ότι κάθε fold έχει αναλογική κατανομή των κλάσεων, έτσι ώστε το μοντέλο να εκπαιδεύεται και να αξιολογείται σε μια αντιπροσωπευτική κατανομή των δεδομένων.
- Leave-One-Out Cross-Validation (LOO-CV): Μια ειδική περίπτωση k-fold, όπου k ισούται με τον αριθμό των δειγμάτων του συνόλου δεδομένων, δηλαδή κάθε δείγμα χρησιμοποιείται μια φορά ως σύνολο αξιολόγησης, και το υπόλοιπο σύνολο ως δεδομένα εκπαίδευσης.



Εικόνα 9: Διασταυρούμενη Επικύρωση 5-fold [59]

Κεφάλαιο 4: Κατασκευή και Προεπεξεργασία Δεδομένων

4.1 Σύνολο Δεδομένων (Dataset)

Τα δεδομένα που χρησιμοποιήθηκαν είναι ένας συνδυασμός από datasets που βρέθηκαν στην πλατφόρμα του Kaggle καθώς και στην διπλωματική εργασία «Challenging LGD Models with Machine Learning». Το dataset μας περιλαμβάνει πάνω από 50.000 παραδείγματα καταναλωτικών δανείων που εκδόθηκαν από το 2014 έως το 2018. Αποτελείται από 43 χαρακτηριστικά που σχετίζονται με τους δανειολήπτες όσο και με τη συμπεριφορά πληρωμών τους.

Χαρακτηριστικά	**Περιγραφή**
AGE	Ηλικία Δανειολήπτη
DATE_OF_LOAN	Ημερομηνία Λήψης Δανείου
DATE_FROM	Ημερομηνία Αθέτησης Όρων Συμβολαίου από τον Δανειολήπτη
DATE_WO	Ημερομηνία Διαγραφής μέρους των Χρεών
DEFAULT_L12M	Αδυναμία πληρωμής δανειολήπτη πριν 12 μήνες
DEFAULT_L6M	Αδυναμία πληρωμής δανειολήπτη πριν 6 μήνες
DEFAULT_L3M	Αδυναμία πληρωμής δανειολήπτη πριν 3 μήνες
DELQ_L12M	Δείχνει αν ο δανειολήπτης ήταν εκπρόθεσμος τους τελευταίους 12 μήνες
DELQ_L6M	Δείχνει αν ο δανειολήπτης ήταν εκπρόθεσμος τους τελευταίους 6 μήνες
DELQ_L3M	Δείχνει αν ο δανειολήπτης ήταν εκπρόθεσμος τους τελευταίους 3 μήνες
loan_amnt	Το αναγραφόμενο ποσό που ζητήθηκε από τον δανειολήπτη
int_rate	Επιτόκιο του δανείου
installment	Η μηνιαία πληρωμή που οφείλει ο δανειολήπτης
grade	Τυποποιημένο μέτρο πιστοληπτικής ικανότητας, ανατίθεται από την λέσχη δανεισμού
sub_grade	Υποκατηγορία τυποποιημένου μέτρου πιστοληπτικής ικανότητας, ανατίθεται από την λέσχη δανεισμού
home_ownership	Το καθεστώς ιδιοκτησίας κατοικίας που παρέχεται από τον δανειολήπτη. Τιμές: RENT, OWN, MORTGAGE, OTHER
annual_inc	Ετήσιο εισόδημα
Dti	Ο λόγος του χρέους προς το εισόδημα του δανειολήπτη

postal_code	Ταχυδρομικός Κώδικας
revol_bal	Συνολικό ανακυκλώμενο πιστωτικό υπόλοιπο
revol_util	Ποσοστό χρήσης της ανακυκλωμένης γραμμής ή το ποσό της πίστωσης που χρησιμοποιεί ο δανειολήπτης σε σχέση με το σύνολο της διαθέσιμης ανακυκλωμένης πίστωσης.
total_acc	Ο συνολικός αριθμός των πιστωτικών γραμμών που βρίσκονται σήμερα στον πιστωτικό φάκελο του δανειολήπτη
out_prncp	Υπολειπόμενο ανεξόφλητο κεφάλαιο για το συνολικό χρηματοδοτούμενο ποσό
total_pymnt	Πληρωμές που έχουν ληφθεί μέχρι σήμερα για το συνολικό ποσό χρηματοδότησης
total_rec_int	Τόκοι που εισπράχθηκαν μέχρι σήμερα
total_rec_late_fee	Μέχρι σήμερα εισπραχθέντα τέλη καθυστέρησης
recoveries	Ακαθάριστη ανάκτηση μετά τη χρέωση
collection_recovery_fee	Αμοιβή είσπραξης μετά τη χρέωση
total_rev_hi_lim	Συνολική ανακυκλωμένη υψηλή πίστωση/όριο πίστωσης
arrears	Ποσό κατά το οποίο είναι ληξιπρόθεσμος ο δανειολήπτης
marital_status_code	Οικογενειακή Κατάσταση
writeoff_amnt	Διαγραφόμενο Ποσό
EAD	Έκθεση σε περίπτωση Αθέτησης Συμβολαίου
ITI	Λόγος Εισοδήματος προς Επιτόκιο
mortg_market_value	Αξία Στεγαστικού Δανείου
RESIDENTIAL	Δυαδική Τιμή για το αν το ενέχυρο είναι κατοικήσιμο
URBANISATION	Δυαδική Τιμή για το αν το ενέχυρο είναι σε αστική ζώνη
outdef_healthy	Δυαδική Τιμή για το αν ο δανειολήπτης είναι ξανά σε θέση να αποπληρώνει το δάνειό του.
Euribor_3	Δείκτης Euribor
LTV	Λόγος Δανείου προς Συνολική Αξία
NHG	Δυαδική Τιμή αν το δάνειο έχει ενέχυρο
Cure	Δυαδική Τιμή για την Ίαση του Δανείου
rec_rate	Ποσοστό Ανάκτησης

Πίνακας 2: Τα χαρακτηριστικά του Dataset και η Περιγραφή τους

4.2 Προεπεξεργασία των Δεδομένων

Προεπεξεργασία των δεδομένων είναι η διαδικασία μορφοποίησης των δεδομένων ακολουθώντας συγκεκριμένες ενέργειες, με σκοπό το μετασχηματισμό τους σε μορφή ικανή να δοθεί ως είσοδος στα προς μελέτη μοντέλα μηχανικής μάθησης.

Για την αποτελεσματική πρόβλεψη του δείκτη **LGD**, απαιτήθηκε η χρήση ενός καλά προεπεξεργασμένου και καθαρού συνόλου δεδομένων. Η διαδικασία προεπεξεργασίας περιλάμβανε διάφορα στάδια, με στόχο την αφαίρεση θορύβου, την κανονικοποίηση και την προετοιμασία των χαρακτηριστικών για τη χρήση σε αλγορίθμους μηχανικής μάθησης.

Η διαδικασία που ακολουθήθηκε περιλαμβάνει:

1. Φόρτωση και αρχικός έλεγχος του συνόλου δεδομένων
2. Χειρισμός ελλειπουσών τιμών
3. Ανίχνευση και αντιμετώπιση ακραίων τιμών
4. Διαχείριση Δεδομένων Τύπου Ημερομηνιών
5. Κανονικοποίηση και κλιμάκωση αριθμητικών μεταβλητών
6. Κωδικοποίηση κατηγορικών μεταβλητών
7. Αφαίρεση ισχυρά συσχετισμένων μεταβλητών
8. Διάσπαση των δεδομένων σε σύνολο εκπαίδευσης και δοκιμής

Κάθε ένα από αυτά τα στάδια αναλύεται στη συνέχεια με λεπτομέρεια.

1. Φόρτωση και Αρχικός Έλεγχος των Δεδομένων

Η διαδικασία ξεκίνησε με τη φόρτωση του συνόλου δεδομένων και τη διενέργεια ενός αρχικού ελέγχου για την κατανόηση της δομής και του περιεχομένου των μεταβλητών. Καταγράφηκαν βασικά στατιστικά στοιχεία, όπως το πλήθος των εγγραφών, ο τύπος των μεταβλητών (αριθμητικές ή κατηγορικές), καθώς και η ύπαρξη ελλειπουσών ή μη έγκυρων τιμών. Αναλύθηκε επίσης η κατανομή των τιμών στις αριθμητικές μεταβλητές, προκειμένου να εντοπιστούν πιθανά μοτίβα ή αποκλίσεις, ενώ πραγματοποιήθηκαν διαγράμματα διασποράς και ιστογράμματα για την καλύτερη κατανόηση της συμπεριφοράς των δεδομένων.

Αριθμητικές Μεταβλητές	**Κατηγορικές Μεταβλητές**
AGE	DATE_OF_LOAN
DEFAULT_L12M	DATE_FROM
DEFAULT_L6M	DATE_WO
DEFAULT_L3M	grade
DELQ_L12M	sub_grade
DELQ_L6M	home_ownership
DELQ_L3M	postal_code
loan_amnt	marital_status_code
int_rate	
installment	
annual_inc	
dti	
revol_bal	
revol_util	
total_acc	
out_prncp	
total_pymnt	
total_rec_int	
total_rec_late_fee	
recoveries	
collection_recovery_fee	
total_rev_hi_lim	
arrears	
writeoff_amnt	
EAD	
ITI	
mortg_market_value	
RESIDENTIAL	
URBANISATION	
outdef_healthy	
Euribor_3	
LTV	
NHG	
cure	
rec_rate	

Πίνακας 3: Αριθμητικές και Κατηγορικές Μεταβλητές

2. **Χειρισμός Ελλειπουσών Τιμών**

Κατά την ανάλυση των δεδομένων, διαπιστώθηκε ότι ορισμένες μεταβλητές περιείχαν ελλείπουσες τιμές. Ο χειρισμός αυτών των τιμών εξαρτήθηκε από τον αριθμό και τη φύση τους.

- Στις περιπτώσεις όπου οι ελλείπουσες τιμές αποτελούσαν μικρό ποσοστό των δεδομένων (κάτω από 5%), αντικαταστάθηκαν με την διάμεσο της αντίστοιχης μεταβλητής.
- Για τις κατηγορικές μεταβλητές, οι ελλείπουσες τιμές αντικαταστάθηκαν είτε με την πιο συχνά εμφανιζόμενη τιμή (modus) είτε με μια νέα κατηγορία "Άγνωστο", ώστε να διατηρηθούν όλα τα δείγματα στο σύνολο δεδομένων.

3. **Ανίχνευση και Αντιμετώπιση Ακραίων Τιμών**

Οι ακραίες τιμές είναι πιθανό να επηρεάσουν σημαντικά την απόδοση των μοντέλων μηχανικής μάθησης, καθώς μπορούν να προκαλέσουν στρέβλωση στη διαδικασία εκπαίδευσης.

Για τον εντοπισμό των ακραίων τιμών, χρησιμοποιήθηκε η μέθοδος του διαμεσολαδικού εύρους (Interquartile Range - IQR). Οι παρατηρήσεις που βρίσκονταν εκτός του καθορισμένου αποδεκτού εύρους αφαιρέθηκαν ή περιορίστηκαν εντός ανεκτών ορίων.

Η ανάλυση ακραίων τιμών πραγματοποιήθηκε με τη χρήση διαγραμμάτων boxplot και ιστογραμμάτων, επιτρέποντας την αναγνώριση τιμών που υπερέβαιναν τα φυσιολογικά όρια.

4. **Διαχείριση Δεδομένων Τύπου Ημερομηνιών**

Για τα χαρακτηριστικά των οποίων ο τύπος ήταν Ημερομηνία χρειάστηκε ειδική μεταχείριση. Στα πλαίσια της εργασίας αυτής, τα χαρακτηριστικά αυτά αντικαταστάθηκαν με 6 καινούργια, τα οποία και κατασκευάστηκαν. Τα χαρακτηριστικά που προέκυψαν αποτελούν τις διαφορές των ημερομηνιών σε ημέρες μεταξύ των τριών αυτών χαρακτηριστικών και τους μήνες κατά τους οποίους πραγματοποιήθηκε το κάθε γεγονός.

Τελικά Χαρακτηριστικά	Περιγραφή
loan_to_default_days	DATE_FROM - DATE_OF_LOAN
loan_to_wo_days	DATE_WO - DATE_OF_LOAN
default_to_wo_days	DATE_WO - DATE_FROM
loan_month	Μήνας του DATE_OF_LOAN
default_month	Μήνας του DATE_FROM
wo_month	Μήνας του DATE_WO

Πίνακας 4: Τελικές Κατασκευασμένες Μεταβλητές

Η μέθοδος αυτή χρησιμοποιείται συχνά σε χαρακτηριστικά τύπου Datetime για οικονομικά δεδομένα. Με αυτό τον τρόπο χρησιμοποιείται η πληροφορία του πόσος καιρός έχει περάσει από την λήψη του δανείου μέχρι την αθέτηση των όρων του και τη διαγραφή μέρους του Χρέους.

5. Εξαγωγή Δεδομένων από ταχυδρομικό κώδικα

Το χαρακτηριστικό του ταχυδρομικού κώδικα απομακρύνθηκε, ενώ δημιουργήθηκαν δύο νέα χαρακτηριστικά, εκείνο της χώρας και της πόλης. Περαιτέρω επεξεργασία σε αυτά γίνεται παρακάτω.

6. Κωδικοποίηση Κατηγορικών Μεταβλητών

Οι αλγόριθμοι μηχανικής μάθησης δεν μπορούν να διαχειριστούν απευθείας κατηγορικές μεταβλητές, επομένως ήταν απαραίτητο να μετατραπούν σε αριθμητικές μορφές.

- Για τις μεταβλητές όπου δεν υπάρχει ιεραρχία μεταξύ των τιμών, χρησιμοποιήθηκε η μέθοδος του One-Hot Encoding, η οποία δημιουργεί ξεχωριστές δυαδικές στήλες για κάθε κατηγορία. Τα χαρακτηριστικά στα οποία χρησιμοποιήθηκε η μέθοδος είναι τα παρακάτω: ‘marital_status_code’, ‘city’, ‘country’.
- Για κατηγορικές μεταβλητές οι οποίες είναι διατεταγμένες ή στις οποίες υπάρχει ψευδο-ιεραρχία, χρησιμοποιήθηκε η μέθοδος της κωδικοποίησης ετικετών (Label Encoding), όπου κάθε κατηγορία αντιστοιχήθηκε σε έναν ακέραιο αριθμό. Τα χαρακτηριστικά στα οποία χρησιμοποιήθηκε η μέθοδος είναι τα παρακάτω: ‘grade’, ‘subgrade’, ‘home_ownership’.

Η επιλογή της κατάλληλης μεθόδου κωδικοποίησης για κάθε περίπτωση βασίστηκε στη διατήρηση της πληροφορίας των δεδομένων και στη μείωση της πολυπλοκότητας του μοντέλου.

7. Κανονικοποίηση και Κλιμάκωση Δεδομένων

Δεδομένου ότι οι αριθμητικές μεταβλητές του dataset είχαν διαφορετικές κλίμακες, εφαρμόστηκε η τεχνική της κανονικοποίησης μέσω min-max scaling. Με αυτόν τον τρόπο, όλες οι τιμές μετατράπηκαν σε ένα εύρος από 0 έως 1, διατηρώντας την κατανομή των δεδομένων και επιτρέποντας στα μοντέλα μηχανικής μάθησης να εκπαιδευτούν πιο αποτελεσματικά. Τα χαρακτηριστικά στα οποία εφαρμόστηκε η παραπάνω μέθοδος είναι τα εξής:

'AGE', 'DEFAULT_L12M', 'DEFAULT_L6M', 'DEFAULT_L3M', 'DELQ_L12M', 'DELQ_L6M', 'DELQ_L3M', 'loan_amnt', 'int_rate', 'installment', 'annual_inc', 'dti', 'revol_bal', 'revol_util', 'total_acc', 'total_pymnt', 'total_rec_int', 'total_rec_late_fee', 'total_rev_hi_lim', 'recoveries', 'arrears', 'writeoff_amnt', 'EAD', 'ITI', 'mortg_market_value'

Η διαδικασία αυτή κρίθηκε απαραίτητη για να αποφευχθεί η ανεπιθύμητη επίδραση μεταβλητών με μεγάλες αριθμητικές τιμές στις υπολογιστικές διαδικασίες των αλγορίθμων.

8. Αφαίρεση Ισχυρά Συσχετισμένων Μεταβλητών

Η ύπαρξη μεταβλητών με υψηλή συσχέτιση μπορεί να δημιουργήσει προβλήματα πολυπλοκότητας και υπερεκπαίδευσης στα μοντέλα. Για τον λόγο αυτό, κατασκευάστηκε ένας πίνακας συσχέτισης, μέσω του οποίου αφαιρέθηκαν μεταβλητές που παρουσίαζαν συσχέτιση άνω του 80% με άλλες μεταβλητές.

Με αυτόν τον τρόπο, διατηρήθηκαν μόνο οι μεταβλητές που προσέφεραν μοναδική πληροφορία στη διαδικασία της πρόβλεψης. Τα αποτελέσματα παρουσιάζονται παρακάτω:

Χαρακτηριστικό 1	**Χαρακτηριστικό 2**	**Συσχέτιση**
loan_amnt	installment	0.9437023858242986
loan_amnt	out_prncp	0.8740495136163786
installment	out_prncp	0.8115874112934265
total_pymnt	total_rec_prncp	0.8459377112049535
recoveries	collection_recovery_fee	0.8260039430326294
revol_bal	total_rev_hi_lim	0.8244902781548844
LTV	cure	0.8429708737063563
revol_util	rec_rate	0.8257317736667380
int_rate	grade	0.9737166344345559
int_rate	sub_grade	0.9974057002522476
grade	sub_grade	0.9754844135881052

Πίνακας 5: Πίνακας Συσχέτισης

Με βάση την παραπάνω ανάλυση, τα χαρακτηριστικά που αφαιρέθηκαν είναι τα παρακάτω: ‘out_prncp’, ‘total_rec_prncp’, ‘collection_recovery_fee’, ‘total_rev_hi_lim’, ‘grade’, ‘sub_grade’.

9. Τελική Μορφή Χαρακτηριστικών Δεδομένων

Το τελικό σύνολο δεδομένων περιλαμβάνει 62 διαφορετικά χαρακτηριστικά (στήλες).

AGE	EAD	country_Belgium
DEFAULT_L12M	ITI	country_France
DEFAULT_L6M	mortg_market_value	country_Germany
DEFAULT_L3M	RESIDENTIAL	country_Netherlands
DELQ_L12M	URBANISATION	city_Amsterdam
DELQ_L6M	outdef_healthy	city_Antwerp
DELQ_L3M	Euribor_3	city_Berlin
loan_amnt	LTV	city_Brussels
int_rate	NHG	city_Charleroi
installment	cure	city_Cologne
annual_inc	rec_rate	city_Ghent
dti	loan_month	city_Hamburg
revol_bal	default_month	city_Lyon
revol_util	wo_month	city_Marseille
total_acc	loan_to_default_days	city_Munich
total_pymnt	loan_to_wo_days	city_Paris
total_rec_int	default_to_wo_days	city_Rotterdam
total_rec_late_fee	marital_status_code_Divorced	city_The Hague
recoveries	marital_status_code_Married	city_Toulouse
arrears	marital_status_code_Unmarried	city_Utrecht
writeoff_amnt	home_ownership	

Πίνακας 6: Τελικά Χαρακτηριστικά Συνόλου

10. Διάσπαση Δεδομένων σε Σύνολα Εκπαίδευσης και Δοκιμής

Το τελικό dataset χωρίστηκε σε σύνολο εκπαίδευσης (80%) και σύνολο δοκιμών (20%). Η τυχαία επιλογή των δεδομένων διασφάλισε ότι το μοντέλο θα εκπαιδευόταν σε ένα αντιπροσωπευτικό υποσύνολο του dataset και θα δοκιμαζόταν με άγνωστα δεδομένα για την αξιολόγηση της απόδοσής του.

Η ολοκληρωμένη διαδικασία προεπεξεργασίας διαμόρφωσε ένα καθαρό, ομοιογενές και έτοιμο για ανάλυση σύνολο δεδομένων, κατάλληλο για την εφαρμογή των μεθόδων μηχανικής μάθησης.

Κεφάλαιο 5: Μεθοδολογία

5.1 Προετοιμασία Δεδομένων και Διαχωρισμός

Το σύνολο δεδομένων που χρησιμοποιείται περιλαμβάνει διάφορες μεταβλητές, ενώ οι εξαρτημένες μεταβλητές είναι το cure και το rec_rate. Η κάθε μια από τις εξαρτημένες μεταβλητές αυτές, χρησιμοποιείται σε διαφορετικά στάδια της υλοποίησης, η cure στο στάδιο της κατηγοριοποίησης και η rec_rate στο στάδιο της παλινδρόμησης. Τέλος, για την εκπαίδευση των μοντέλων, τα δεδομένα χωρίζονται σε εκπαίδευση (train), επικύρωση (validation) και δοκιμή (test), με ποσοστά 70%-10%-20%.

5.2. Περιγραφή των Μοντέλων Κατηγοριοποίησης

5.2.1 Λογιστική Παλινδρόμηση

Η Λογιστική Παλινδρόμηση είναι μια γραμμική μέθοδος ταξινόμησης που χρησιμοποιεί τη συνάρτηση sigmoid για την πρόβλεψη της πιθανότητας ότι ένα δείγμα ανήκει σε μια συγκεκριμένη κατηγορία. Στην εκπαίδευση του μοντέλου χρησιμοποιείται ο solver "liblinear", καθώς ο "lbfgs" παρουσίαζε προβλήματα σύγκλισης. Επιπλέον, ο αριθμός των μέγιστων επαναλήψεων της διαδικασίας βελτιστοποίησης ορίζεται στις 200, ώστε να διασφαλιστεί ότι το μοντέλο συγκλίνει σε μια σταθερή λύση. Η απόδοση του μοντέλου αξιολογείται μέσω διασταυρούμενης επικύρωσης 5-fold, όπου υπολογίζεται η μέση ακρίβεια.

5.2.2 Τυχαίο Δάσος (Random Forest)

Ο αλγόριθμος Τυχαίο Δάσος (Random Forest) αποτελεί έναν από τους πιο διαδεδομένους ταξινομητές μηχανικής μάθησης, ο οποίος βασίζεται στη χρήση πολλαπλών δέντρων απόφασης (decision trees) για τη βελτίωση της ακρίβειας και της γενίκευσης του μοντέλου. Η διαδικασία εκπαίδευσης του μοντέλου ξεκινά με την αρχικοποίηση ενός αντικειμένου RandomForestClassifier και τον ορισμό των βασικών υπερπαραμέτρων του. Για τη βελτιστοποίηση των υπερπαραμέτρων χρησιμοποιείται η μέθοδος Grid Search με 5-Fold Cross-Validation, ώστε να επιλεγούν οι τιμές που οδηγούν στη βέλτιστη απόδοση του μοντέλου.

Συγκεκριμένα, το πλέγμα υπερπαραμέτρων (param_grid) περιλαμβάνει:

- τον αριθμό των δέντρων του δάσους (n_estimators) [50, 100, 200, 400],
- το μέγιστο βάθος των δέντρων (max_depth) [None, 10, 20, 30],
- Μέγιστος αριθμός χαρακτηριστικών που χρησιμοποιούνται για κάθε διαχωρισμό (max_features) [5, 10, sqrt, log2]. Το sqrt και το log2 επιλέγουν την τετραγωνική ρίζα του συνολικού αριθμού χαρακτηριστικών που επιλέγονται για το μοντέλο,
- τον ελάχιστο αριθμό δειγμάτων σε ένα φύλλο (min_samples_leaf) [1, 2, 4]

Η αναζήτηση των βέλτιστων τιμών αυτών των παραμέτρων πραγματοποιείται μέσω της κλάσης GridSearchCV, η οποία εφαρμόζει διασταυρούμενη επικύρωση (cross-validation) με πέντε υποσύνολα (5-folds) και χρησιμοποιεί την ακρίβεια (accuracy) ως κριτήριο αξιολόγησης. Συνολικά, μέσω του Grid Search εξετάζονται 192 (4x4x4x3) πιθανά μοντέλα, από 5 φορές το καθένα, λόγω του cross-validation πραγματοποιώντας, έτσι, 960 υπολογισμούς. Το τελικό μοντέλο με την καλύτερη απόδοση έχει τα εξής χαρακτηριστικά:

max_depth	None
max_features	10
min_samples_leaf	1
n_estimators	200

Πίνακας 7: Υπερπαράμετροι τελικού μοντέλου Τυχαίου Δάσους

5.2.3 XGBoost

Υλοποιείται ένας XGBoost Classifier χρησιμοποιώντας Grid SearchCV για τη βελτιστοποίηση των υπερπαραμέτρων. Για την αρχικοποίηση του ταξινομητή, χρησιμοποιείται η παράμετρος booster='gbtree', η οποία καθορίζει ότι το XGBoost θα χρησιμοποιήσει δέντρα απόφασης για την ταξινόμηση. Επιπλέον, η επιλογή της συνάρτησης κόστους eval_metric='mlogloss' στοχεύει στη βελτιστοποίηση της πολυταξικής λογιστικής απώλειας (Multiclass Logarithmic Loss), η οποία είναι ιδανική για προβλήματα κατηγοριοποίησης. Για να διασφαλιστεί η επαναληψιμότητα των αποτελεσμάτων, ορίζεται το random_state=42.

Για τη βελτιστοποίηση της απόδοσης του μοντέλου, χρησιμοποιείται ένα πλέγμα υπερπαραμέτρων (param_grid), το οποίο περιλαμβάνει:

- Το max_depth των δέντρων ρυθμίζεται στις τιμές [3, 4, 5, 7], ελέγχοντας πόσο βαθιά μπορούν να αναπτυχθούν τα δέντρα. Οι μικρότερες τιμές βοηθούν στην αποφυγή του overfitting, ενώ οι μεγαλύτερες επιτρέπουν στο μοντέλο να μάθει πιο σύνθετα μοτίβα.
- Το subsample, με τιμές [0.7, 0.8, 0.9, 1.0], καθορίζει το ποσοστό των δεδομένων που χρησιμοποιούνται για την εκπαίδευση κάθε δέντρου, με χαμηλότερες τιμές να βοηθούν στην αποφυγή υπερπροσαρμογής.
- Το min_child_weight ([1, 3, 5]) ελέγχει το ελάχιστο άθροισμα βαρών των δειγμάτων που απαιτείται για τη δημιουργία ενός νέου κόμβου, με υψηλότερες τιμές να οδηγούν σε πιο σταθερά μοντέλα.
- Ο ρυθμός εκμάθησης (learning_rate), ο οποίος ορίζεται στις τιμές [0.001, 0.01, 0.1, 0.2], καθορίζει πόσο γρήγορα το μοντέλο προσαρμόζεται στα δεδομένα. Χαμηλές τιμές, όπως 0.001 και 0.01, απαιτούν περισσότερα δέντρα αλλά μειώνουν τον κίνδυνο overfitting, ενώ υψηλότερες τιμές επιταχύνουν τη σύγκλιση αλλά μπορεί να οδηγήσουν σε υπερπροσαρμογή.
- Ο αριθμός των δέντρων (n_estimators) καθορίζεται στις τιμές [50, 100, 200, 400], με μεγαλύτερο αριθμό δέντρων να οδηγεί σε αυξημένη ακρίβεια αλλά και μεγαλύτερο υπολογιστικό κόστος.

Η διαδικασία Grid SearchCV εφαρμόζεται για να δοκιμάσει όλους τους πιθανούς συνδυασμούς αυτών των παραμέτρων, χρησιμοποιώντας 5-Fold Cross-Validation (cv=5). Η ακρίβεια (accuracy) επιλέγεται ως μετρική αξιολόγησης. Συνολικά, μέσω του Grid Search εξετάζονται 768 (4x4x4x4x3) πιθανά μοντέλα, από 5 φορές το καθένα, λόγω του cross-validation πραγματοποιώντας, έτσι, 3840 υπολογισμούς. Το τελικό μοντέλο με την καλύτερη απόδοση έχει τα εξής χαρακτηριστικά:

learning_rate	0.1
max_depth	7
min_child_weight	1
n_estimators	100
subsample	0.7

Πίνακας 8: Υπερπαράμετροι τελικού μοντέλου XGBoost

5.2.4 Βαθύ Νευρωνικό Δίκτυο (Neural Network)

Υλοποιείται η βελτιστοποίηση υπερπαραμέτρων ενός νευρωνικού δικτύου μέσω Bayesian Optimization, με σκοπό την εύρεση των βέλτιστων ρυθμίσεων για τον ρυθμό εκμάθησης (learning_rate), τον αριθμό νευρώνων (neurons) και το ποσοστό dropout (dropout). Η διαδικασία περιλαμβάνει δύο βασικά στάδια: τη δημιουργία μιας συνάρτησης αξιολόγησης του μοντέλου και την εκτέλεση της Bayesian Optimization.

Αρχικά, ορίζεται η συνάρτηση αξιολόγησης nn_evaluate, η οποία κατασκευάζει και εκπαιδεύει ένα feedforward νευρωνικό δίκτυο για δυαδική ταξινόμηση. Το μοντέλο αποτελείται από μία είσοδο με πλήρως συνδεδεμένους νευρώνες (Dense layer), ακολουθούμενη από Dropout layer, το οποίο μειώνει τον κίνδυνο overfitting, απενεργοποιώντας τυχαία ορισμένους νευρώνες κατά την εκπαίδευση. Έπειτα, υπάρχει μια δεύτερη πλήρως συνδεδεμένη κρυφή στρώση, η οποία περιέχει τους μισούς νευρώνες της πρώτης. Η έξοδος του δικτύου περιλαμβάνει έναν νευρώνα με ενεργοποίηση sigmoid, ο οποίος μετατρέπει την έξοδο του δικτύου σε τιμές μεταξύ 0 και 1. Το μοντέλο συμπιέζεται χρησιμοποιώντας τον βελτιστοποιητή Adam, με ρυθμό εκμάθησης (learning_rate) που καθορίζεται δυναμικά από την Bayesian Optimization. Ως συνάρτηση κόστους, χρησιμοποιείται η binary_crossentropy, κατάλληλη για δυαδική ταξινόμηση, και ως μετρική αξιολόγησης χρησιμοποιείται η ακρίβεια (accuracy). Το μοντέλο εκπαιδεύεται για 50 εποχές, με batch size 32. Επιπλέον, εφαρμόζεται ένας Early Stopping μηχανισμός, ο οποίος διακόπτει την εκπαίδευση εάν δεν υπάρχει βελτίωση μετά από 5 συνεχόμενες εποχές, διασφαλίζοντας τη διατήρηση των καλύτερων βαρών του μοντέλου.

Στη συνέχεια, γίνεται η βελτιστοποίηση των υπερπαραμέτρων μέσω της Bayesian Optimization. Καθορίζεται ένα σύνορο αναζήτησης (pbounds) για τις υπερπαραμέτρους, όπου:

- Ο ρυθμός εκμάθησης (learning_rate) κυμαίνεται μεταξύ 0.0001 και 0.01, επιτρέποντας στο μοντέλο να μάθει είτε πολύ αργά είτε με μεγαλύτερα βήματα.
- Ο αριθμός νευρώνων (neurons) κυμαίνεται μεταξύ 16 και 128, καθορίζοντας τον αριθμό των μονάδων που θα χρησιμοποιηθούν στις κρυφές στρώσεις.
- Το ποσοστό dropout ρυθμίζεται μεταξύ 0.1 και 0.5, επιτρέποντας τη δοκιμή διαφορετικών επιπέδων τυχαίας απενεργοποίησης νευρώνων.

Η Bayesian Optimization εκτελείται με init_points=10 (δηλαδή αρχικοποιεί 10 τυχαίες τιμές για να δημιουργήσει το μοντέλο) και n_iter=20, που σημαίνει ότι θα βελτιστοποιήσει τις υπερπαραμέτρους σε 20 επιπλέον επαναλήψεις. Η μέθοδος χρησιμοποιεί ως αντικειμενική συνάρτηση τη nn_evaluate, όπου αξιολογεί διαφορετικούς συνδυασμούς παραμέτρων και επιλέγει εκείνον που μεγιστοποιεί την ακρίβεια. Στο τέλος της διαδικασίας, εμφανίζονται οι βέλτιστες υπερπαράμετροι (optimizer.max['params']), οι οποίες στη συνέχεια χρησιμοποιούνται για την κατασκευή του τελικού μοντέλου.

Το τελικό μοντέλο κατασκευάζεται χρησιμοποιώντας τις βέλτιστες τιμές των υπερπαραμέτρων που προέκυψαν από την Bayesian Optimization. Το μοντέλο κατασκευάζεται με τις εξής υπερπαραμέτρους:

learning_rate	0.008625
neurons	27
dropout rate	0.2233

Πίνακας 9: Υπερπαράμετροι Τελικού Μοντέλου Νευρωνικού Δικτύου

5.3 Περιγραφή των Μοντέλων Παλινδρόμησης

5.3.1 Γραμμική Παλινδρόμηση

Η Γραμμική Παλινδρόμηση είναι η βασική μέθοδος παλινδρόμησης που αναζητά μια γραμμική σχέση μεταξύ των χαρακτηριστικών και της μεταβλητής-στόχου. Η εκτίμηση της εξαρτημένης μεταβλητής y γίνεται μέσω της εξίσωσης:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

όπου β_0 είναι η σταθερά (intercept) και β_i οι συντελεστές των χαρακτηριστικών.

Μετρώνται οι εξής δείκτες απόδοσης:

- Mean Squared Error (MSE): Εκτιμά το μέσο τετράγωνο των σφαλμάτων πρόβλεψης.
- Mean Absolute Error (MAE): Υπολογίζει την απόλυτη διαφορά μεταξύ των πραγματικών και προβλεπόμενων τιμών.
- R-squared (R^2): Μετρά πόσο καλά το μοντέλο εξηγεί τη διακύμανση της εξαρτημένης μεταβλητής.

Επιπλέον, παρουσιάζεται η γραφική απεικόνιση των υπολειμμάτων (residuals) και των πραγματικών vs προβλεπόμενων τιμών.

5.3.2 Τυχαίο Δάσος (Random Forest)

Υλοποιείται η βελτιστοποίηση υπερπαραμέτρων για έναν Random Forest Regressor χρησιμοποιώντας τη μέθοδο Grid Search. Το Random Forest Regressor είναι ένας ensemble αλγόριθμος παλινδρόμησης, ο οποίος αποτελείται από πολλαπλά Δέντρα Απόφασης, συνδυάζοντας τα αποτελέσματά τους για να κάνει πιο ακριβείς προβλέψεις. Ο στόχος της βελτιστοποίησης είναι να βρεθούν οι καλύτερες δυνατές υπερπαράμετροι, ώστε το μοντέλο να αποδίδει με τη μέγιστη ακρίβεια.

Αρχικά, ορίζεται το πλέγμα υπερπαραμέτρων (param_grid), το οποίο περιέχει διάφορες ρυθμίσεις που το Grid Search θα εξετάσει:

- n_estimators: Ο αριθμός των δέντρων στο δάσος, με πιθανά μεγέθη [50, 100, 200, 400].
- max_features: Ο αριθμός των χαρακτηριστικών που θα χρησιμοποιεί κάθε δέντρο για τον διαχωρισμό. Οι πιθανές τιμές είναι [5, 10, 'sqrt', 'log2'], όπου 'sqrt' και 'log2' σημαίνουν ότι το μοντέλο επιλέγει τετραγωνική ρίζα ή λογάριθμο του συνολικού αριθμού χαρακτηριστικών.
- max_depth: Το μέγιστο επιτρεπτό βάθος των δέντρων, με τιμές [None, 10, 20, 30]. Όταν είναι None, το δέντρο αναπτύσσεται μέχρι να μην μπορεί να γίνει άλλος διαχωρισμός.
- min_samples_leaf: Ο ελάχιστος αριθμός δειγμάτων που απαιτούνται σε κάθε φύλλο του δέντρου ([1, 2, 4]).
- bootstrap: Καθορίζει εάν το μοντέλο θα χρησιμοποιήσει δειγματοληψία με αντικατάσταση (True) ή χωρίς αντικατάσταση (False). Όταν είναι True, κάθε δέντρο λαμβάνει τυχαίο υποσύνολο των δεδομένων, κάτι που βοηθά στη μείωση του overfitting.

Αφού ο αλγόριθμος ολοκληρώσει τη διαδικασία της διασταυρούμενης επικύρωσης (Cross-Validation) και συγκρίνει όλους τους πιθανούς συνδυασμούς υπερπαραμέτρων, το μοντέλο εκπαιδεύεται με τα καλύτερα δυνατά χαρακτηριστικά. Δοκιμάζονται 384 (4x4x4x3x3) διαφορετικά μοντέλα, το καθένα από τα οποία δοκιμάζεται από 5 φορές, λόγω της διασταυρομένης επικύρωσης, δίνοντας 1920 διαφορετικά fits. Το τελικό μοντέλο έχει τις εξής τιμές για υπερπαραμέτρους:

n_estimators	400
max_features	10
max_depth	20
min_samples_leaf	1
bootstrap	False

Πίνακας 10: Υπερπαράμετροι Τελικού Μοντέλου Τυχαίου Δάσους Παλινδρόμησης

5.3.3 XGBoost

Υλοποιείται η βελτιστοποίηση υπερπαραμέτρων για έναν XGBoost Regressor χρησιμοποιώντας τη μέθοδο Grid Search. Το XGBoost είναι ένας από τους πιο ισχυρούς αλγορίθμους gradient boosting, ο οποίος βελτιώνει τη διαδικασία μάθησης, χρησιμοποιώντας διαδοχικά δέντρα απόφασης και διορθώνοντας τα σφάλματα των προηγούμενων επαναλήψεων. Ο κύριος στόχος του συγκεκριμένου κώδικα είναι η εύρεση του βέλτιστου συνδυασμού υπερπαραμέτρων, ώστε το μοντέλο να έχει τη χαμηλότερη μέση τετραγωνική απόκλιση (Mean Squared Error - MSE), η οποία χρησιμοποιείται ως κριτήριο αξιολόγησης.

Αρχικά, ορίζεται το πλέγμα υπερπαραμέτρων (param_grid), το οποίο περιλαμβάνει διαφορετικές τιμές που θα εξεταστούν για να βρεθεί η καλύτερη ρύθμιση του XGBoost μοντέλου:

- n_estimators: [50, 100, 200, 400]. Καθορίζει τον αριθμό των δέντρων απόφασης που θα χρησιμοποιηθούν στο μοντέλο. Περισσότερα δέντρα αυξάνουν τη δυνατότητα μάθησης, αλλά μπορεί να οδηγήσουν σε μεγαλύτερο υπολογιστικό κόστος και πιθανό overfitting.
- max_depth: [1, 3, 5, 7]. Ρυθμίζει το μέγιστο βάθος κάθε δέντρου. Τα χαμηλότερα βάθη (1, 3) μειώνουν την πιθανότητα overfitting, ενώ τα μεγαλύτερα επιτρέπουν στο μοντέλο να μάθει πιο σύνθετα μοτίβα.
- learning_rate: [0.01, 0.05, 0.1]. Ορίζει πόσο γρήγορα το μοντέλο προσαρμόζει τα βάρη του. Χαμηλές τιμές (0.01, 0.05) οδηγούν σε πιο σταθερή εκμάθηση, αλλά απαιτούν μεγαλύτερο αριθμό δέντρων.
- subsample: [0.7, 0.9]. Ορίζει το ποσοστό των δεδομένων εκπαίδευσης που χρησιμοποιούνται σε κάθε δέντρο. Χαμηλές τιμές (0.7) εισάγουν περισσότερη τυχαιότητα, μειώνοντας την πιθανότητα overfitting.
- colsample_bytree: [0.7, 1.0]. Ελέγχει το ποσοστό των χαρακτηριστικών που χρησιμοποιούνται σε κάθε δέντρο. Αν είναι 0.7, μόνο το 70% των χαρακτηριστικών θα χρησιμοποιείται κάθε φορά. Αν είναι 1.0, κάθε δέντρο χρησιμοποιεί όλα τα διαθέσιμα χαρακτηριστικά.
- min_child_weight: [1, 3]. Καθορίζει το ελάχιστο άθροισμα των βαρών των δειγμάτων που απαιτούνται για να δημιουργηθεί ένας νέος κόμβος στο δέντρο. Μεγαλύτερες τιμές (3) σημαίνουν ότι το μοντέλο θα δημιουργεί πιο ισορροπημένα δέντρα, αποφεύγοντας πολύπλοκους διαχωρισμούς.

Για την εύρεση του τελικού μοντέλου δοκιμάζονται συνολικά 394 μοντέλα (4x4x3x2x2x2) δημιουργώντας έτσι 1920 fits. Το τελικό μοντέλο έχει τις εξής τιμές για υπερπαραμέτρους:

n_estimators	400
learning_rate	0.1
max_depth	3
subsample	0.7
cosample_bytree	0.7
min_child_weight	1

Πίνακας 11: Υπερπαράμετροι Τελικού Μοντέλου XGBoost Παλινδρόμησης

5.3.4 Βαθύ Νευρωνικό Δίκτυο (Neural Network)

Υλοποιείται η βελτιστοποίηση υπερπαραμέτρων ενός νευρωνικού δικτύου (Neural Network) χρησιμοποιώντας Bayesian Optimization. Η Bayesian Optimization είναι μια τεχνική που χρησιμοποιείται για τη βελτιστοποίηση σύνθετων συναρτήσεων, όπου η αξιολόγηση κάθε συνδυασμού υπερπαραμέτρων είναι χρονοβόρα. Αντί να δοκιμάζει όλους τους δυνατούς συνδυασμούς όπως το Grid Search, η μέθοδος αυτή αναζητά τις καλύτερες τιμές υπερπαραμέτρων πιο αποδοτικά, χρησιμοποιώντας ένα στοχαστικό μοντέλο πρόβλεψης. Ο στόχος της βελτιστοποίησης σε αυτό το σενάριο είναι η ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος (MSE) του νευρωνικού δικτύου.

Αρχικά, ορίζεται το εύρος των υπερπαραμέτρων, που περιλαμβάνει:

- num_layers: Το πλήθος των κρυφών στρώσεων (1 έως 8), επιτρέποντας στο μοντέλο να δοκιμάσει τόσο απλές όσο και βαθύτερες αρχιτεκτονικές.
- num_neurons: Ο αριθμός των νευρώνων σε κάθε κρυφή στρώση (4 έως 256), καθορίζοντας τη χωρητικότητα μάθησης του δικτύου.
- learning_rate: Ο ρυθμός εκμάθησης (0.0001 έως 0.1), ο οποίος επηρεάζει το πόσο γρήγορα το μοντέλο προσαρμόζει τα βάρη του.
- batch_size: Το μέγεθος του batch (32 έως 512), δηλαδή πόσα δείγματα θα χρησιμοποιούνται σε κάθε ενημέρωση των βαρών.
- dropout_rate: Το ποσοστό dropout (0.0 έως 0.5), το οποίο βοηθά στην αποφυγή overfitting, απενεργοποιώντας τυχαία ένα ποσοστό των νευρώνων.

Στη συνέχεια, ορίζεται η συνάρτηση train_nn(), η οποία δημιουργεί, εκπαιδεύει και αξιολογεί ένα νευρωνικό δίκτυο. Η συνάρτηση δέχεται ως εισόδους τις υπερπαραμέτρους

και διαμορφώνει ένα μοντέλο. Το νευρωνικό δίκτυο κατασκευάζεται δυναμικά με βάση τις δοθείσες υπερπαραμέτρους. Προστίθεται μια είσοδος με το πλήθος χαρακτηριστικών του X_train.

- Δημιουργούνται num_layers κρυφές στρώσεις, όπου κάθε μία έχει num_neurons νευρώνες με ReLU ενεργοποίηση.
- Σε κάθε στρώση προστίθεται ένα Dropout Layer με dropout_rate, ώστε να μειωθεί η πιθανότητα υπερπροσαρμογής.
- Στην έξοδο του δικτύου, χρησιμοποιείται ένας νευρώνας με sigmoid ενεργοποίηση, κατάλληλος για δυαδική ταξινόμηση.

Ο αλγόριθμος Bayesian Optimization αξιολογεί 5 τυχαίους αρχικούς συνδυασμούς (init_points=5) και στη συνέχεια εκτελεί 25 επιπλέον επαναλήψεις (n_iter=25) προσπαθώντας να βελτιστοποιήσει το MSE. Κατά τη διάρκεια αυτών των 25 επαναλήψεων, η Bayesian Optimization χρησιμοποιεί ένα **Gaussian Process Regression (GPR) μοντέλο**, το οποίο προβλέπει ποιες ρυθμίσεις υπερπαραμέτρων είναι πιθανό να βελτιώσουν την απόδοση του μοντέλου.

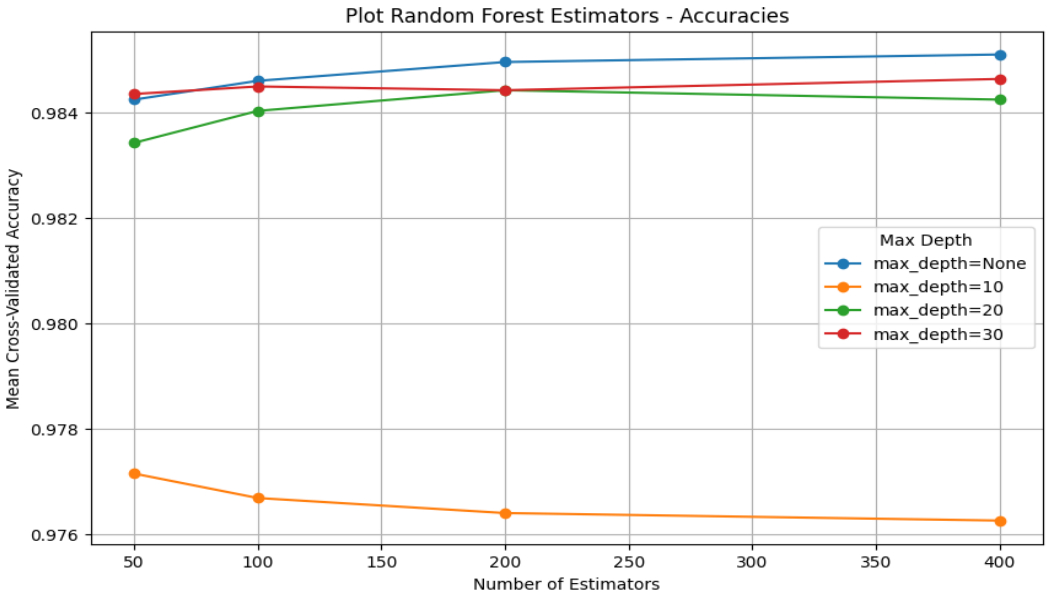
Το τελικό μοντέλο έχει τις εξής υπερπαραμέτρους:

batch_size	512
dropout_rate	0.412736
num_of_layers	2
num_neurons	54
learning_rate	0.006147

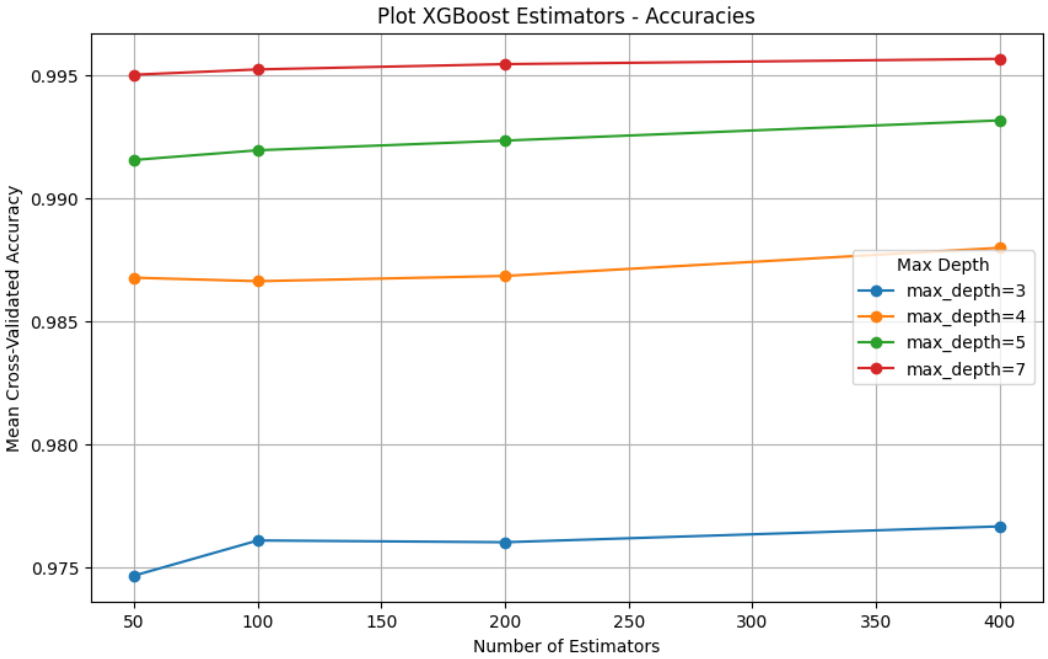
Πίνακας 12: Υπερπαράμετροι Τελικού Νευρωνικού Δικτύου Παλινδρόμησης

Το τελικό νευρωνικό δίκτυο εκπαιδεύεται στο τέλος για 50 εποχές ώστε να δώσει τα τελικά αποτελέσματα και τις προβλέψεις.

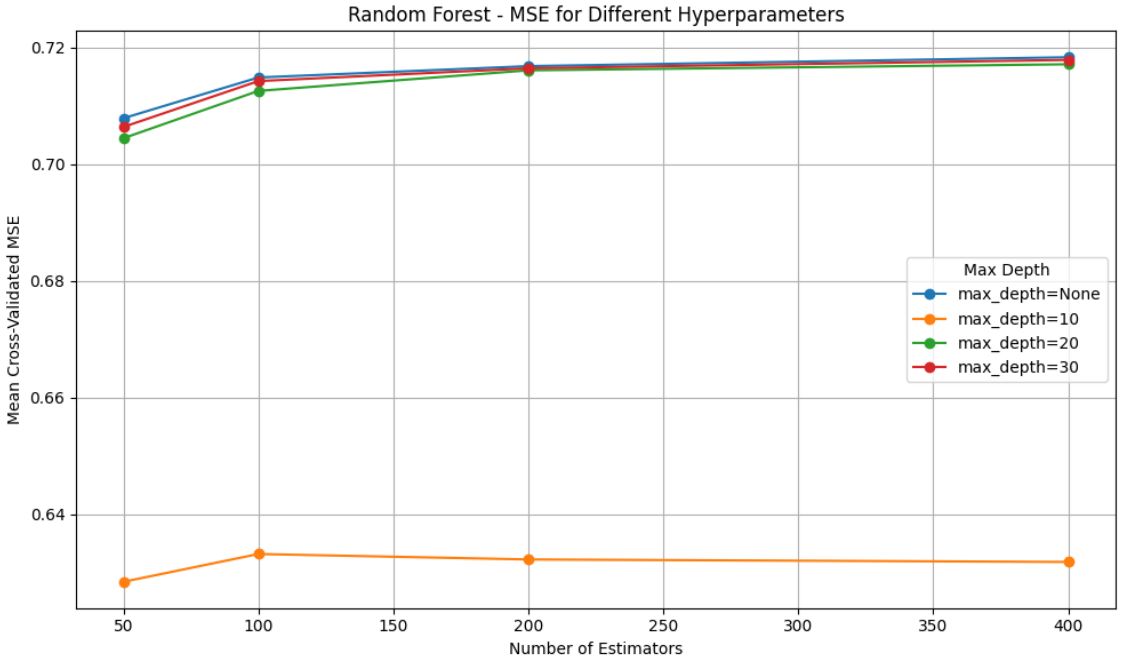
Παρατίθονται παρακάτω διαγράμματα εύρεσης βέλτιστων υπερπαραμέτρων για κάθε μοντέλο που δημιουργήθηκε σε αυτή την εργασία.



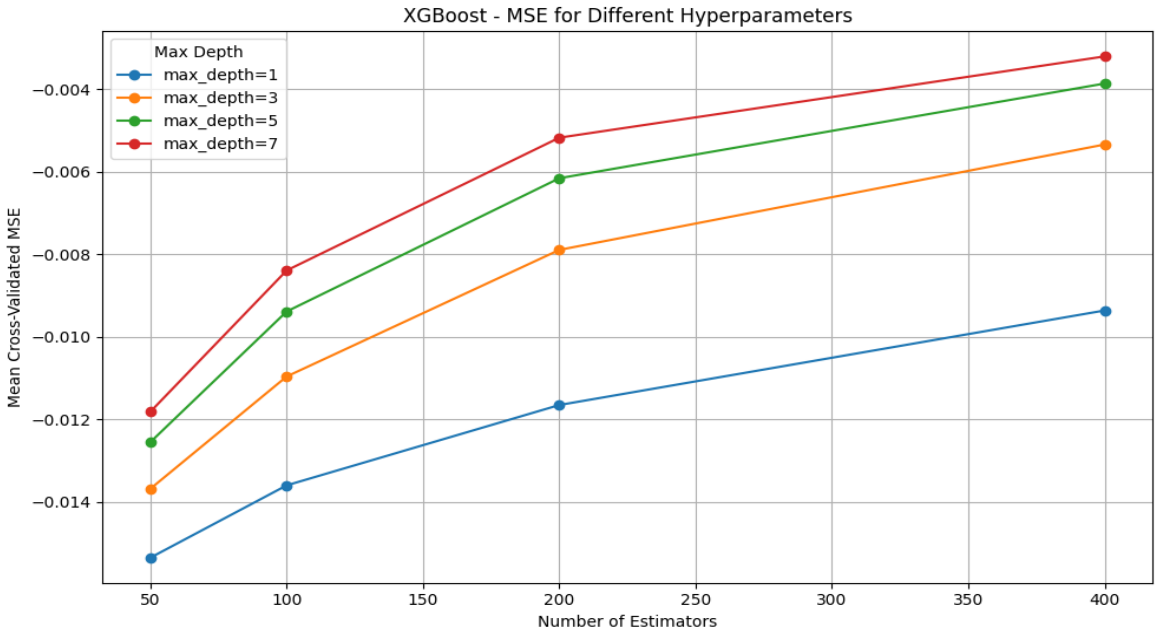
Εικόνα 10: Διάγραμμα αναζήτησης πλέγματος για Τυχαίο Δάσος Κατηγοριοποίησης



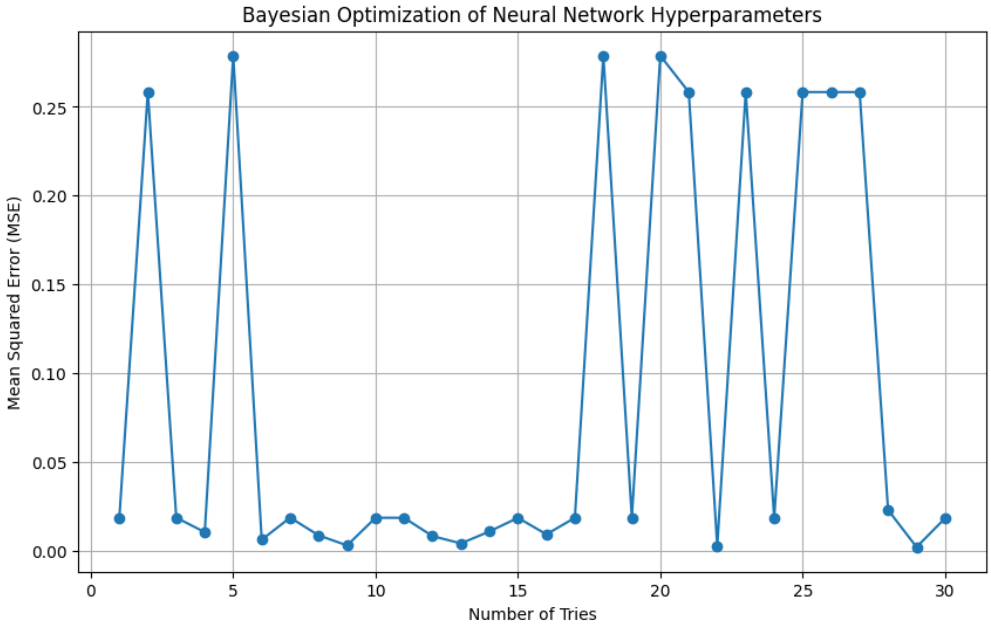
Εικόνα 11: Διάγραμμα αναζήτησης πλέγματος για XGBoost Κατηγοριοποίησης



Εικόνα 12: Διάγραμμα Αναζήτησης Πλέγματος για Τυχαίο Δάσος Παλινδρόμησης



Εικόνα 13: Διάγραμμα Αναζήτησης Πλέγματος για XGBoost Παλινδρόμησης



Εικόνα 14: Διάγραμμα Bayesian Βελτιστοποίησης για Νευρωνικό Δίκτυο Παλινδρόμησης

Κεφάλαιο 6: Αποτελέσματα Μοντέλων

Το παρόν κεφάλαιο συνοψίζει τα αποτελέσματα των μοντέλων που παρουσιάστηκαν εκτενώς στο προηγούμενο κεφάλαιο. Τόσο τα αποτελέσματα των μοντέλων κατηγοριοποίησης, όσο και των μοντέλων παλινδρόμησης, βρίσκονται

6.1 Μοντέλα Κατηγοριοποίησης

Εδώ βρίσκονται μαζεμένα τα αποτελέσματα όλων των μοντέλων που χρησιμοποιήθηκαν σε πίνακες. Υπολογίστηκαν οι τιμές του accuracy, precision, recall και F1 score.

Accuracy

Logistic Regression	0.982
Random Forest	0.9925
XGBoost	0.99725
Neural Network	0.985

Precision

Logistic Regression	0.982002
Random Forest	0.9925
XGBoost	0.99725
Neural Network	0.983

Recall

Logistic Regression	0.982
Random Forest	0.9925
XGBoost	0.99725
Neural Network	0.987

F1 Score

Logistic Regression	0.982
Random Forest	0.9925
XGBoost	0.99725
Neural Network	0.985

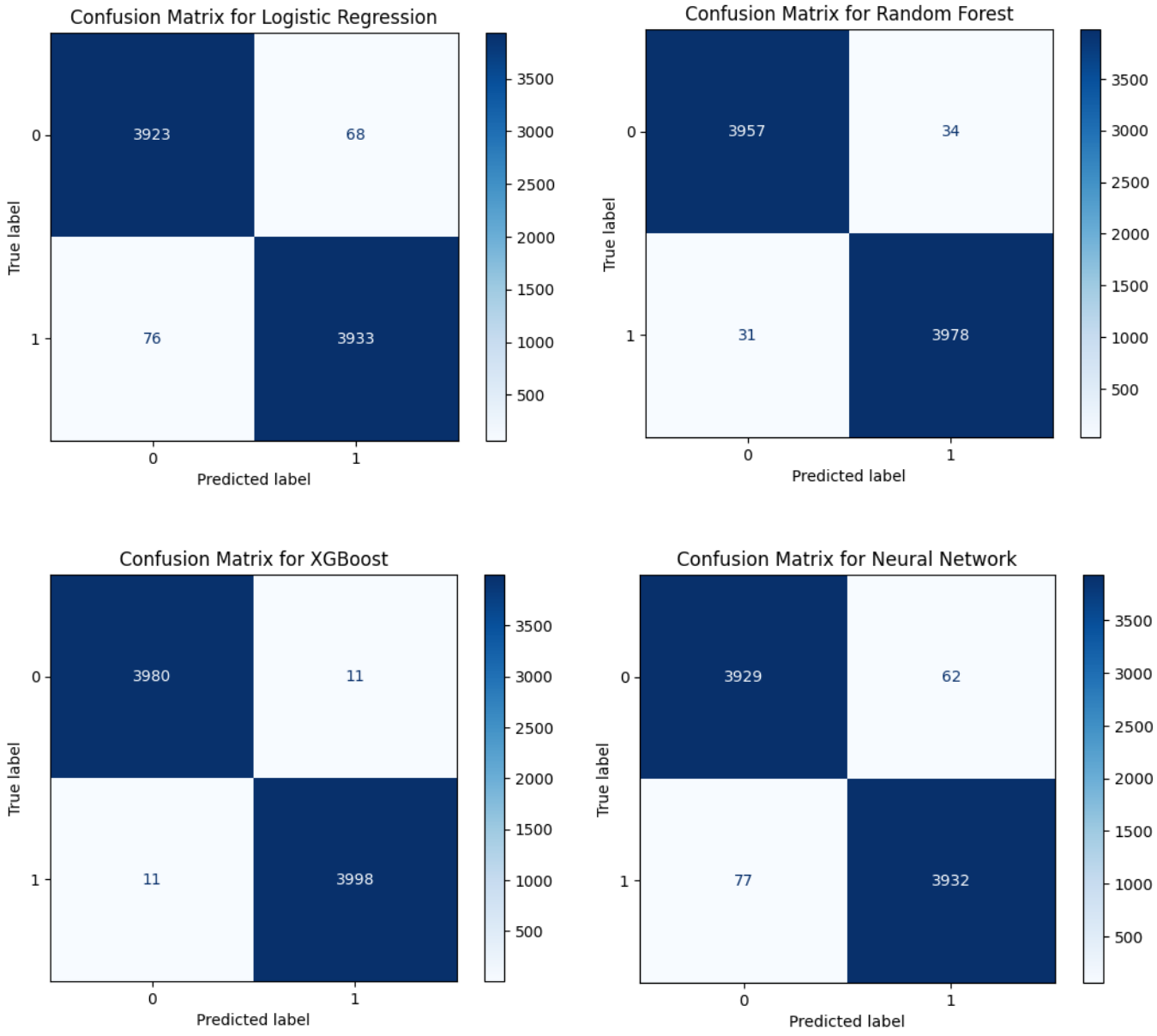
Πίνακας 13: Αποτελέσματα Μοντέλων Ταξινόμησης

Όλα τα μοντέλα κατηγοριοποίησης παρουσίασαν ιδιαίτερα υψηλή απόδοση, επιτυγχάνοντας εξαιρετικές τιμές στις μετρικές accuracy, precision, recall και F1-score. Το μοντέλο XGBoost ξεχώρισε, εμφανίζοντας τις καλύτερες τιμές, με όλες τις μετρικές να φτάνουν σχεδόν το 100% (99,7%), γεγονός που αποδεικνύει την ισχυρή προβλεπτική του ικανότητα.

Συγκριτικά, το μοντέλο Random Forest καθώς και το Νευρωνικό Δίκτυο σημείωσαν επίσης πολύ υψηλές επιδόσεις, αν και υπολείπονται ελαφρώς του XGBoost. Από την άλλη πλευρά, η Λογιστική Παλινδρόμηση παρουσίασε τη χαμηλότερη απόδοση μεταξύ των εξεταζόμενων μοντέλων, με τιμές περίπου στο 98,2%. Ωστόσο, ακόμη και αυτή η επίδοση θεωρείται εξαιρετικά ικανοποιητική, επιβεβαιώνοντας ότι ακόμη και απλούστερα μοντέλα μπορούν να προσφέρουν πολύ αξιόπιστες προβλέψεις για τα συγκεκριμένα δεδομένα.

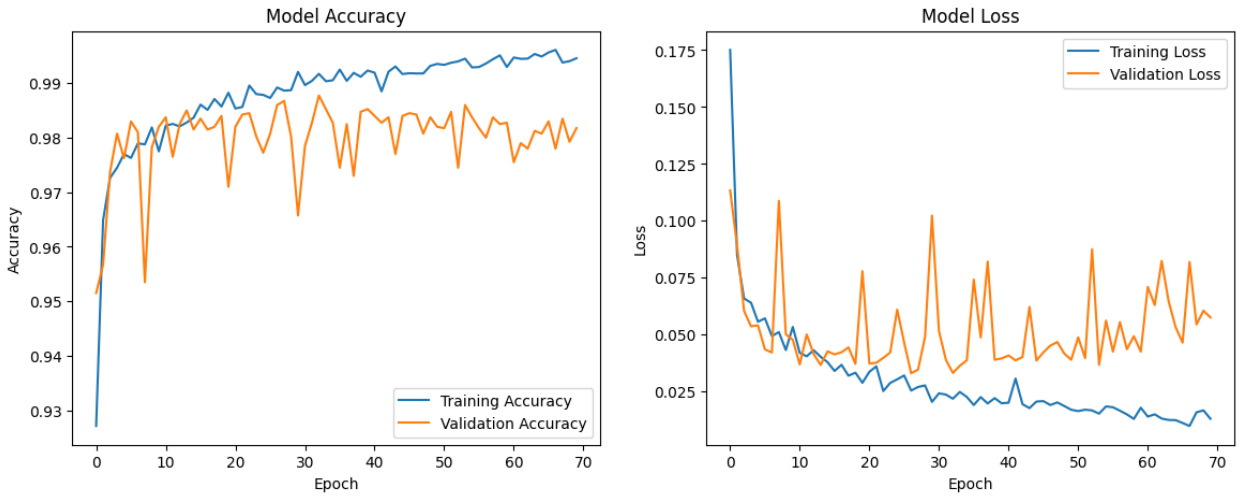
Τέλος, η μικρή διαφορά μεταξύ των τιμών precision και recall σε όλα τα μοντέλα καταδεικνύει την ισορροπία μεταξύ λανθασμένων θετικών και λανθασμένων αρνητικών προβλέψεων. Αυτό οδηγεί στη διαμόρφωση μοντέλων με σταθερή και συνεπή συμπεριφορά, τα οποία μπορούν να θεωρηθούν ισορροπημένα και αξιόπιστα στη διαδικασία κατηγοριοποίησης.

Παρακάτω παρουσιάζονται οι πίνακες σύγχυσης για τα μοντέλα κατηγοριοποίησης.



Εικόνα 15: Πίνακες Σύγχυσης Μοντέλων Ταξινόμησης

Τέλος, δίνονται τα διαγράμματα Accuracy και Loss για την εκπαίδευση του Νευρωνικού Δικτύου.



Εικόνα 16: Διαγράμματα Ακρίβειας και Loss για Νευρωνικό Δίκτυο Ταξινόμησης

6.2 Αποτελέσματα Μοντέλων Παλινδρόμησης

Εδώ βρίσκονται μαζεμένα τα αποτελέσματα όλων των μοντέλων που χρησιμοποιήθηκαν σε πίνακες. Υπολογίστηκαν οι τιμές του MSE, MAE, R^2 και Training Time.

MSE

Linear Regression	0.001893312
Random Forest	0.003851679
XGBoost	0.002034968
Neural Network	0.001957819

MAE

Linear Regression	0.034692525
Random Forest	0.049416653
XGBoost	0.035988906
Neural Network	0.035272741

R^2

Linear Regression	0.897472365
Random Forest	0.79142187
XGBoost	0.889801341
Neural Network	0.893979168

Training Time

Linear Regression	0.003000736
Random Forest	63.03792763
XGBoost	0.673271179
Neural Network	7.437022448

Πίνακας 14: Αποτελέσματα Μοντέλων Παλινδρόμησης

Τα αποτελέσματα που προκύπτουν από την αξιολόγηση των μοντέλων παλινδρόμησης αποδεικνύουν ότι όλα τα μοντέλα που εξετάστηκαν έχουν υψηλή προβλεπτική ικανότητα, κάτι που διαφαίνεται ξεκάθαρα από τις χαμηλές τιμές των μετρικών σφάλματος, δηλαδή το Mean Squared Error (MSE) και το Mean Absolute Error (MAE). Ειδικότερα, οι χαμηλές τιμές των σφαλμάτων αυτών υποδηλώνουν ότι τα μοντέλα είναι σε θέση να προβλέπουν με ιδιαίτερα υψηλή ακρίβεια τη μεταβλητή στόχο rec_rate.

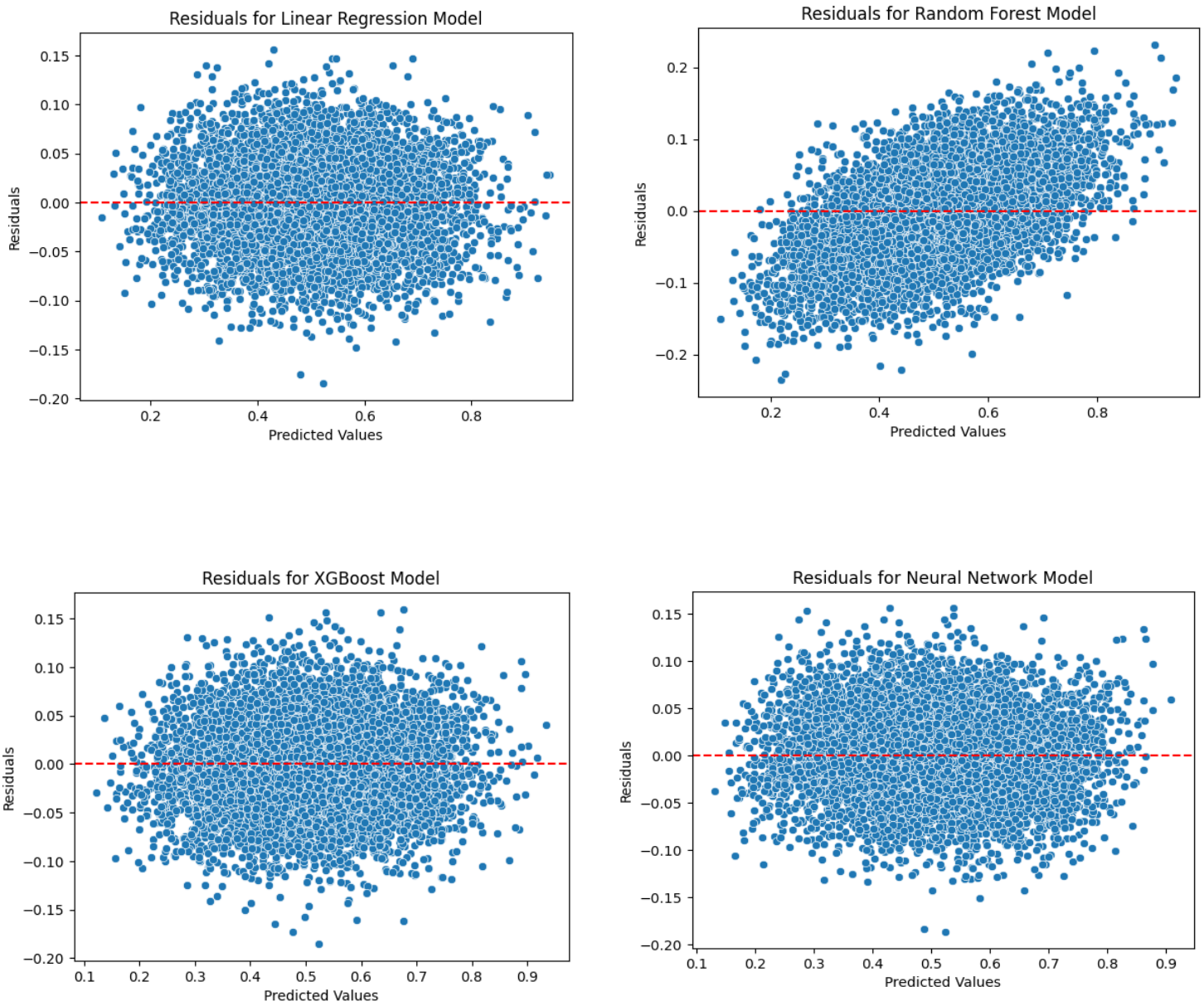
Ειδικότερα, η Γραμμική Παλινδρόμηση ξεχωρίζει, καθώς παρουσιάζει το χαμηλότερο σφάλμα με τιμή MSE ίση με 0.00189 και MAE 0.0347. Αυτό σημαίνει πως οι προβλέψεις της είναι πολύ κοντά στις πραγματικές τιμές, γεγονός που την καθιστά ιδιαίτερα αποτελεσματική και αξιόπιστη. Επιπλέον, ο συντελεστής προσδιορισμού R^2 της Γραμμικής Παλινδρόμησης είναι ο υψηλότερος από όλα τα μοντέλα ($R^2 = 0.8974$), στοιχείο που υπογραμμίζει τη σημαντική της ικανότητα να εξηγεί μεγάλο μέρος της μεταβλητότητας που εμφανίζεται στα δεδομένα.

Ωστόσο, παρά τη γενική καλή απόδοση όλων των μοντέλων, το μοντέλο Τυχαίου Δάσους εμφανίζει αυξημένο χρόνο εκπαίδευσης (περίπου 63 δευτερόλεπτα), γεγονός που μπορεί να αποτελέσει σημαντικό μειονέκτημα στην περίπτωση εφαρμογής του σε πολύ μεγάλα σύνολα δεδομένων ή σε εφαρμογές που απαιτούν συχνή επανεκπαίδευση και γρήγορες προβλέψεις. Επιπρόσθετα, από το διάγραμμα των residuals για το μοντέλο αυτό, προκύπτει μια έντονη τάση που υποδηλώνει πρόβλημα ετεροσκεδαστικότητας, γεγονός που επηρεάζει αρνητικά την αξιοπιστία των προβλέψεών του και απαιτεί περαιτέρω διερεύνηση.

Από την άλλη, τα μοντέλα XGBoost και Νευρωνικού Δικτύου εμφανίζουν αρκετά χαμηλές τιμές σφάλματος και υψηλές τιμές R^2, γεγονός που τα καθιστά ελκυστικές και αξιόπιστες εναλλακτικές επιλογές. Παράλληλα, παρουσιάζουν σημαντικά μικρότερο χρόνο εκπαίδευσης συγκριτικά με το Τυχαίο Δάσος, κάτι που τα καθιστά κατάλληλα για ευέλικτες και γρήγορες πρακτικές εφαρμογές.

Εν κατακλείδι, με βάση την ανάλυση των αποτελεσμάτων φαίνεται ότι η Γραμμική Παλινδρόμηση και το XGBoost είναι οι καλύτερες επιλογές για αξιόπιστες και αποδοτικές προβλέψεις της μεταβλητής rec_rate, ενώ το Νευρωνικό Δίκτυο προσφέρει επίσης πολύ καλή εναλλακτική επιλογή με ικανοποιητική ακρίβεια και σταθερότητα. Το μοντέλο Random Forest, λόγω των χαρακτηριστικών των residuals και του αυξημένου χρόνου εκπαίδευσης, είναι λιγότερο προτεινόμενο χωρίς περαιτέρω βελτιστοποίηση.

Στη συνέχεια παρουσιάζονται τα Υπολειμματικά Σύνολα (Residuals) των μοντέλων μας.

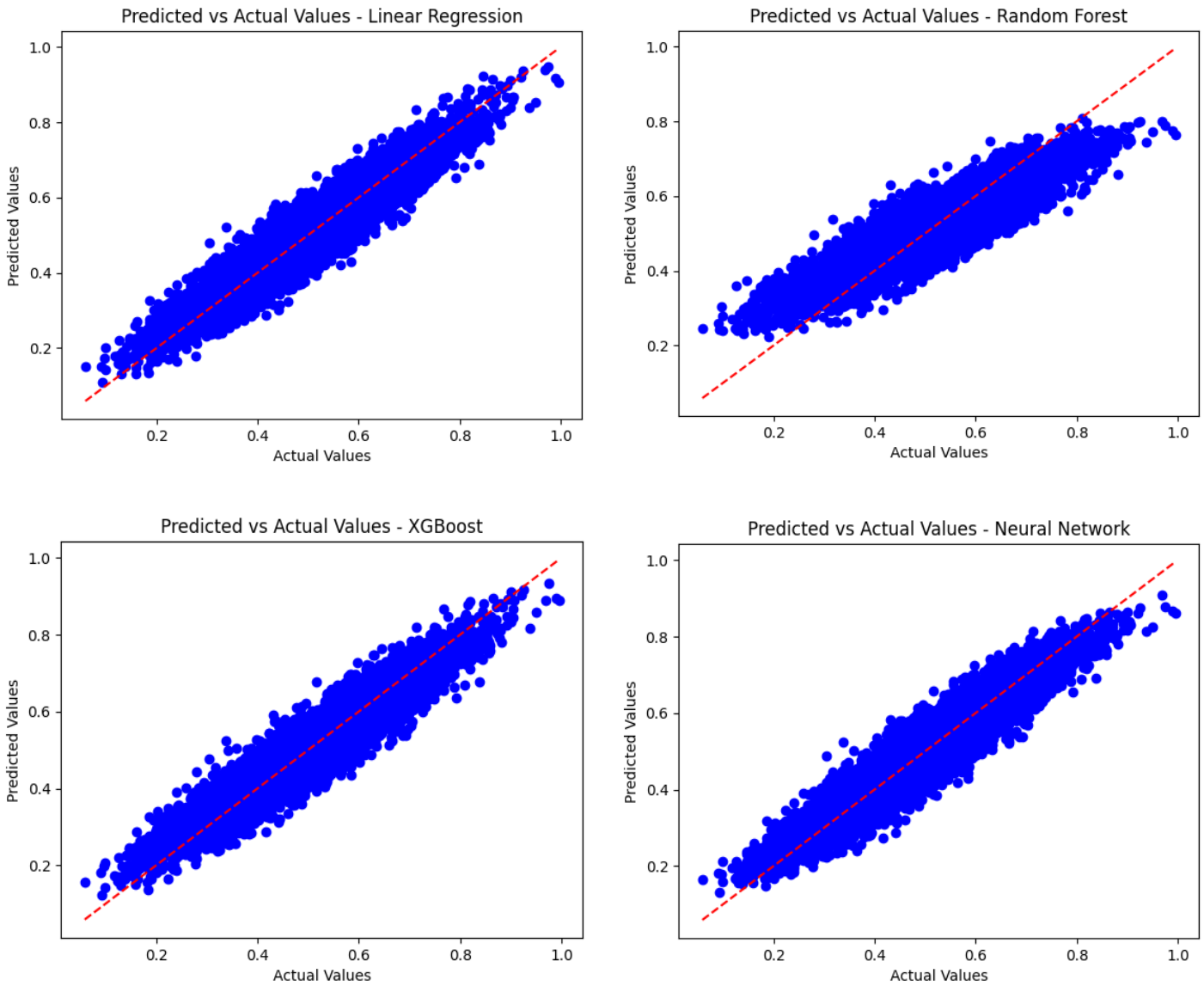


Εικόνα 17: Διαγράμματα Υπολειμμάτων Μοντέλων Παλινδρόμησης

Από την ανάλυση των διαγραμμάτων υπολειμμάτων των μοντέλων προκύπτει ότι η Γραμμική Παλινδρόμηση, το Νευρωνικό Δίκτυο και το XGBoost εμφανίζουν συμμετρική και ομοιόμορφη κατανομή γύρω από το μηδέν, χωρίς σαφή τάση ή ένδειξη ετεροσκεδαστικότητας, επιβεβαιώνοντας ότι πληρούν τις υποθέσεις σταθερής διακύμανσης σφάλματος και είναι ικανά να παρέχουν αξιόπιστες προβλέψεις.

Αντίθετα, το Τυχαίο Δάσος παρουσιάζει διαγώνια κατανομή στα υπολείμματα, γεγονός που υποδεικνύει έντονη ετεροσκεδαστικότητα και ανάγκη περαιτέρω βελτιστοποίησης ή επανεξέτασης του μοντέλου. Επομένως, για το συγκεκριμένο πρόβλημα, η Γραμμική Παλινδρόμηση, το Νευρωνικό Δίκτυο και το XGBoost είναι καταλληλότερα για πρακτική εφαρμογή, ενώ το Τυχαίο Δάσος απαιτεί επιπλέον διερεύνηση.

Παρακάτω παρατίθενται πίνακες σύγκρισης προβλεπόμενων και πραγματικών τιμών για την τιμή rec_rate για κάθε μοντέλο.

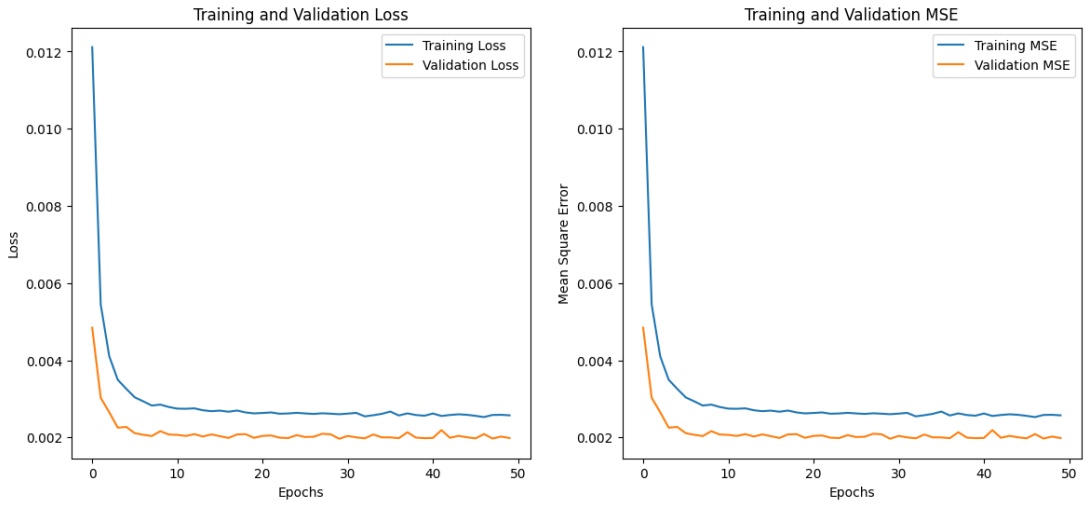


Εικόνα 18: Διαγράμματα Πραγματικών vs Προβλεπόμενων Τιμών Μοντέλων Παλινδρόμησης

Από τα διαγράμματα Predicted vs Actual values προκύπτει ότι τα μοντέλα Γραμμικής Παλινδρόμησης και XGBoost εμφανίζουν την υψηλότερη ακρίβεια και αξιοπιστία, καθώς οι προβλεπόμενες τιμές είναι πολύ κοντά στις πραγματικές και συγκεντρώνονται πυκνά γύρω από την ιδανική διαγώνιο χωρίς εμφανείς τάσεις ή σημαντικές αποκλίσεις.

Το Νευρωνικό Δίκτυο παρουσιάζει επίσης ικανοποιητική απόδοση, με μικρή διασπορά και τυχαίες αποκλίσεις, αν και λίγο υψηλότερη μεταβλητότητα συγκριτικά με τα δύο προηγούμενα μοντέλα. Αντίθετα, το Τυχαίο Δάσος εμφανίζει μεγαλύτερη διασπορά και αξιοσημείωτες αποκλίσεις, ιδιαίτερα στις ακραίες τιμές, γεγονός που υποδεικνύει μειωμένη σταθερότητα και ακρίβεια. Συνεπώς, τα μοντέλα Γραμμικής Παλινδρόμησης και XGBoost αποτελούν τις πιο κατάλληλες επιλογές για το συγκεκριμένο πρόβλημα, ενώ το μοντέλο Τυχαίου Δάσους χρήζει περαιτέρω βελτίωσης.

Παρακάτω βρίσκεται οι καμπύλες του MSE και του Loss κατά την εκπαίδευση του τελικού Νευρωνικού Δικτύου.



Εικόνα 19: Διαγράμματα Εκπαίδευσης Νευρωνικού Δικτύου Παλινδρόμησης

Από τα διαγράμματα της εκπαίδευσης του Νευρωνικού Δικτύου (Training and Validation Loss και MSE) παρατηρείται απότομη πτώση τόσο του Loss όσο και του MSE στις πρώτες εποχές (epochs), γεγονός που υποδηλώνει ότι το μοντέλο προσαρμόζεται ταχύτατα στα δεδομένα. Επίσης, μετά τις πρώτες 10 εποχές, οι τιμές του Loss και του MSE σταθεροποιούνται, δείχνοντας ότι το Νευρωνικό Δίκτυο έχει φτάσει σε μία κατάσταση ισορροπίας και περαιτέρω εκπαίδευση δεν προσφέρει σημαντική βελτίωση στην απόδοση.

Η απόσταση μεταξύ των καμπυλών training και validation είναι μικρή, με τις δύο καμπύλες να συγκλίνουν σχετικά γρήγορα. Αυτό είναι ιδιαίτερα θετικό, διότι σημαίνει ότι δεν υπάρχει υπερπροσαρμογή (overfitting) και το μοντέλο γενικεύει αποτελεσματικά σε νέα, αθέατα δεδομένα. Δεν παρατηρείται αυξητική τάση της καμπύλης του Validation Loss μετά από ορισμένο αριθμό εποχών, στοιχείο που δείχνει ότι δεν έχουμε πρόβλημα υπερπροσαρμογής (overfitting) και ότι το μοντέλο έχει γενικευτική ικανότητα.

Κεφάλαιο 7: Συμπεράσματα και Μελλοντικές Προεκτάσεις

Στην παρούσα διπλωματική εργασία αναζητήθηκαν νέοι τρόποι υπολογισμού της Ζημίας Λόγω Αθέτησης (LGD) και επιλέχθηκε η χρήση της μαθηματικής εξίσωσης (1) που χρησιμοποιείται από την ING. Με αυτό τον τρόπο, το πρόβλημα υπολογισμού του LGD μετατράπηκε σε πρόβλημα εύρεσης των δύο άλλων χαρακτηριστικών, των πιθανοτήτων θεραπείας (Cure Rate) και ανάκτησης (Recovery Rate).

Η εργασία αυτή ανέδειξε την αποτελεσματικότητα των μεθόδων μηχανικής μάθησης στην πρόβλεψη της LGD, με έμφαση στα Cure Rate και Recovery Rate. Τα μοντέλα που χρησιμοποιήθηκαν παρείχαν υψηλή ακρίβεια, ενισχύοντας τη δυνατότητα ακριβέστερης αξιολόγησης του πιστωτικού κινδύνου. Συγκεκριμένα, το μοντέλο XGBoost παρουσίασε την καλύτερη απόδοση στην ταξινόμηση, με ακρίβεια 99,7%, ενώ η Γραμμική Παλινδρόμηση σημείωσε την υψηλότερη ακρίβεια στην παλινδρόμηση με συντελεστή προσδιορισμού R^2=0.8974. Τα μοντέλα Τυχαίου Δάσους και Νευρωνικών Δικτύων εμφάνισαν, επίσης, ισχυρή προβλεπτική ικανότητα, γεγονός που υπογραμμίζει τη σημασία της κατάλληλης επιλογής και ρύθμισης των αλγορίθμων.

Η επιτυχία των μοντέλων επηρεάστηκε σημαντικά από την προεπεξεργασία των δεδομένων, η οποία συνέβαλε στη βελτίωση της ποιότητας των προβλέψεων. Παράλληλα, η χρήση τεχνικών βελτιστοποίησης υπερπαραμέτρων, όπως η αναζήτηση πλέγματος και η Μπεϋζιανή Βελτιστοποίηση, ενίσχυσε την απόδοση των μοντέλων, επιβεβαιώνοντας τη σημασία της προσεκτικής ρύθμισης των παραμέτρων για τη βελτίωση της ακρίβειας και της σταθερότητάς τους.

Για μελλοντική έρευνα, προτείνεται η εφαρμογή των μεθόδων αυτών σε μεγαλύτερα και πιο διαφοροποιημένα σύνολα δεδομένων, ώστε να διασφαλιστεί η γενίκευση των αποτελεσμάτων και η εγκυρότητα των προβλέψεων σε διαφορετικά χρηματοπιστωτικά περιβάλλοντα. Επιπλέον, η διερεύνηση πρόσθετων αλγορίθμων μηχανικής μάθησης και προηγμένων τεχνικών βελτιστοποίησης, όπως η αυτοματοποιημένη μηχανική μάθηση (AutoML) και οι πιο πολύπλοκες αρχιτεκτονικές νευρωνικών δικτύων, μπορεί να βελτιώσει περαιτέρω την απόδοση των μοντέλων. Σημαντική, επίσης, είναι η ανάγκη αντιμετώπισης ζητημάτων όπως η ετεροσκεδαστικότητα και άλλες στατιστικές προκλήσεις, οι οποίες επηρεάζουν την αξιοπιστία των μοντέλων.

Τέλος, η βελτίωση της ερμηνευσιμότητας των μοντέλων μέσω τεχνικών Explainable AI (XAI) αποτελεί ένα κρίσιμο ζήτημα, προκειμένου να ενισχυθεί η διαφάνεια και η αποδοχή τους από τα χρηματοπιστωτικά ιδρύματα. Η περαιτέρω διερεύνηση αυτών των κατευθύνσεων θα συμβάλλει σημαντικά στη βελτίωση της αξιοπιστίας και της πρακτικής χρησιμότητας των μεθόδων μηχανικής μάθησης στη διαχείριση πιστωτικού κινδύνου.

Βιβλιογραφία

1. Basel Committee on Banking Supervision. (2000). *Principles for the Management of Credit Risk*. Bank for International Settlements.
2. Hull, J. C. (2018). *Διαχείριση Κινδύνων και Παραγώγων* (10η έκδ.). Εκδόσεις Κριτική.
3. Saunders, A., & Allen, L. (2020). *Credit Risk Management In and Out of the Financial Crisis* (3η έκδ.). Wiley Finance.
4. Bessis, J. (2015). *Risk Management in Banking* (4η έκδ.). Wiley.
5. Crouhy, M., Galai, D., & Mark, R. (2014). *The Essentials of Risk Management* (2η έκδ.). McGraw-Hill Education.
6. European Banking Authority. (2018). *Guidelines on management of non-performing and forborne exposures*.
7. Altman, E. I., & Saunders, A. (1998). Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance*, 21(11–12), 1721–1742.
8. Basel Committee on Banking Supervision. (2005). *Studies on the Validation of Internal Rating Systems*.
9. Treacy, W. F., & Carey, M. S. (2000). Credit Risk Rating Systems at Large U.S. Banks. *Journal of Banking & Finance*, 24(1–2), 167–201.
10. Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit Scoring and Its Applications*. SIAM.
11. Cantor, R., & Packer, F. (1997). Differences of Opinion and Selection Bias in the Credit Rating Industry. *Journal of Banking & Finance*, 21(10), 1395–1417.
12. White, L. J. (2010). Markets: The Credit Rating Agencies. *Journal of Economic Perspectives*, 24(2), 211–226.
13. Basel Committee on Banking Supervision. (2006). *Guidelines on the Implementation of the Basel II Framework*.
14. Frye, J. (2000). Depressing Recoveries. *Risk*, 13(11), 108–111.
15. Gupton, G. M., Gates, D., & Carty, L. V. (2000). *Bank Loan Loss Given Default*. Moody's Investors Service.
16. Dermine, J., & Neto de Carvalho, C. (2006). Bank loan losses-given-default: A case study. *Journal of Banking & Finance*, 30(4), 1219–1243.
17. Altman, E. I., Brady, B., Resti, A., & Sironi, A. (2005). The Link between Default and Recovery Rates: Theory, Empirical Evidence, and Implications. *Journal of Business*, 78(6), 2203–2227.
18. Acharya, V. V., Bharath, S. T., & Srinivasan, A. (2007). Does Industry-wide Distress Affect Defaulted Firms? *Journal of Financial Economics*, 85(3), 787–821.

19. Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring. *European Journal of Operational Research*, 247(1), 124–136.
20. Moscatelli, M., Narizzano, S., & Vacca, V. (2020). Machine learning for credit risk prediction: Challenges and opportunities. *Bank of Italy Occasional Papers*, No. 591.
21. Zhang, D., Zhou, X., & Zheng, X. (2021). Credit risk evaluation using machine learning: A survey. *Applied Soft Computing*, 98, 106852.
22. Rudin, C. (2019). Stop explaining black box machine learning models. *Nature Machine Intelligence*, 1(5), 206–215.
23. European Banking Authority. (2021). *Report on the use of Big Data and Advanced Analytics in the banking sector*.
24. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2η έκδ.). Springer.
25. Provost, F., & Fawcett, T. (2013). *Data Science for Business*. O'Reilly Media.
26. Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice* (3η έκδ.). OTexts.
27. Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.
28. Freedman, D. A. (2009). *Statistical Models: Theory and Practice* (2η έκδ.). Cambridge University Press.
29. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2η έκδ.). O'Reilly Media.
30. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
31. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2η έκδ.). Springer.
32. Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.
33. Tan, P.-N., Steinbach, M., & Kumar, V. (2018). *Introduction to Data Mining* (2η έκδ.). Pearson.
34. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning* (2η έκδ.). Springer.
35. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3η έκδ.). Wiley.
36. Menard, S. (2010). *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE.
37. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
38. Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, 25(2), 197–227.
39. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD*, 785–794.
40. Nielsen, D. (2016). *Tree Boosting With XGBoost*. NTNU Technical Report.

41. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
42. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
43. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5η έκδ.). Wiley.
44. Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13, 281–305.
45. Feurer, M., & Hutter, F. (2019). Hyperparameter Optimization. In *Automated Machine Learning*. Springer.
46. Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for Hyper-Parameter Optimization. *NeurIPS*, 24, 2546–2554.
47. Bergstra, J., & Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13, 281–305.
48. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2η έκδ.). O'Reilly Media.
49. Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian Optimization of Machine Learning Algorithms. *NeurIPS*, 25, 2951–2959.
50. Shahriari, B., et al. (2016). Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, 104(1), 148–175.
51. Mitchell, M. (1998). *An Introduction to Genetic Algorithms*. MIT Press.
52. Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *IJCAI*, 2, 1137–1143.
53. Severeijns, L. (2018). *Challenging LGD models with Machine Learning*. Vrije Universiteit Amsterdam.
54. Tpoint Tech. (2025). *Supervised Machine Learning*.
55. Wikipedia. (2025). *Logistic function*.
56. MindManager. (2025). *Decision Tree.*
57. Assalamu Al. (2025). *Introduction: Random Forest Classification by Example*. Medium
58. MDPI. (2025). *Impact of Data Pre-Processing Techniques on XGBoost Model Performance for Predicting All-Cause Readmission and Mortality Among Patients with Heart Failure.*
59. ResearchGate. (2025). *Diagram of the 5-fold cross-validation method*.
60. Evidently AI. (2025). *Confusion Matrix*.
61. Raschka, S. (2018, October 22). *Hyperparameter tuning: Always tune your models*. Towards Data Science.
62. Osco, L. P., Marcato Junior, J., Ramos, A. P. M., & Li, J. (2021). *A review on deep learning in UAV remote sensing*. ResearchGate.

Παράρτημα 1: Σύστημα και Βιβλιοθήκες

Παρακάτω παρατίθεται μία σύντομη περιγραφή των βιβλιοθηκών Python που χρησιμοποιήθηκαν στην παρούσα εργασία:

- **NumPy**: Αποτελεί μία από τις βασικότερες βιβλιοθήκες στην Python για επιστημονικούς υπολογισμούς, παρέχοντας αποδοτικές δομές δεδομένων, όπως οι πίνακες πολλαπλών διαστάσεων (arrays), και ένα μεγάλο εύρος συναρτήσεων για γραμμική άλγεβρα και ανάλυση δεδομένων.
- **Pandas**: Χρησιμοποιήθηκε για τη διαχείριση, προεπεξεργασία και ανάλυση των δεδομένων. Παρέχει ευέλικτες δομές δεδομένων, όπως το DataFrame και το Series, που είναι ιδιαίτερα χρήσιμες για την επεξεργασία δεδομένων πίνακα.
- **Matplotlib**: Μία από τις δημοφιλέστερες βιβλιοθήκες οπτικοποίησης δεδομένων στην Python. Χρησιμοποιήθηκε για τη δημιουργία γραφημάτων και διαγραμμάτων που επιτρέπουν την καλύτερη κατανόηση των αποτελεσμάτων της ανάλυσης.
- **Seaborn**: Βιβλιοθήκη που βασίζεται στη Matplotlib και παρέχει υψηλού επιπέδου διεπαφές για στατιστικά γραφήματα, κάνοντας εύκολη την παραγωγή καλαίσθητων και πληροφοριακών γραφημάτων.
- **Scikit-learn**: Η πλέον διαδεδομένη βιβλιοθήκη μηχανικής μάθησης στην Python. Χρησιμοποιήθηκε για την υλοποίηση μοντέλων πρόβλεψης και την αξιολόγησή τους, παρέχοντας πληθώρα αλγορίθμων, μετρικών και τεχνικών προεπεξεργασίας.
- **XGBoost**: Μία ισχυρή και αποδοτική βιβλιοθήκη για την ανάπτυξη μοντέλων βασισμένων σε δενδρικές δομές και μεθόδους ενίσχυσης (boosting). Χρησιμοποιήθηκε για την αύξηση της ακρίβειας των προβλέψεων μέσω προηγμένων μοντέλων boosting.
- **TensorFlow**: Δημοφιλής πλατφόρμα ανοιχτού κώδικα για την ανάπτυξη και εκπαίδευση μοντέλων βαθιάς μάθησης. Παρέχει δυνατότητες για υπολογισμούς σε γράφους και αξιοποίηση GPU για γρήγορη εκπαίδευση νευρωνικών δικτύων.
- **Keras**: Υψηλού επιπέδου API του TensorFlow, το οποίο διευκολύνει τη γρήγορη υλοποίηση και εκπαίδευση μοντέλων βαθιάς μάθησης, κάνοντας τον κώδικα πιο ευανάγνωστο και προσβάσιμο.
- **PyTorch**: Μία ακόμα διαδεδομένη βιβλιοθήκη βαθιάς μάθησης, η οποία παρέχει ευέλικτη ανάπτυξη δυναμικών υπολογιστικών γράφων και απλοποιεί τη διαδικασία δημιουργίας και εκπαίδευσης νευρωνικών δικτύων.

- **skopt.space**: Αποτελεί μέρος του Scikit-Optimize, μιας βιβλιοθήκης που χρησιμοποιήθηκε για τη βελτιστοποίηση των υπερπαραμέτρων των μοντέλων μηχανικής μάθησης μέσω τεχνικών Bayesian optimization, καθιστώντας πιο αποτελεσματική την αναζήτηση του βέλτιστου συνδυασμού παραμέτρων.
- **time**: Πρόκειται για βασική βιβλιοθήκη της Python, η οποία χρησιμοποιήθηκε για τη μέτρηση και καταγραφή του χρόνου εκτέλεσης των διάφορων τμημάτων του κώδικα, συμβάλλοντας στη βελτιστοποίηση της αποδοτικότητας του μοντέλου και της διαδικασίας.

Παράρτημα 2

Παρακάτω παρατίθενται οι συντελεστές (coefficients) και οι σταθερές (intercepts) που ανακτήθηκαν από τα εκπαιδευμένα μοντέλα λογιστικής και γραμμικής παλινδρόμησης:

Γραμμική Παλινδρόμηση

Intercept	6.1736E+11
Coefficients	0.040493693
	0.083766398
	0.066196428
	0.056271803
	0.020800564
	0.022766967
	0.013518103
	0.079899364
	0.000280577
	0.050036806
	0.046307385
	0.204923487
	0.075119263
	0.071495146
	0.024514156
	0.040309291
	0.016644763
	0.022964318
	0.060177505
	0.042569875
	0.028396298
	0.057696792
	0.020961857
	0.033327251
	0.037378343
	0.044194111
	0.070927268
	0.025188469
	0.048773814
	0.054580963
	0.012734267

	0.056786779
	0.022066531
	568862.93
	-1100326.46
	539337.244
	0.262069466
	0.032870122
	0.016866591
	-6.25463E+11
	-5.9028E+11
	-1.79344E+11
	-4.96699E+11
	-1.2066E+11
	8103055990
	-4.38016E+11
	8103055990
	8103055990
	-4.38016E+11
	8103055990
	-4.38016E+11
	-27079375700
	-27079375700
	-4.38016E+11
	-27079375700
	-1.2066E+11
	-1.2066E+11
	-27079375700
	-1.20660480e+11

Λογιστική Παλινδρόμηση

intercept	-7.95772406
coefficients	-2.41E-01
	-0.093003161
	0.043830929
	-0.053850792
	-0.144504161
	-0.093604861
	-0.051427464
	0.720369383
	0.794104505
	-3.02005389
	0.034837661
	0.103714203
	-0.072863299
	-0.119547089
	-0.197882473
	-15.0095447
	-19.3393174
	-1.76937987
	-6.62730543
	0.027012381
	-0.000368243
	-0.164693571
	-0.08448301
	-0.061926024
	-0.068914119
	-0.097821527
	0.023746361
	0.07973991
	24.9352975
	0.01353574
	-0.049807418
	0.015931373
	0.007172148
	-0.12474167
	-0.061964708
	-0.062612963
	-0.008950074

	-0.054176592
	-0.086430657
	-0.026850243
	-1.94711934
	-1.98865853
	-1.98144778
	-2.04049841
	-0.540583219
	-0.416422039
	-0.515690044
	-0.601099101
	-0.460895813
	-0.413319941
	-0.468702385
	-0.641306332
	-0.328175969
	-0.526071564
	-0.411131466
	-0.572954415
	-0.493832078
	-0.510702778
	-0.561456586
	-0.495