2025-06

þÿ ¤ µ Ç ½ ¹ º - Â ± ¾ ¹ ¿ » Ì ³ · Ã · Â þÿ Ç Á · ¼ ± Ä ¿ ¿ ¹ º ¿ ½ ¿ ¼ ¹ º Î ½ º ¹ ½ ´ Í ½ É þÿ µ À ¹ Ç µ ¹ Á · ¼ ± Ä ¹ º - Â ± À ¿ Æ ¬ Ã µ ¹ Â þÿ Ä · Â ± ½ ¬ » Å Ã · Â ´ µ ´ ¿ ¼ - ½ É ½

þÿ – Î Ä ¿ Å , • Å Á Å ´ ¯ º ·

þÿ œ µ Ä ± À Ä Å Ç ¹ ± º Ì Á Ì ³ Á ± ¼ ¼ ± Ã Ä · ½ • ³ º » · ¼ ± Ä ¿ » ¿ ³ ¹ º ® › ¿ ³ ¹ Ã Ä ¹ º ® º ± ¹ Ä ± § Á · ¼ ±
þÿ £ Ç ¿ » ® Ÿ ¹ º ¿ ½ ¿ ¼ ¹ º Î ½ • À ¹ Ã Ä · ¼ Î ½ º ± ¹ ” ¹ ¿ ¯ º · Ã · Â , ± ½ µ À ¹ Ã Ä ® ¼ ¹ ¿ • µ ¬ À ¿ » ¹ Â

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

Πανεπιστήμιο Νεάπολις Πάφος

«Τεχνικές αξιολόγησης χρηματοοικονομικών κινδύνων στις επιχειρηματικές αποφάσεις μέσω της ανάλυσης δεδομένων»

Ζώτου Ευρυδίκη

Ιούνιος, 2025

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ ΤΕΧΝΟΛΟΓΙΑ

Διπλωματική Εργασία η οποία υποβλήθηκε προς απόκτηση εξ αποστάσεως μεταπτυχιακού τίτλου σπουδών στην ανάλυση δεδομένων και χρηματοοικονομική τεχνολογία στο Πανεπιστήμιο Νεάπολις Πάφος.

«Τεχνικές αξιολόγησης χρηματοοικονομικών κινδύνων στις επιχειρηματικές αποφάσεις μέσω της ανάλυσης δεδομένων»

Ζώτου Ευρυδίκη

Ιούνιος, 2025

Η ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

Η Ζώτου Ευρυδίκη, γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Τεχνικές αξιολόγησης χρηματοοικονομικών κινδύνων στις επιχειρηματικές αποφάσεις μέσω της ανάλυσης δεδομένων», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.

Η Δηλούσα

Ζώτου Ευρυδίκη

Πίνακας περιεχομένων

Πίνακας Εικόνων

Πίνακας Πινάκων

Τίτλος Διπλωματικής Εργασίας: Τεχνικές αξιολόγησης χρηματοοικονομικών κινδύνων στις επιχειρηματικές αποφάσεις μέσω της ανάλυσης δεδομένων

Η παρούσα Διπλωματική Εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση εξ αποστάσεως μεταπτυχιακού τίτλου στο Πανεπιστήμιο Νεάπολις και εγκρίθηκε στις 20/06/2025 από τα μέλη της Εξεταστικής Επιτροπής.

Εξεταστική Επιτροπή:

Πρώτος επιβλέπων (Πανεπιστήμιο Νεάπολις Πάφος): Δρ. Κωνσταντίνος Παναγιωτάκης, Καθηγητής

Μέλος Εξεταστικής Επιτροπής: Χρήστος Λεμονάκης, Αναπληρωτής Καθηγητής

Μέλος Εξεταστικής Επιτροπής: Γεώργιος Μαστοράκης, Αναπληρωτής Καθηγητής

Περίληψη

Η αποτελεσματική αξιολόγηση και διαχείριση του χρηματοοικονομικού κινδύνου αποτελεί θεμέλιο λίθο για τη βιωσιμότητα και την ανάπτυξη των σύγχρονων επιχειρήσεων. Στην παρούσα διπλωματική εργασία εξετάζεται η εφαρμογή προηγμένων τεχνικών ανάλυσης δεδομένων στη διαχείριση πιστωτικού κινδύνου, με έμφαση στη συγκριτική αξιολόγηση δύο ευρέως χρησιμοποιούμενων μεθόδων: της λογιστικής παλινδρόμησης και των δέντρων αποφάσεων. Αρχικά, παρουσιάζεται το θεωρητικό πλαίσιο του χρηματοοικονομικού κινδύνου και η σημασία της αξιοποίησης δεδομένων στις επιχειρηματικές αποφάσεις. Στη συνέχεια, αναλύονται οι βασικές αρχές, τα πλεονεκτήματα και οι περιορισμοί των επιλεγμένων μεθόδων, με έμφαση στη δυνατότητα πρόβλεψης και ερμηνευσιμότητας των αποτελεσμάτων.

Για την εμπειρική αξιολόγηση, χρησιμοποιήθηκε πραγματικό σύνολο δεδομένων πιστωτικής αξιολόγησης από την πλατφόρμα Kaggle, το οποίο υπέστη εκτενή προεπεξεργασία, ανάλυση και αντιμετώπιση ανισορροπίας τάξεων. Τα μοντέλα εκπαιδεύτηκαν και αξιολογήθηκαν με βάση πολλαπλές μετρικές απόδοσης (accuracy, precision, recall, F1-score, AUC-ROC), ενώ εξετάστηκαν και ζητήματα ερμηνευσιμότητας, κόστους σφαλμάτων και πρακτικής εφαρμογής σε πραγματικές επιχειρηματικές συνθήκες. Τα αποτελέσματα ανέδειξαν τη σχετική υπεροχή κάθε μεθόδου υπό διαφορετικές συνθήκες: η λογιστική παλινδρόμηση υπερτερεί σε περιβάλλοντα όπου προέχει η διαφάνεια και η ερμηνευσιμότητα, ενώ τα δέντρα αποφάσεων προσφέρουν αυξημένη ευελιξία και ικανότητα ανίχνευσης μη γραμμικών συσχετίσεων. Επιπλέον, διαπιστώθηκε ότι η ενσωμάτωση συνδυαστικών μεθόδων (ensemble learning) μπορεί να ενισχύσει περαιτέρω την ακρίβεια και την αξιοπιστία της πρόβλεψης.

Η εργασία καταλήγει σε πρακτικές προτάσεις για την επιλογή και εφαρμογή των κατάλληλων τεχνικών ανάλογα με το επιχειρηματικό πλαίσιο και τους διαθέσιμους πόρους, ενώ προτείνει μελλοντικές ερευνητικές κατευθύνσεις. Τα ευρήματα αναδεικνύουν τη σημασία της πολυδιάστατης και επιστημονικά τεκμηριωμένης προσέγγισης στη διαχείριση χρηματοοικονομικού κινδύνου, προσφέροντας πολύτιμα εργαλεία για τη στρατηγική λήψη επιχειρηματικών αποφάσεων.

Λέξεις κλειδιά: Χρηματοοικονομικός κίνδυνος, πιστωτικός κίνδυνος, ανάλυση δεδομένων, λογιστική παλινδρόμηση, δέντρα αποφάσεων, credit scoring, ensemble learning, διαχείριση επιχειρηματικού κινδύνου.

Abstract

Effective assessment and management of financial risk are fundamental for the sustainability and growth of modern enterprises. This thesis investigates the application of advanced data analysis techniques in the context of credit risk management, with a particular focus on the comparative evaluation of two widely used methods: logistic regression and decision trees. Initially, the theoretical framework of financial risk is presented, highlighting the importance of data-driven decision-making in business environments. Subsequently, the fundamental principles, advantages, and limitations of the selected methods are analyzed, emphasizing predictive performance and interpretability.

For the empirical analysis, a real-world credit scoring dataset from Kaggle was utilized, which underwent extensive preprocessing, exploratory data analysis, and class imbalance treatment. The models were trained and evaluated using multiple performance metrics (accuracy, precision, recall, F1-score, AUC-ROC), while issues of interpretability, error cost, and practical applicability in real business contexts were also examined. The results reveal the relative superiority of each method under different conditions: logistic regression excels in environments where transparency and interpretability are priorities, whereas decision trees offer enhanced flexibility and the ability to capture non-linear relationships. Furthermore, it was demonstrated that the integration of ensemble methods can further improve predictive accuracy and reliability.

The thesis concludes with practical recommendations for the selection and application of appropriate techniques according to business context and available resources and proposes future research directions related to the adoption of more sophisticated models, the use of heterogeneous data, and the enhancement of algorithmic decision interpretability. The findings underscore the importance of a multidimensional and scientifically validated approach to financial risk management, providing valuable tools for strategic business decision-making.

Keywords: Financial risk, credit risk, data analysis, logistic regression, decision trees, credit scoring, ensemble learning, business risk management.

ΚΕΦΑΛΑΙΟ 1: ΕΙΣΑΓΩΓΗ

1.1 Εισαγωγή

Ο χρηματοοικονομικός κίνδυνος αποτελεί έναν από τους σημαντικότερους παράγοντες που επηρεάζουν τη βιωσιμότητα και ανάπτυξη μιας επιχείρησης. Αναφέρεται στην πιθανότητα απώλειας χρηματοοικονομικών πόρων λόγω μεταβολών στις αγορές, επιτόκια, συναλλαγματικές ισοτιμίες ή οικονομικές συνθήκες. Στον παγκοσμιοποιημένο οικονομικό χώρο, οι κίνδυνοι γίνονται όλο και πιο πολύπλοκοι, καθώς οι αλληλεπιδράσεις μεταξύ οικονομικών μεταβλητών εντείνονται από παγκόσμιες κρίσεις, γεωπολιτικές αναταραχές και τεχνολογικές ανακατατάξεις. Η αποτελεσματική αξιολόγηση και διαχείριση αυτού του κινδύνου συνιστά θεμέλιο λίθο για την μακροπρόθεσμη στρατηγική κάθε οργανισμού.

Οι επιχειρήσεις, προκειμένου να επιτύχουν βιώσιμη ανάπτυξη και να διασφαλίσουν τη χρηματοπιστωτική τους σταθερότητα, καλούνται να αναλύσουν, να μετρήσουν και να διαχειριστούν τον κίνδυνο με την βοήθεια προηγμένων τεχνικών ανάλυσης δεδομένων. Η διαχείριση κινδύνου δεν αφορά μόνο την αποφυγή απωλειών, αλλά και τη βελτιστοποίηση των ευκαιριών, ισορροπώντας μεταξύ απόδοσης και ασφάλειας.

Η παραδοσιακή αξιολόγηση του χρηματοοικονομικού κινδύνου βασιζόταν σε στατιστικές τεχνικές και στην χρηματοοικονομική ανάλυση. Ωστόσο, με την ραγδαία ανάπτυξη της τεχνολογίας και τη διαθεσιμότητα μεγάλων όγκων δεδομένων, νέες μέθοδοι που αξιοποιούν τεχνικές ανάλυσης δεδομένων και μηχανικής μάθησης έχουν αναδυθεί. Αυτές οι τεχνικές επιτρέπουν την πιο ακριβή και δυναμική πρόβλεψη πιθανών κινδύνων, διευκολύνοντας τη λήψη αποφάσεων με βάση τα δεδομένα.

1.2 Στόχοι

Ο βασικός σκοπός της παρούσας μελέτης είναι να διευρυνθεί και να συγκριθεί η αποτελεσματικότητα δύο διαφορετικών τεχνικών ανάλυσης δεδομένων, της Λογιστικής Παλινδρόμησης (Logistic Regression) και των Δέντρων Αποφάσεων (Decision Trees), στην αξιολόγηση χρηματοοικονομικού κινδύνου σε επιχειρηματικές αποφάσεις. Η μελέτη εστιάζει στην ικανότητα των δύο μεθόδων να εντοπίζουν κινδύνους που ενδέχεται να επηρεάσουν την οικονομική ανάπτυξη και τη στρατηγική κατεύθυνση των επιχειρήσεων.

Η ανάγκη για ακριβείς και διαφανείς τεχνικές πρόβλεψης κινδύνου γίνεται ολοένα και πιο επιτακτική, καθώς οι οργανισμοί αντιμετωπίζουν ένα ολοένα αυξανόμενο δυναμικό περιβάλλον αβεβαιότητας. Η χρήση τεχνικών μηχανικής μάθησης, όπως η Λογιστική Παλινδρόμηση και τα Δέντρα Αποφάσεων, προσφέρει τη δυνατότητα για καλύτερη πρόβλεψη και διαχείριση του χρηματοοικονομικού κινδύνου.

Συγκεκριμένα, η μελέτη στοχεύει να εξετάσει συγκριτικά τις δυνατότητες αυτών των δύο μεθόδων, απαντώντας σε τέσσερα βασικά ερωτήματα: την ακρίβεια πρόβλεψης, την ερμηνευσιμότητα, την επίδραση των χαρακτηριστικών του δείγματος και τις πρακτικές εφαρμογές τους. Το πρώτο ερευνητικό ερώτημα αφορά τη συγκριση της ακρίβειας μεταξύ της Λογιστικής Παλινδρόμησης και των Δέντρων Αποφάσεων. Η αξιολόγηση της ακρίβειας θα πραγματοποιήθεί μέσω πολλαπλών μετρικών, όπως το accuracy, το precision, το recall, το F1-score και η καμπύλη AUC-ROC. Η ανάλυση θα διερευνήσει υπό ποιες συνθήκες η κάθε μέθοδος υπερτερεί, λαμβάνοντας υπόψη τη φύση των δεδομένων και την πολυπλοκότητα του προβλήματος. Το δεύτερο ερευνητικό ερώτημα επικεντρώνεται στην ερμηνευσιμότητα (interpretability) των δύο μεθόδων και στην σχέση της με τη λήψη επιχειρηματικών αποφάσεων. Η απάντηση θα περιλαμβάνει την ανάλυση της διαφάνειας των μοντέλων και συζήτηση μεταξύ ακρίβειας και ερμηνευσιμότητας. Το τρίτο ερευνητικό ερώτημα διερευνά πως η απόδοση των δύο μεθόδων επηρεάζεται από τα χαρακτήριστικά του δείγματος, όπως η ισορροπία των κατηγοριών και το μέγεθος του δείγματος. Τέλος, το τέταρτο ερευνητικό ερώτημα αφορά τις πρακτικές προεκτάσεις της χρήσης κάθε μεθόδου σε πραγματικές επιχειρηματικές συνθήκες.

Η μελέτη θα συνδυάσει ποσοτική ανάλυση, δηλαδή πειράματα με datasets χρηματοοικονομικών δεδομένων και ποιοτική αξιολόγηση, δηλαδή ερμηνευσιμότητα, εφαρμογή στην πράξη για να δώσει μια ολοκληρωμένη απάντηση στα ερωτήματα. Τα αποτελέσματα θα μπορούσαν να καθοδηγήσουν επιχειρήσεις στην επιλογή της βέλτιστης τεχνικής ανάλογα με τις ανάγκες τους.

ΚΕΦΑΛΑΙΟ 2: ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

2.1 Θεωρητική Ανάλυση του Χρηματοοικονομικού Κινδύνου

Το κεφάλαιο αυτό ξεκινά με την θεωρητική προσέγγιση του χρηματοοικονομικού κινδύνου. Ο χρηματοοικονομικός κίνδυνος αναφέρεται στην πιθανότητα μιας επιχείρησης, ενός ατόμου ή ενός οργανισμού να υποστεί απώλειες λόγω δυσμενών οικονομικών συνθηκών, αστάθειας των αγορών ή λανθασμένων αποφάσεων. Είναι ένα κρίσιμο στοιχείο στη λήψη επενδυτικών, δανειακών και επιχειρηματικών αποφάσεων και επηρεάζει όλους τους τομείς της οικονομίας, από τις τράπεζες μέχρι τις εταιρείες και τα κράτη (Hull, 2018). Στη συνέχεια, αναλύονται όλες οι κατηγορίες του χρηματοοικονομικού κινδύνου.

Ο πιστωτικός κίνδυνος, ίσως ο πιο γνωστός και διαχρονικός, αναφέρεται στην πιθανότητα μη αποπληρωμής χρηματοοικονομικών υποχρεώσεων. Αυτός εμφανίζεται σε τρεις κύριες μορφές: τον κίνδυνο αθέτησης, όπου ο οφειλέτης δεν μπορεί να εκπληρώσει τις υποχρεώσεις του, τον κίνδυνο υποβάθμισης της πιστοληπτικής ικανότητας και τον κίνδυνο συγκεντρωτικού εκθέματος, όταν υπάρχει υπερβολική έκθεση σε συγκεκριμένους οφειλέτες. Οι επιπτώσεις του πιστωτικού κινδύνου γίνονται ιδιαίτερα εμφανείς κατά περιόδους οικονομικής ύφεσης, όπου η μείωση των εσόδων και η αύξηση της ανεργίας οδηγούν σε αυξημένα ποσοστά αθέτησης (Saunders & Cornett, 2011).

Παράλληλα, ο αγοραστικός κίνδυνος αποτελεί μια άλλη κρίσιμη διάσταση, σχετιζόμενη με τις διακυμάνσεις στις τιμές χρηματοοικονομικών μέσων. Αυτός περιλαμβάνει τον κίνδυνο τιμών μετοχών, τον κίνδυνο επιτοκίων, τον συναλλαγματικό κίνδυνο και τον κίνδυνο εμπορευμάτων. Μια ιδιαίτερη διάκριση γίνεται μεταξύ απόλυτου κινδύνου, που μετρά τις απώλειες σε νομισματικές μονάδες, και σχετικού κινδύνου, που αφορά αποκλίσεις από το σημείο αναφοράς (benchmark). Η μέτρηση του αγοραστικού κινδύνου γίνεται συχνά μέσω της ανάλυσης Value-at-Risk (VaR) και των δοκιμών στρες (stress testing), οι οποίες προσφέρουν πολύτιμες πληροφορίες για τις πιθανές απώλειες υπό ακραίες συνθήκες (Jorion, 2020).

Ο λειτουργικός κίνδυνος περιλαμβάνει τον κίνδυνο απωλειών από εσωτερικές διαδικασίες, ανθρώπινα λάθη, τεχνολογικές αστοχίες και εξωτερικά γεγονότα. Αυτός εκδηλώνεται μέσω διαφόρων μορφών, όπως ο κίνδυνος νομισματικού αδικήματος, ο κίνδυνος συστημικών αστοχιών και ο κίνδυνος καταστροφικών γεγονότων. Οι τράπεζες και οι χρηματοοικονομικοί οργανισμοί δίνουν ιδιαίτερη προσοχή σε αυτόν τον κίνδυνο, καθώς μπορεί να οδηγήσει σε σημαντικές οικονομικές ζημίες και υποβάθμιση της φήμης (Jorion, 2020).

Ένας εξίσου σημαντικός κίνδυνος είναι ο κίνδυνος ρευστότητας, ο οποίος διακρίνεται σε δύο βασικές μορφές: τη ρευστότητα αγοράς, που αφορά την αδυναμία πώλησης περιουσιακών στοιχείων χωρίς σημαντική απώλεια αξίας, και τη χρηματοπιστωτική ρευστότητα, που σχετίζεται με την αδυναμία κάλυψης βραχυπρόθεσμων υποχρεώσεων. Οι κύριες αιτίες αυτού του κινδύνου περιλαμβάνουν την ασυμμετρία πληροφοριών, τις οικονομικές κρίσεις και τις απρόβλεπτες μαζικές αναλήψεις, φαινόμενα που μπορούν να δημιουργήσουν σοβαρές δυσκολίες στους χρηματοοικονομικούς οργανισμούς (Hull, 2018).

Τέλος, ο κίνδυνος επιτοκίων αποτελεί μια κρίσιμη διάσταση του χρηματοοικονομικού κινδύνου, αναφερόμενη στις δυνητικές απώλειες που προκύπτουν από τις διακυμάνσεις των επιτοκίων στην οικονομία. Αυτός ο κίνδυνος εκδηλώνεται κυρίως μέσω τριών μηχανισμών: του κινδύνου τιμής, όπου οι αυξήσεις των επιτοκίων οδηγούν σε μείωση της αγοραστικής αξίας των υφιστάμενων ομολόγων λόγω της αντιστρόφου σχέσης μεταξύ τιμών ομολόγων και επιτοκίων, του κινδύνου επανεπένδυσης, που σχετίζεται με τη δυσκολία επανεπένδυσης των εσόδων από κουπόνια ή λήξεις σε συγκρίσιμα ή υψηλότερα επιτόκια, και του κινδύνου καμπύλης απόδοσης, όπου αλλαγές στο σχήμα της καμπύλης επιτοκίων επηρεάζουν διαφορετικά χρεόγραφα ανάλογα με τη διάρκειά τους. Οι κύριες αιτίες διακυμάνσεων των επιτοκίων περιλαμβάνουν τις αποφάσεις της χρηματοοικονομικής πολιτικής των κεντρικών τραπεζών, τις μεταβολές του πληθωρισμού, την οικονομική ανάπτυξη και τις συνθήκες της παγκόσμιας κεφαλαιαγοράς (Hull, 2018).

2.2 Σημασία της Ανάλυσης Δεδομένων στην Διαχείριση Κινδύνου

Στη σύγχρονη εποχή, η ανάλυση δεδομένων αναδεικνύεται σε έναν από τους πιο καθοριστικούς παράγοντες για την επιτυχημένη διαχείριση κινδύνων, καθιστώντας τη διαδικασία όχι μόνο πιο ακριβή αλλά και ουσιαστικά στρατηγική. Σε έναν κόσμο όπου η αβεβαιότητα και η πολυπλοκότητα των αγορών και των επιχειρησιακών περιβαλλόντων αυξάνονται εκθετικά, η ικανότητα των οργανισμών να συλλέγουν, να επεξεργάζονται και να αναλύουν μεγάλους όγκους δεδομένων αποτελεί ζωτικής σημασίας εργαλείο. Μέσω της ανάλυσης δεδομένων, είναι δυνατή η εξαγωγή έγκαιρων, τεκμηριωμένων συμπερασμάτων, τα οποία στηρίζουν τη λήψη αποφάσεων και την υιοθέτηση προληπτικών μέτρων με στόχο την αποτελεσματική θωράκιση απέναντι σε δυνητικούς κινδύνους (Hull, 2018) .

Πρόσφατες επιστημονικές μελέτες, όπως αυτές που δημοσιεύονται στο πλαίσιο ερευνητικών έργων του Bangor University, τονίζουν ότι ο τομέας της διαχείρισης κινδύνων βρίσκεται σε τροχιά σημαντικού μετασχηματισμού. Συγκεκριμένα, αναδεικνύεται μία στροφή από παραδοσιακές, υποκειμενικές προσεγγίσεις σε πιο αντικειμενικές και εμπεριστατωμένες μεθόδους, που βασίζονται στην αξιοποίηση των Big Data και στις τεχνικές εξόρυξης δεδομένων (data mining). Η τεχνολογική πρόοδος, σε συνδυασμό με τις αναδυόμενες μεθόδους στατιστικής ανάλυσης, επιτρέπει πλέον τη διαχείριση του κινδύνου με τρόπους που μειώνουν τις προσωπικές προκαταλήψεις και ενισχύουν την ακρίβεια των προβλέψεων (Cornwell et al., 2023).

Ένα από τα βασικά πλεονεκτήματα που προσφέρει η ανάλυση δεδομένων στον τομέα της διαχείρισης κινδύνων είναι η δυνατότητα έγκαιρης αναγνώρισης μοτίβων που σχετίζονται με πιθανές απειλές. Μέσω της εφαρμογής εξελιγμένων αλγορίθμων μηχανικής μάθησης και της αξιοποίησης στατιστικών μεθόδων όπως οι καμπύλες ROC (Receiver Operating Characteristic curves) και η επιλογή μεταβλητών πρόβλεψης (predictive variables), οι οργανισμοί μπορούν να αποκτήσουν μια πιο καθαρή και τεκμηριωμένη εικόνα για την πιθανότητα εμφάνισης συγκεκριμένων συμβάντων κινδύνου. Αυτή η γνώση επιτρέπει τη λήψη στοχευμένων και προληπτικών μέτρων, ενισχύοντας ουσιαστικά την ανθεκτικότητα του οργανισμού (Cornwell et al. 2023).

Παράλληλα, η συνεχής υποστήριξη της λήψης αποφάσεων μέσω της εξαγωγής και ανάλυσης κρίσιμων μετρικών μετατρέπει τη διαχείριση κινδύνου από μία αντιδραστική σε μία δυναμική και προληπτική λειτουργία. Ουσιαστικά, δημιουργείται ένα πλαίσιο συνεχούς ανατροφοδότησης (feedback loop), όπου τα δεδομένα όχι μόνο χρησιμοποιούνται για την αποτίμηση υπαρχόντων κινδύνων, αλλά και για την πρόβλεψη μελλοντικών, συμβάλλοντας στη δημιουργία ενός πιο ευέλικτου και προσαρμοστικού οργανισμού (Cornwell et al. 2023).

Η ολοένα και πιο έντονη τάση προς την επιστημονική προσέγγιση της διαχείρισης κινδύνου δεν περιορίζεται αποκλειστικά στις τεχνολογικές εξελίξεις. Αντιθέτως, αντανακλά και μια βαθύτερη πολιτισμική μεταβολή στον τρόπο με τον οποίο οι οργανισμοί κατανοούν και διαχειρίζονται τον κίνδυνο. Η χρήση τεκμηριωμένων δεδομένων επιτρέπει έναν πιο δομημένο και συστηματικό σχεδιασμό πολιτικών ασφαλείας, οδηγεί στην υιοθέτηση πιο ξεκάθαρων και αποδοτικών μηχανισμών ελέγχου και συμβάλλει στην ανάπτυξη στρατηγικών που περιορίζουν τόσο την πιθανότητα εμφάνισης ζημιογόνων γεγονότων όσο και την έντασή τους όταν αυτά συμβούν (Cornwell et al. 2023).

Συμπερασματικά, η χρήση Big Data στην ανάλυση κινδύνων διευκολύνει την ενσωμάτωση εξωτερικών παραγόντων, όπως οι μακροοικονομικές τάσεις, οι αλλαγές στο κανονιστικό πλαίσιο ή τα δεδομένα για φυσικές καταστροφές, προσφέροντας μία ολιστική προσέγγιση. Έτσι, οι οργανισμοί δεν περιορίζονται στην απλή ανασκόπηση του παρελθόντος αλλά προετοιμάζονται ενεργά για μελλοντικές ανατροπές.

2.3 Τεχνικές Αξιολόγησης Πιστωτικού Κινδύνου μέσω Ανάλυσης Δεδομένων

Στη σύγχρονη τραπεζική πρακτική, η αξιολόγηση του πιστωτικού κινδύνου αποτελεί αναπόσπαστο στοιχείο της διαχείρισης χαρτοφυλακίων, καθώς επιτρέπει την ποσοτικοποίηση και τον έλεγχο των ενδεχόμενων ζημιών από μη εξυπηρετούμενα δάνεια. Στο πέρασμα των τελευταίων ετών, η εξέλιξη των τεχνικών αξιολόγησης κινδύνου κινείται από παραδοσιακές στατιστικές προσεγγίσεις σε εξελιγμένα αλγοριθμικά μοντέλα μηχανικής μάθησης, ανταποκρινόμενη στην ανάγκη για μεγαλύτερη ακρίβεια, επεκτασιμότητα και ευελιξία μεταξύ διαφορετικών συστημάτων (Roy & Vasa, 2025).

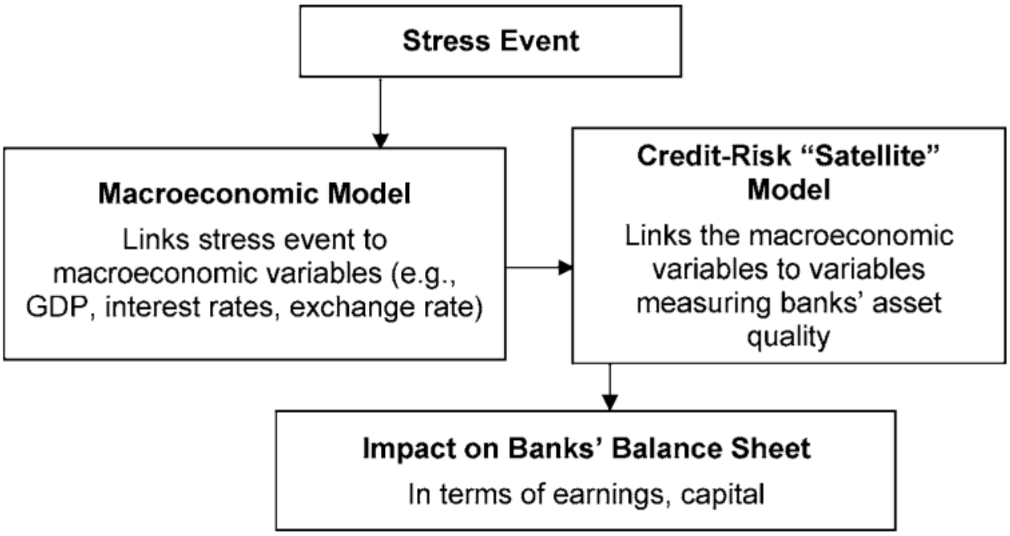
Οι κλασικές στατιστικές τεχνικές, όπως η λογιστική παλινδρόμηση (Logistic Regression) και η Γραμμική Ανάλυση Διακριτότητας (Linear Discriminant Analysis, LDA), αποτέλεσαν για πολλά χρόνια το κύριο εργαλείο των τραπεζών για τον υπολογισμό της πιθανότητας αθέτησης. Συγκεκριμένα, η λογιστική παλινδρόμηση προσφέρει μια απευθείας εκτίμηση των odds ratios (λόγος πιθανοτήτων) των μεταβλητών κινδύνου, ενώ η ανάλυση διακριτότητας, που συνήθως εφαρμόζεται σε μοντέλα τύπου Z-score, επιτρέπει τον διαχωρισμό των δανειοληπτών σε ομάδες υψηλού και χαμηλού κινδύνου. Ωστόσο, με την αύξηση της πολυπλοκότητας των διαθέσιμων δεδομένων, έγινε σαφές ότι αυτές οι μέθοδοι δεν αξιοποιούν πλήρως τις μη γραμμικές αλληλεπιδράσεις μεταξύ μεταβλητών, γεγονός που οδήγησε στην αναζήτηση εναλλακτικών προσεγγίσεων (Chen et al., 2015).

Οι τεχνικές μηχανικής μάθησης έχουν ανταποκριθεί σε αυτή την ανάγκη, εισάγοντας αλγορίθμους που μαθαίνουν προσαρμόσιμους κανόνες ταξινόμησης και μπορούν να αντιμετωπίσουν τόσο την αποτίμηση μη γραμμικών συσχετίσεων όσο και την ανισορροπία κλάσεων. Μοντέλα όπως τα δέντρα αποφάσεων και τα τυχαία δάση προσφέρουν ευελιξία στην ανίχνευση σύνθετων αλληλεπιδράσεων και μειώνουν τον κίνδυνο υπερεφαρμογής (overfitting) μέσω της συμπερίληψης πολλών δέντρων (ensembles) (Lessmann et al., 2015).

Παράλληλα, οι Μηχανές Υποστηρικτικών Διανυσμάτων (SVM) παρέχουν ισχυρά εργαλεία για την εύρεση βέλτιστων ορίων διαχωρισμού σε υψηλής διαστατικότητας χώρους (high-dimensional spaces), ενώ τα βαθιά νευρωνικά δίκτυα (deep neural networks) και οι αλγόριθμοι ενίσχυσης βαθμίδωσης (gradient boosting) όπως XGBoost και LightGBM προσφέρουν υψηλού επιπέδου εκμάθηση μη γραμμικών κατανομών, βελτιώνοντας σημαντικά την πρόβλεψη αθέτησης (Roy & Vasa, 2025).

Για την απόδοση μιας ολιστικής εκτίμησης του συνολικού κινδύνου, υιοθετούνται ποσοτικές μεθοδολογίες όπως η Αξία σε Κίνδυνο (Value at Risk, VaR) και η Υπό Όρους Αξία σε Κίνδυνο (Conditional Value at Risk, CVaR). Ο δείκτης VaR συστηματοποιεί τη μέγιστη αναμενόμενη ζημιά σε ορισμένο χρονικό ορίζοντα και επίπεδο εμπιστοσύνης, ενώ το CvaR, η αναμενόμενη ζημία πέραν του VaR, παρέχει πληροφορίες για το μέγεθος των ακραίων ζημιών, προσφέροντας ένα συνεκτικό (coherent) μέτρο κινδύνου (Rockafellar & Uryasev, 2002). Η τεχνική που χρησιμοποιεί αυτούς τους δείκτες είναι οι Monte Carlo προσομοιώσεις, οι οποίες επιτρέπουν την εκτίμηση της κατανομής ζημιών μέσω επαναλαμβανόμενων τυχαίων δειγματοληψιών, λαμβάνοντας υπόψη την αλληλεξάρτηση μεταξύ των στοιχείων του χαρτοφυλακίου (Chen et al., 2015).

Στον ποιοτικό τομέα, η ανάλυση σεναρίων (scenario analysis) και η ανάλυση ευαισθησίας (sensitivity analysis) λειτουργούν ως εργαλεία ελέγχου αντοχής (stress testing), εξετάζοντας πώς ακραίες μεταβολές μακροοικονομικών παραμέτρων (π.χ. επιτόκια, ποσοστά ανεργίας) επιδρούν στην οικονομική ευρωστία ενός οργανισμού. (Foglia, 2009).

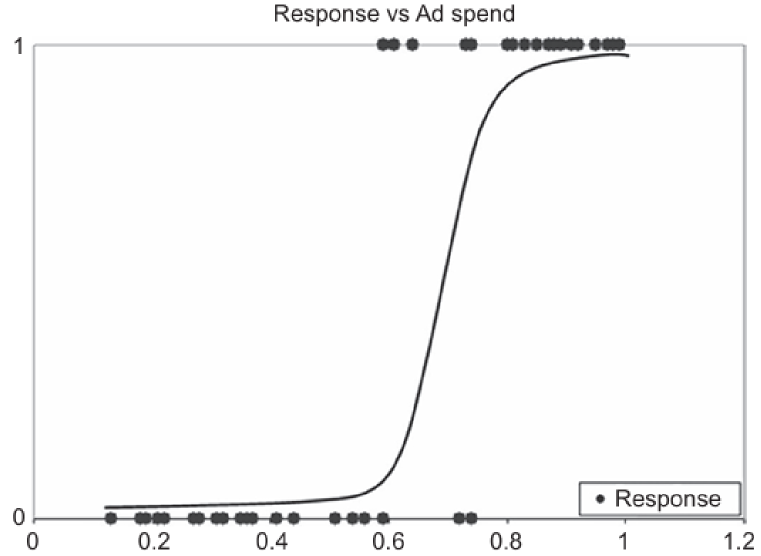


Εικόνα 1 Μια τυπική διαδικασία Macro Stress Testing. Πηγή: Foglia, 2009

Συνοψίζοντας, η σύγχρονη αξιολόγηση πιστωτικού κινδύνου βασίζεται σε ένα πολυεπίπεδο πλέγμα τεχνικών: από παραδοσιακές στατιστικές, μέσω αλγορίθμων μηχανικής μάθησης, μέχρι ποσοτικές και ποιοτικές μεθοδολογίες, διασφαλίζοντας έτσι αρτιότητα, ακρίβεια και ανθεκτικότητα των χρηματοπιστωτικών ιδρυμάτων. Με τα παραπάνω, γίνονται σαφή τα πλεονεκτήματα και οι περιορισμοί τόσο των παραδοσιακών στατιστικών μεθόδων όσο και των πιο εξελιγμένων τεχνικών μηχανικής μάθησης. Στα επόμενα κεφάλαια θα αναλυθεί λεπτομερώς η υλοποίηση και η αξιολόγηση μοντέλων της λογιστικής παλινδρόμησης και των δέντρων αποφάσεων όπου θα αναδειχθούν οι θεωρητικές βάσεις, οι τεχνικές εφαρμογές και τα πλεονεκτήματα καθεμιάς προσέγγισης στο πεδίο του πιστωτικού κινδύνου.

2.3.1 Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση (Logistic Regression) αποτελεί μία από τις πιο θεμελιώδεις και διαδεδομένες στατιστικές τεχνικές για την πρόβλεψη ενός δυαδικού αποτελέσματος, όπου η εξαρτημένη μεταβλητή παίρνει δύο τιμές, όπως για παράδειγμα αθέτηση ή μη αθέτηση ενός δανείου. Σε αντίθεση με τη γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση χρησιμοποιεί τη λογαριθμική συνάρτηση logit, η οποία περιορίζει το αποτέλεσμα μεταξύ 0 και 1, καθιστώντας τη μέθοδο ιδανική για την εκτίμηση πιθανοτήτων (Vidovic & Yue, 2020). Μέσω της εφαρμογής της, είναι δυνατή η εκτίμηση της πιθανότητας εμφάνισης ενός συμβάντος ως συνάρτηση διαφόρων ανεξάρτητων μεταβλητών, οι οποίες μπορεί να είναι δημογραφικά χαρακτηριστικά, χρηματοοικονομικοί δείκτες ή στοιχεία συναλλακτικής συμπεριφοράς.



Εικόνα 2 Λογιστικό Μοντέλο Παλινδρόμησης. Πηγή: Kotu & Deshpande, 2019

Στο χρηματοοικονομικό περιβάλλον, η λογιστική παλινδρόμηση έχει καθιερωθεί ως το βασικό εργαλείο εκτίμησης της Πιθανότητας Αθέτησης (Probability of Default - PD), κυρίως μέσω της κατασκευής μοντέλων πιστοληπτικής αξιολόγησης. Ιστορικά δεδομένα πελατών, όπως εισοδηματικά στοιχεία, ιστορικό πληρωμών και πιστωτική συμπεριφορά, χρησιμοποιούνται ως είσοδοι σε λογιστικά μοντέλα, με σκοπό τη δημιουργία συστημάτων κατάταξης (scorecards) που κατατάσσουν τους δανειολήπτες σε διαφορετικά επίπεδα κινδύνου (Costa-e-Silva & Lopes, 2020).

Ένα από τα κυριότερα πλεονεκτήματα της λογιστικής παλινδρόμησης έγκειται στην απλότητα και την ερμηνευσιμότητα που προσφέρει. Οι συντελεστές που παράγει το μοντέλο επιτρέπουν την άμεση και ποσοτική αποτίμηση της επίδρασης κάθε ανεξάρτητης μεταβλητής στην πιθανότητα εκδήλωσης του υπό εξέταση γεγονότος, όπως είναι η αθέτηση ενός δανείου. Αυτή η διαφάνεια διευκολύνει σημαντικά τη χρήση του μοντέλου από μη ειδικούς και αποτελεί βασικό κριτήριο αποδοχής του από ρυθμιστικές αρχές, δεδομένου ότι επιτρέπει την επαρκή τεκμηρίωση και αιτιολόγηση των αποτελεσμάτων. Παράλληλα, η λογιστική παλινδρόμηση επιτρέπει τη διενέργεια στατιστικών ελέγχων για την αξιολόγηση της σημαντικότητας κάθε εισόδου, γεγονός που καθιστά δυνατή τη σταδιακή βελτιστοποίηση του μοντέλου μέσω της επιλογής μόνο των ουσιαστικών μεταβλητών (Kotu & Deshpande, 2019).

Ένα ακόμα πλεονέκτημα είναι ότι το μοντέλο παρέχει συνεχή αποτελέσματα εκτιμήσεων πιθανοτήτων και όχι απλές δυαδικές ταξινομήσεις. Με τον τρόπο αυτό, οι οργανισμοί μπορούν να προσαρμόσουν δυναμικά τα όρια αποδοχής ή απόρριψης αιτήσεων με βάση το επίπεδο κινδύνου που είναι διατεθειμένοι να αναλάβουν. Η δυνατότητα διαβάθμισης της πιθανότητας αθέτησης μέσω scorecards επιτρέπει την πιο στοχευμένη κατανομή πιστωτικών πόρων και την αποτελεσματικότερη διαχείριση του χαρτοφυλακίου. Επιπλέον, η μέθοδος λειτουργεί ικανοποιητικά ακόμα και σε σχετικά μικρού μεγέθους δείγματα, σε αντίθεση με πιο περίπλοκες τεχνικές μηχανικής μάθησης που συχνά απαιτούν τεράστιους όγκους δεδομένων για να εκπαιδευτούν αποτελεσματικά. (Vidovic & Yue, 2020)

Ωστόσο, παρά τις ανωτέρω θετικές ιδιότητες, η λογιστική παλινδρόμηση εμφανίζει σημαντικούς περιορισμούς που πρέπει να ληφθούν υπόψη. Ένας βασικός περιορισμός αφορά την υπόθεση γραμμικότητας που διέπει τη σχέση ανάμεσα στις ανεξάρτητες μεταβλητές και στα log-odds της εξαρτημένης μεταβλητής. Στην πράξη, πολλές σχέσεις στον χρηματοοικονομικό χώρο είναι έντονα μη γραμμικές ή επηρεάζονται από περίπλοκες αλληλεπιδράσεις, τις οποίες η λογιστική παλινδρόμηση αδυνατεί να καταγράψει χωρίς κατάλληλους μετασχηματισμούς ή τη δημιουργία διαδραστικών όρων. Η αποτυχία αποτύπωσης αυτών των πολυπλοκοτήτων μπορεί να οδηγήσει σε απώλεια πληροφορίας και μειωμένη ακρίβεια προβλέψεων (Vidovic, L., & Yue, L. 2020).

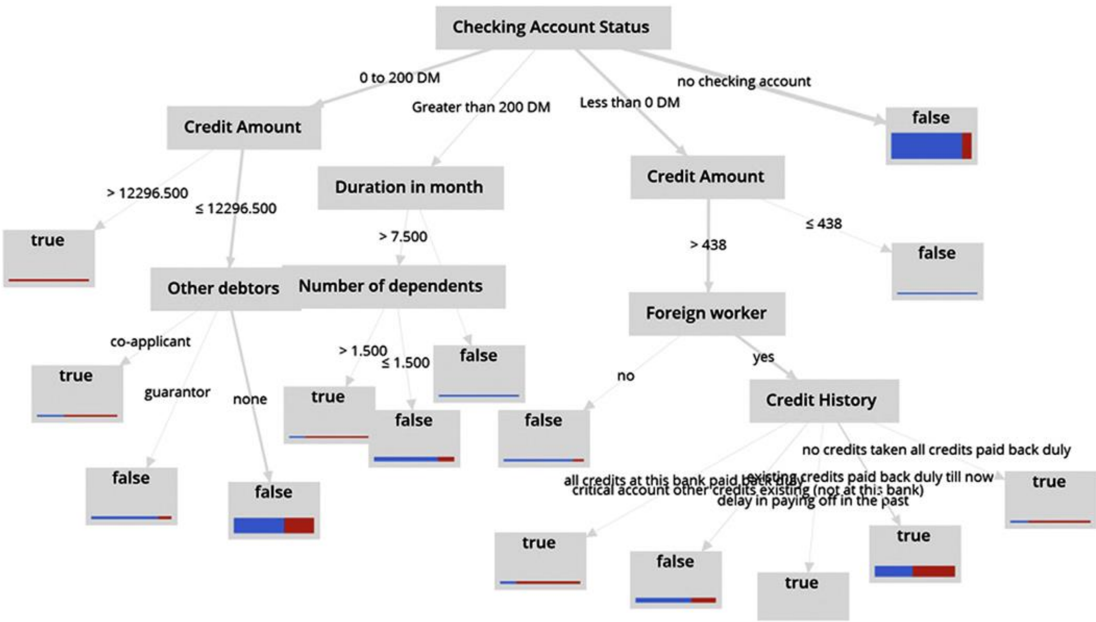
Ένας ακόμα περιορισμός αφορά την προϋπόθεση ανεξαρτησίας των παρατηρήσεων. Η ύπαρξη συσχετισμένων δεδομένων, όπως ομάδες δανειοληπτών που ανήκουν στην ίδια εταιρεία ή οικογένεια, μπορεί να παραβιάσει αυτήν την υπόθεση, οδηγώντας σε υποεκτίμηση της αβεβαιότητας και σε μεροληπτικές εκτιμήσεις των πιθανοτήτων. Η λογιστική παλινδρόμηση είναι επίσης ευάλωτη στην ύπαρξη ισχυρής συσχέτισης μεταξύ των ανεξάρτητων μεταβλητών (multicollinearity), ένα φαινόμενο που μπορεί να καταστήσει τους εκτιμώμενους συντελεστές ασταθείς και να δυσκολέψει την ερμηνεία τους (Kotu & Deshpande, 2019).

Σε σύγκριση με πιο σύγχρονες τεχνικές μηχανικής μάθησης, όπως τα Τυχαία Δάση (Random Forests) ή τα Βαθιά Νευρωνικά Δίκτυα (Deep Neural Networks), η λογιστική παλινδρόμηση ενδέχεται να υστερεί σε προβλήματα υψηλής πολυπλοκότητας και μεγάλης διάστασης. Τα μη παραμετρικά αυτά μοντέλα μπορούν να ανιχνεύσουν περίπλοκες αλληλεπιδράσεις χωρίς την ανάγκη προκαθορισμένων υποθέσεων, ενώ η λογιστική παλινδρόμηση απαιτεί τη ρητή αναπαράσταση τέτοιων σχέσεων μέσω πρόσθετων όρων ή διασταυρούμενων μεταβλητών (Kotu & Deshpande, 2019). Παρόλα αυτά, το γεγονός ότι η λογιστική παλινδρόμηση είναι εύχρηστη, τεκμηριωμένη και ευρέως αποδεκτή, την καθιστά ακόμα ένα από τα πρώτα εργαλεία επιλογής στη φάση ανάπτυξης βασικών μοντέλων αξιολόγησης κινδύνου.

2.3.2 Δέντρα Αποφάσεων

Τα Δέντρα Αποφάσεων (Decision Trees) αποτελούν μία από τις πλέον αναγνωρίσιμες και χρηστικές τεχνικές στον χώρο της επιβλεπόμενης μηχανικής μάθησης, ιδίως όταν ο στόχος είναι η επίλυση προβλημάτων πρόβλεψης και ταξινόμησης με παράλληλη απαίτηση για υψηλή ερμηνευσιμότητα. Η λειτουργία τους βασίζεται στην αναπαράσταση του συνόλου των αποφάσεων υπό τη μορφή ιεραρχικής δομής, που παραπέμπει οπτικά σε δέντρο, από την κορυφή (ρίζα) του οποίου ξεκινά η διαδικασία, διασπώντας το σύνολο των δεδομένων σε υποσύνολα με βάση λογικές συνθήκες (π.χ. τιμές χαρακτηριστικών), μέχρι την κατάληξη στους τελικούς κόμβους (φύλλα), όπου αποδίδεται η πρόβλεψη ή η κατηγορία (Kotu & Deshpande, 2019).

Κάθε εσωτερικός κόμβος αναπαριστά έναν έλεγχο σε μια μεταβλητή (π.χ. "το εισόδημα > 12296.5 ευρώ;"), ενώ κάθε διακλάδωση αντιστοιχεί σε μία από τις δυνατές απαντήσεις του ερωτήματος. Η διαδικασία συνεχίζεται αναδρομικά, παράγοντας ένα μονοπάτι κανόνων που οδηγεί από την ρίζα προς ένα φύλλο, το οποίο αντιστοιχεί στο τελικό συμπέρασμα. Η απλότητα αυτής της δομής προσδίδει στον αλγόριθμο εξαιρετική διαισθητικότητα. Οι κανόνες που παράγονται είναι απόλυτα κατανοητοί και μπορούν εύκολα να μεταφραστούν σε «αν–τότε» συνθήκες, παρέχοντας στον χρήστη μια ξεκάθαρη λογική αλληλουχία που οδηγεί στην τελική απόφαση (Sen et al, 2022).



Εικόνα 3 Οπτική Αναπαράσταση του Δέντρου Αποφάσεων. Πηγή: Kotu & Deshpande, 2019

Στον χρηματοοικονομικό τομέα, τα Δέντρα Αποφάσεων έχουν βρει ιδιαίτερα ευρεία εφαρμογή. Χρησιμοποιούνται σε συστήματα πρόβλεψης πιστωτικού κινδύνου, ανάλυσης αφερεγγυότητας επιχειρήσεων, πρόβλεψης χρεοκοπίας, ανίχνευσης απάτης, καθώς και στην ανάπτυξη μηχανισμών πρώιμης προειδοποίησης για επικείμενες κρίσεις. Λόγω της ικανότητάς τους να εντοπίζουν πολύπλοκες αλληλεπιδράσεις μεταξύ μεταβλητών χωρίς την ανάγκη προκαθορισμένων υποθέσεων γραμμικότητας ή ανεξαρτησίας, μπορούν να αποκαλύψουν μοτίβα που δεν θα ήταν εμφανή με πιο παραδοσιακές στατιστικές μεθόδους. Ένα χαρακτηριστικό παράδειγμα είναι η χρήση των δέντρων για την πρόβλεψη πιθανότητας αθέτησης δανείου με βάση ένα συνδυασμό ποιοτικών και ποσοτικών μεταβλητών, όπως δημογραφικά στοιχεία, εισοδηματικά επίπεδα, και συμπεριφορά πληρωμών (Sen et al 2022).

Ένα σημαντικό πλεονέκτημα των δέντρων αποφάσεων είναι η φυσική τους ικανότητα να χειρίζονται τόσο κατηγορικές όσο και συνεχείς μεταβλητές, χωρίς να απαιτείται ιδιαίτερη μετατροπή ή κανονικοποίηση των δεδομένων. Επιπλέον, είναι σε θέση να ενσωματώνουν αλληλεπιδράσεις μεταξύ μεταβλητών αυτόματα, χωρίς την ανάγκη χειροκίνητης δημιουργίας διασταυρούμενων όρων, κάτι που τα καθιστά ιδιαίτερα ευέλικτα και αποδοτικά σε εφαρμογές με μεγάλη ετερογένεια δεδομένων. Επίσης, λόγω της οπτικής τους φύσης, αποτελούν εξαιρετικά εργαλεία παρουσίασης και επικοινωνίας αποτελεσμάτων, ακόμα και σε μη τεχνικά κοινά, όπως στελέχη επιχειρήσεων ή ρυθμιστικές αρχές (Kotu & Deshpande, 2019).

Ωστόσο, τα δέντρα αποφάσεων δεν στερούνται περιορισμών. Μία από τις πιο συχνές προκλήσεις που αντιμετωπίζουν είναι η τάση υπερπροσαρμογής στα δεδομένα εκπαίδευσης, το λεγόμενο overfitting. Καθώς το δέντρο αναπτύσσεται, είναι πιθανό να αναπαριστά με ακρίβεια ακόμη και θόρυβο ή ασυνήθιστα μοτίβα του δείγματος, γεγονός που οδηγεί σε μειωμένη ικανότητα γενίκευσης όταν εφαρμόζεται σε νέα, άγνωστα δεδομένα. Αυτό μπορεί να αποβεί κρίσιμο στην πρόβλεψη χρηματοοικονομικού κινδύνου, όπου η ακρίβεια σε ανεξερεύνητα σενάρια έχει καθοριστική σημασία. Για την αντιμετώπιση του φαινομένου αυτού, εφαρμόζονται στρατηγικές όπως το «κλάδεμα» (pruning), που περιορίζει το μέγεθος και την πολυπλοκότητα του δέντρου, εξαλείφοντας διακλαδώσεις που δεν συμβάλλουν ουσιαστικά στην απόδοση. Επιπλέον, αναπτύχθηκαν ensemble μέθοδοι, όπως τα Random Forests και το Gradient Boosting, που συνδυάζουν πολλαπλά δέντρα για να ενισχύσουν τη σταθερότητα και την ακρίβεια του συνολικού συστήματος (Vidovic & Yue, 2020).

Ένα ακόμα ζήτημα που έχει εντοπιστεί στη χρήση των δέντρων είναι η φύση των εξόδων που παράγουν. Συχνά, το τελικό αποτέλεσμα των προβλέψεων τείνει να είναι «σκληρό» ή απόλυτο, για παράδειγμα, πιθανότητα αθέτησης 0% ή 100%, κάτι που ενδέχεται να μην ανταποκρίνεται ρεαλιστικά στη φύση του κινδύνου, η οποία χαρακτηρίζεται συχνότερα από διαβαθμίσεις και αβεβαιότητα. Αυτό περιορίζει τη χρηστικότητα των δέντρων σε περιβάλλοντα όπου απαιτούνται πιο λεπτομερείς και συνεχείς εκτιμήσεις πιθανοτήτων, όπως συμβαίνει στη λογιστική παλινδρόμηση, η οποία παρέχει ένα πιο ομαλό αποτέλεσμα (Vidovic & Yue, 2020).

Παρά τα παραπάνω μειονεκτήματα, τα δέντρα αποφάσεων εξακολουθούν να αποτελούν βασικό λίθο της αναλυτικής στρατηγικής σε περιβάλλοντα όπου η κατανόηση της απόφασης είναι εξίσου σημαντική με την ακρίβεια της. Η διαφάνεια, η ευκολία στην εφαρμογή, η ευελιξία στην επεξεργασία δεδομένων και η ικανότητά τους να παράγουν σαφείς επιχειρησιακούς κανόνες, τα καθιστούν απαραίτητα εργαλεία για οργανισμούς που στοχεύουν σε γρήγορη και κατανοητή υποστήριξη αποφάσεων. Ο συνδυασμός τους με πιο σύνθετες μεθόδους μπορεί να οδηγήσει σε υβριδικές προσεγγίσεις ιδιαίτερης ισχύος, ειδικά στο πεδίο της χρηματοοικονομικής ανάλυσης κινδύνου, όπου απαιτείται ισορροπία μεταξύ προβλεπτικής ισχύος και λογοδοσίας.

2.4 Σύνδεση Θεωρίας και Πρακτικής στις Επιχειρηματικές Αποφάσεις

Η εφαρμογή των ανωτέρω τεχνικών ανάλυσης δεδομένων έχει καταστεί θεμέλιο στη σύγχρονη διαχείριση χρηματοοικονομικού κινδύνου, συνδέοντας άρρηκτα τη θεωρητική γνώση με την πρακτική λήψη επιχειρηματικών αποφάσεων. Οι μέθοδοι της λογιστικής παλινδρόμησης και τα δέντρα αποφάσεων δεν αποτελούν πλέον μόνο στατιστικά εργαλεία, αλλά κρίσιμα στρατηγικά μέσα που επηρεάζουν τη βιωσιμότητα και την ανάπτυξη οργανισμών σε ένα διαρκώς μεταβαλλόμενο και απρόβλεπτο οικονομικό περιβάλλον.

Η λογιστική παλινδρόμηση, με την ικανότητά της να παρέχει διαφανείς και ερμηνεύσιμες εκτιμήσεις πιθανοτήτων, χρησιμοποιείται εκτεταμένα σε πιστωτικά ιδρύματα για την αξιολόγηση της αξιόχρεας δανειοληπτών. Η πρακτική αυτή όχι μόνο βελτιώνει τη συνοχή των αποφάσεων δανεισμού, αλλά επιτρέπει και την εύκολη συμμόρφωση με ρυθμιστικές απαιτήσεις (Costa-e-Silva & Lopes, 2020).

Παράλληλα, τα δέντρα αποφάσεων, χάρη στην οπτική και λογική τους δομή, αποτελούν ισχυρό εργαλείο για την έγκαιρη ανίχνευση κινδύνων σε οργανισμούς διαφόρων μεγεθών (Sen et al, 2022). Η ικανότητά τους να αναλύουν και να απεικονίζουν διακριτά σενάρια και να παράγουν κανόνες τύπου «εάν-τότε» τα καθιστά ιδανικά για την ανάπτυξη συστημάτων έγκαιρης προειδοποίησης. Μελέτες έχουν δείξει ότι η εφαρμογή δέντρων αποφάσεων σε μικρομεσαίες επιχειρήσεις συμβάλλει ουσιαστικά στην ταχύτερη αναγνώριση οικονομικών δυσχερειών και στη βελτίωση της διαχείρισης ρευστότητας και κεφαλαίων (Jun & Liu, 2022).

Η θεωρητική έρευνα αναδεικνύει επίσης το γεγονός ότι η επιλογή της κατάλληλης τεχνικής εξαρτάται από τις προτεραιότητες του εκάστοτε οργανισμού. Όταν ο βασικός στόχος είναι η αξιόπιστη ανίχνευση των πιο επικίνδυνων πελατών και η μείωση του σφάλματος τύπου I (false negative), τα δέντρα αποφάσεων προσφέρουν σημαντικά πλεονεκτήματα λόγω της ικανότητάς τους να συλλαμβάνουν πολύπλοκες σχέσεις μεταξύ μεταβλητών (Vidovic & Yue, 2020). Αντιθέτως, όταν ζητούμενο είναι η ομαλή διαβάθμιση του επιπέδου κινδύνου και η εύκολη ερμηνεία των αποτελεσμάτων, η λογιστική παλινδρόμηση προτιμάται ως πιο κατάλληλη μέθοδος.

Επιπλέον, η συνολική εφαρμογή ανάλυσης δεδομένων στο risk management έχει αποδειχθεί ότι μειώνει τη συχνότητα και τη σοβαρότητα των απωλειών, ενισχύοντας τη λήψη στρατηγικών αποφάσεων σε επιχειρήσεις (Cornwell et al. 2023).

Συνοψίζοντας, η θεωρία των τεχνικών αξιολόγησης κινδύνου συνδέεται στενά με την πρακτική διαχείριση επενδύσεων, δανείων και εσωτερικών ελέγχων σε επιχειρήσεις, παρέχοντας εργαλείο λήψης αποφάσεων που βασίζεται σε δεδομένα υψηλής εγκυρότητας.

ΚΕΦΑΛΑΙΟ 3: ΜΕΘΟΔΟΛΟΓΙΑ

3.1 Σχεδίαση της Έρευνας

Η παρούσα μελέτη στοχεύει στη συγκριτική αξιολόγηση δύο δημοφιλών τεχνικών ταξινόμησης στη διαχείριση χρηματοοικονομικού κινδύνου: της Λογιστικής Παλινδρόμησης (Logistic Regression) και των Δέντρων Αποφάσεων (Decision Trees). Η επιλογή αυτών των μεθόδων βασίζεται στην ευρεία εφαρμογή τους στον τομέα της πιστοληπτικής αξιολόγησης και στην ικανότητά τους να παρέχουν ερμηνεύσιμα αποτελέσματα, τα οποία είναι κρίσιμα για τη συμμόρφωση με κανονιστικά πλαίσια και τη λήψη επιχειρηματικών αποφάσεων (Rudd & Priestley, 2017).

Το πρώτο στάδιο αφορά τη συλλογή και την προετοιμασία του συνόλου δεδομένων, το οποίο προέρχεται από τη διαδικτυακή πλατφόρμα Kaggle και αφορά πραγματικές περιπτώσεις πιστοληπτικής αξιολόγησης. Το dataset περιλαμβάνει ποικιλία χαρακτηριστικών που έχουν αποδειχθεί σημαντικά στην πρόβλεψη πιστωτικού κινδύνου, όπως η ηλικία του δανειολήπτη, το επίπεδο εισοδήματός του, το αιτούμενο ποσό του δανείου, καθώς και στοιχεία από το ιστορικό πληρωμών. Στο πλαίσιο της προετοιμασίας των δεδομένων εφαρμόζονται διαδικασίες καθαρισμού, χειρισμού ελλιπών τιμών, μετασχηματισμού μεταβλητών και κανονικοποίησης, σύμφωνα με τις βέλτιστες πρακτικές στον τομέα της εφαρμοσμένης μηχανικής μάθησης για πιστωτική ανάλυση (Koc et al., 2023). Στόχος του σταδίου αυτού είναι η δημιουργία ενός αξιόπιστου και συνεκτικού συνόλου δεδομένων, το οποίο να επιτρέπει την ανάπτυξη στατιστικά έγκυρων και προβλεπτικά ισχυρών μοντέλων.

Στη συνέχεια, περνάμε στο στάδιο ανάπτυξης των μοντέλων. Ειδικότερα, επιλέγονται δύο διαφορετικές τεχνικές, με συμπληρωματικά χαρακτηριστικά: η λογιστική παλινδρόμηση και τα δέντρα αποφάσεων. Η λογιστική παλινδρόμηση (logistic regression) προσφέρει ένα γραμμικό μοντέλο πρόβλεψης, το οποίο επιτρέπει την εκτίμηση της πιθανότητας αθέτησης ενός πελάτη με βάση τις εισόδους του, με σημαντικό πλεονέκτημα την υψηλή ερμηνευσιμότητα των συντελεστών της. Αντιθέτως, τα δέντρα αποφάσεων (decision trees) λειτουργούν σε μη γραμμικό πλαίσιο και έχουν τη δυνατότητα να ανιχνεύουν αλληλεπιδράσεις μεταξύ μεταβλητών χωρίς να απαιτείται προεπιλογή σχέσεων. Η οπτική αναπαράσταση των αποφάσεων μέσω των διακλαδώσεων του δέντρου επιτρέπει την αποτύπωση κανόνων τύπου «εάν-τότε», ενισχύοντας έτσι την κατανόηση των αποτελεσμάτων σε λειτουργικό επίπεδο.

Στο τρίτο στάδιο της διαδικασίας πραγματοποιείται η αξιολόγηση των μοντέλων με βάση ποσοτικές μετρικές απόδοσης. Για να επιτευχθεί μια πολυδιάστατη αποτίμηση, χρησιμοποιούνται δείκτες όπως η ακρίβεια (accuracy), η ευαισθησία (recall), η ειδικότητα (precision), η συνδυαστική μέτρηση F1-score, καθώς και η περιοχή κάτω από την καμπύλη ROC (Area Under the Curve AUC). Οι μετρικές αυτές επιτρέπουν την κατανόηση τόσο της ικανότητας γενικής πρόβλεψης όσο και της απόδοσης των μοντέλων στις περιπτώσεις ακραίου κινδύνου, όπου απαιτείται αυξημένη προσοχή σε σφάλματα τύπου I και II. Επιπλέον, στο πλαίσιο της κανονιστικής συμμόρφωσης και της πρακτικής χρησιμότητας, δίνεται ιδιαίτερη έμφαση στην ερμηνευσιμότητα των αποτελεσμάτων, ειδικά στον χρηματοπιστωτικό τομέα όπου η αιτιολόγηση της αποφάσεων έχει καθοριστικό ρόλο (Biecek et al., 2021).

Τέλος, ακολουθεί η σύγκριση και ερμηνεία των αποτελεσμάτων μεταξύ των δύο μοντέλων. Η σύγκριση γίνεται τόσο σε επίπεδο απόδοσης πρόβλεψης όσο και σε επίπεδο χρηστικότητας του μοντέλου, με στόχο να εντοπιστεί ποια τεχνική προσφέρει καλύτερη ισορροπία ανάμεσα στην προβλεπτική ικανότητα και την επιχειρησιακή εφαρμογή. Σε περιβάλλοντα όπως η πιστοληπτική αξιολόγηση, η επιλογή της μεθόδου δεν καθορίζεται μόνο από την ακρίβεια πρόβλεψης αλλά και από την ικανότητά της να μεταφράζει το αποτέλεσμα σε μια απόφαση που μπορεί να υποστηριχθεί, να ελεγχθεί και να αιτιολογηθεί.

Η επιλογή αυτών των δύο τεχνικών δεν είναι τυχαία, αλλά βασίζεται στην ανάγκη των σύγχρονων οργανισμών: αφενός για ακριβείς προβλέψεις και αφετέρου για διαφάνεια, λογοδοσία και κανονιστική συμμόρφωση (Biecek et al., 2021). Η λογιστική παλινδρόμηση παρέχει ένα στιβαρό στατιστικό υπόβαθρο και εύκολη τεκμηρίωση, ενώ τα δέντρα αποφάσεων προσφέρουν ενισχυμένη ικανότητα αποτύπωσης μη γραμμικών συσχετίσεων και δομημένη οπτικοποίηση της διαδικασίας λήψης αποφάσεων. Μέσω αυτής της συγκριτικής προσέγγισης, επιχειρείται να δοθεί μια ολοκληρωμένη εικόνα για το πώς η επιστήμη των δεδομένων μπορεί να ενισχύσει ουσιαστικά τις πρακτικές διαχείρισης πιστωτικού κινδύνου στις σύγχρονες επιχειρήσεις.

3.2 Δεδομένα και Πηγή Δεδομένων

Η παρούσα μελέτη αξιοποιεί ένα εκτεταμένο και αντιπροσωπευτικό σύνολο δεδομένων (dataset) που προέρχεται από την πλατφόρμα Kaggle, με τίτλο «Loan Default Prediction». Το dataset περιέχει συνολικά 255.347 εγγραφές, οι οποίες χαρακτηρίζονται από 18 μεταβλητές που καλύπτουν δημογραφικά και χρηματοοικονομικά στοιχεία των δανειοληπτών. Κεντρικός σκοπός της αξιοποίησης του dataset είναι η ανάπτυξη προβλεπτικών μοντέλων που μπορούν να καθορίσουν με υψηλή αξιοπιστία αν ένα δάνειο θα καταλήξει σε default (μη αποπληρωμή) ή θα εξυπηρετηθεί κανονικά.

Η επιλογή του συγκεκριμένου dataset βασίστηκε στην καταλληλόλητά του για την ανάλυση πιστωτικού κινδύνου. Πιο συγκεκριμένα, η μεγάλη κλίμακα των δεδομένων, ο ικανοποιητικός αριθμός παρατηρήσεων και η ποικιλία σχετικών μεταβλητών δημιουργούν τις κατάλληλες συνθήκες για την αποτελεσματική εφαρμογή και λεπτομερή αξιολόγηση σύγχρονων μοντέλων μηχανικής μάθησης. Επιπλέον, η ποικιλομορφία των δεδομένων επιτρέπει μια εις βάθος κατανόηση των παραγόντων που επηρεάζουν τον πιστωτικό κίνδυνο, προσφέροντας έτσι σημαντική πληροφορία για την υποστήριξη αποφάσεων και την αποτελεσματική διαχείριση του πιστωτικού κινδύνου από χρηματοπιστωτικά ιδρύματα και άλλους οργανισμούς που παρέχουν δάνεια. Ακολουθεί ο σύνδεσμος για το dataset:

https://www.kaggle.com/datasets/nikhil1e9/loan-default

3.3 Προετοιμασία Δεδομένων

Η προετοιμασία των δεδομένων αποτελεί το πιο λεπτομερές και χρονοβόρο στάδιο της μεθοδολογίας, καθώς καθορίζει την ποιότητα και την αξιοπιστία των εισόδων για την εκπαίδευση των μοντέλων. Ο κώδικας υλοποιήθηκε σε περιβάλλον Python, με χρήση βιβλιοθηκών όπως pandas, NumPy, scikit-learn και imblearn. Στην συνέχεια παρουσιάζονται τα επιμέρους βήματα αναλυτικά.

3.3.1 Εξερευνητική Ανάλυση των Δεδομένων (Exploratory Data Analysis)

Η Εξερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis – EDA) είναι ένα κρίσιμο αρχικό στάδιο στην ανάλυση δεδομένων. Στόχος της EDA είναι να εξετάσουμε τα δεδομένα για την κατανομή τους, να εντοπίσουμε τυχόν ανωμαλίες ή ακραίες τιμές, και γενικά να αξιολογήσουμε την ποιότητά τους πριν προχωρήσουμε σε μοντελοποίηση. Η EDA μάς δίνει μια πρώτη εικόνα των χαρακτηριστικών και πιθανών προβλημάτων ενός dataset, βοηθώντας στη διατύπωση υποθέσεων για περαιτέρω έρευνα. (Komorowski et al., 2016)

Στην παρούσα μελέτη, εφαρμόστηκαν προσεκτικά επιλεγμένες τεχνικές ανάλυσης δεδομένων, με έμφαση τόσο στην ακρίβεια όσο και στην ερμηνευσιμότητα των αποτελεσμάτων. Παρακάτω περιγράφονται βασικές τεχνικές και βήματα της EDA, όπως τεκμηριώνονται σε βιβλιογραφία και εγχειρίδια, μαζί με τη σημασία τους στο στάδιο της προεπεξεργασίας και κατανόησης δεδομένων πριν από οποιαδήποτε μοντελοποίηση ή στατιστικό έλεγχο. Ο πλήρης κώδικας βρίσκεται στο παράρτημα Α.

1. Επισκόπηση πρώτων γραμμών & έλεγχος τύπων δεδομένων: Ένα πρώτο βήμα είναι να δούμε μερικές γραμμές του dataset με την εντολή df.head() σε Python ώστε να επαληθεύσουμε τη δομή και τους τύπους δεδομένων σε κάθε στήλη. Η προβολή των πρώτων εγγραφών βοηθά να διαπιστώσουμε αν τα δεδομένα φορτώθηκαν σωστά και αν οι στήλες έχουν τον αναμενόμενο τύπο (αριθμητικό, κατηγορικό κ.λπ.). Είναι σημαντικό να γνωρίζουμε τον τύπο κάθε μεταβλητής, διότι τυχόν ασυμφωνία (π.χ. αριθμοί αποθηκευμένοι ως συμβολοσειρές) μπορεί να επηρεάσει αρνητικά την ανάλυση ή να οδηγήσει σε λάθη στους υπολογισμούς. Με την επισκόπηση αυτή, εντοπίζουμε άμεσα εμφανή σφάλματα στα δεδομένα και παίρνουμε μια αρχική αίσθηση του περιεχομένου.

2. Στατιστική περίληψη αριθμητικών μεταβλητών: γίνεται με την εντολή df.describe(). Για κάθε αριθμητική μεταβλητή, υπολογίζεται μια περιγραφική σύνοψη βασικών στατιστικών δεικτών όπως το πλήθος τιμών (count), ο μέσος

όρος, η τυπική απόκλιση, το ελάχιστο και μέγιστο, καθώς και τα τεταρτημοριακά όρια (Q1 – 25%, διάμεση τιμή – 50%, Q3 – 75%). Αυτή η συνοπτική στατιστική εικόνα παρέχει πληροφορίες για την κεντρική τάση (mean/median) και τη διασπορά (std, εύρος τιμών, IQR) κάθε μεταβλητής. Οι δείκτες αυτοί βοηθούν να εντοπιστούν αρχικά θέματα όπως πιθανές ακραίες τιμές που επηρεάζουν τον μέσο όρο και αυξάνουν την τυπική απόκλιση ή μεταβλητές με χαμηλή διακύμανση. Χαρακτηριστικά όπως ο μέσος, η διασπορά, τα τεταρτημόρια, η κύρτωση και η ασυμμετρία αποτελούν βασικές ποσοτικές περιγραφές της κατανομής μιας μεταβλητής και βοηθούν τον αναλυτή να συνοψίσει τις ιδιότητες της κατανομής της (Komorowski et al., 2016).

3. Έλεγχος ελλείπουσας πληροφορίας (Missing Values): Η διερεύνηση των ελλειπουσών τιμών σε κάθε χαρακτηριστικό είναι ουσιώδης, καθώς η απουσία δεδομένων μπορεί να εισάγει μεροληψία και να μειώσει τη στατιστική ισχύ των αποτελεσμάτων. Με την συνάρτηση df.isnull().sum() μπορούμε να υπολογίσουμε τον αριθμό των missing values ανά στήλη. Η ύπαρξη πολλών κενών τιμών σε ένα dataset μειώνει το διαθέσιμο δείγμα προς ανάλυση, υπονομεύοντας την αξιοπιστία των συμπερασμάτων. Επιπλέον, αν οι ελλείπουσες τιμές δεν είναι τυχαίες, μπορεί να εισάγουν σημαντική μεροληψία στα αποτελέσματα. Για τον λόγο αυτό, οι ελλείπουσες τιμές όσο και οι outliers πρέπει να αντιμετωπιστούν κατάλληλα πριν την ανάλυση, διαφορετικά τα αποτελέσματα μπορεί να είναι παραπλανητικά (Kwak & Kim, 2017).

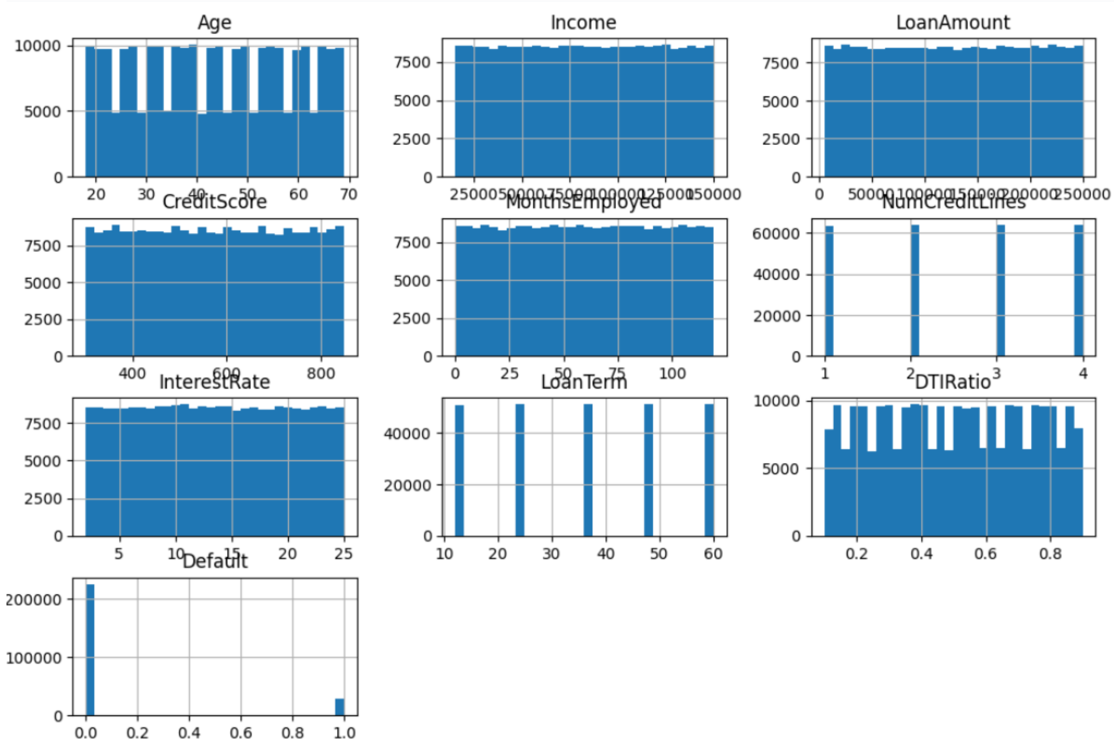
4. Ανάλυση κατηγορικών μεταβλητών (συχνότητες κατηγοριών): Για κατηγορικές μεταβλητές, η EDA περιλαμβάνει την καταμέτρηση της συχνότητας εμφάνισης κάθε κατηγορίας. Με εντολές όπως df['στήλη'].value_counts() μπορούμε να δημιουργήσουμε έναν πίνακα συχνοτήτων που δείχνει πόσες περιπτώσεις αντιστοιχούν σε κάθε κατηγορία, συχνά και σε ποσοστιαία μορφή. Αυτό αντιστοιχεί στην πίνακαλογράφηση των κατηγοριών, όπου καταγράφεται τόσο ο απόλυτος αριθμός όσο και το σχετικό ποσοστό κάθε κατηγορίας. Μια τέτοια ανάλυση βοηθά να εντοπιστούν ανισορροπίες στις κατηγορίες (π.χ. μια κατηγορία που κυριαρχεί συντριπτικά, ή κατηγορίες με ελάχιστες εμφανίσεις) (Komorowski et al., 2016). Στον παρακάτω πίνακα φαίνονται τα κατηγορικά δεδομένα του dataset.

Πίνακας 1 Κατανομή κατηγορικών μεταβλητών

Variable	Category	Count
LoanID	Unique IDs	255347
Education	Bachelor's	64366
	High School	63903
	Master's	63541
	PhD	63537
EmploymentType	Part-time	64161

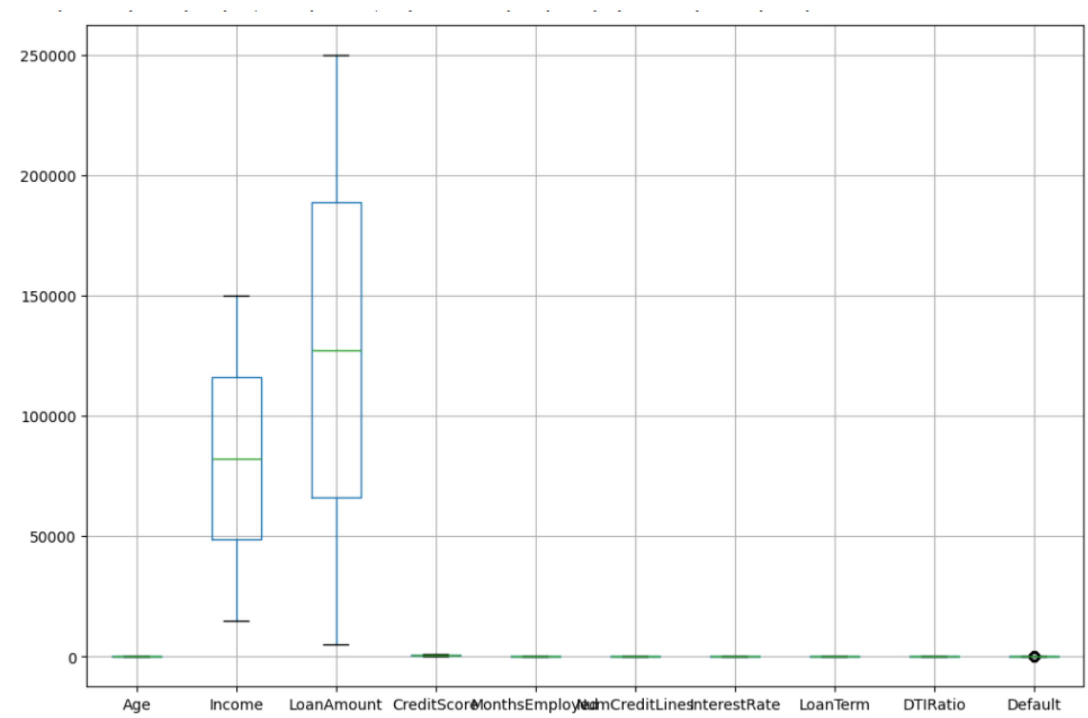
	Unemployed	63824
	Self-employed	63706
	Full-time	63656
MaritalStatus	Married	85302
	Divorced	85033
	Single	85012
HasMortgage	Yes	127677
	No	127670
HasDependents	Yes	127742
	No	127605
LoanPurpose	Business	51298
	Home	51286
	Education	51005
	Other	50914
	Auto	50844
HasCoSigner	Yes	127701
	No	127646

5. Οπτικοποίηση κατανομών αριθμητικών μεταβλητών (ιστογράμματα): Τα ιστογράμματα αποτελούν ένα από τα πιο χρήσιμα γραφικά εργαλεία της EDA για συνεχή (αριθμητικά) δεδομένα. Ένα ιστόγραμμα δείχνει το πλήθος ή την αναλογία των παρατηρήσεων που εμπίπτουν σε διαδοχικά διαστήματα τιμών (bins), δίνοντας μια άμεση εικόνα του σχήματος της κατανομής της μεταβλητής. Μέσω των ιστογραμμάτων μπορούμε να αντιληφθούμε την κεντρική τάση και τη διασπορά (το εύρος) των τιμών, καθώς και χαρακτηριστικά όπως η συμμετρία ή ασυμμετρία (λοξότητα) και η ύπαρξη πολλαπλών κορυφώσεων (πολυτροπικότητα). Τα ιστογράμματα θεωρούνται θεμέλιο της EDA για συνεχή δεδομένα, διότι οπτικοποιούν που συγκεντρώνεται η πλειονότητα των τιμών μεταξύ του ελάχιστου και μέγιστου. Επιπλέον, τα ιστογράμματα μπορούν να αποκαλύψουν ακραίες τιμές ή σφάλματα: π.χ. αν δούμε ένα μικρό πλήθος τιμών απομονωμένο μακριά από τον κύριο όγκο στο διάγραμμα, αυτό σηματοδοτεί πιθανώς outliers. (Komorowski et al., 2016)



Εικόνα 4 Ιστογράμματα για τις αριθμητικές μεταβλητές

6. Ανίχνευση ακραίων τιμών (outliers) με διαγράμματα κουτιού & εύρος Ενδοτεταρτημορίων (IQR): Τα διαγράμματα κουτιού (boxplots) είναι ένα ακόμα πολύτιμο εργαλείο EDA, ιδίως για την ανάδειξη κεντρικής τάσης, διασποράς και πιθανών outliers σε μια απλή εικόνα. Ένα boxplot συνοψίζει μια κατανομή μέσω του μεσαίου 50% (κουτί από Q1 έως Q3), της διαμέσου (γραμμή εντός του κουτιού) και των λεγόμενων "μουστακιών" (whiskers) που εκτείνονται συνήθως έως 1.5 * IQR πέρα από το Q1 και το Q3. Οποιαδήποτε παρατήρηση πέφτει πέρα από 1.5×IQR πάνω από το τρίτο τεταρτημόριο ή κάτω από το πρώτο τεταρτημόριο χαρακτηρίζεται στο γράφημα ως πιθανό outlier. Τα boxplots θεωρούνται εξαιρετική τεχνική EDA επειδή βασίζονται σε στατιστικούς δείκτες, δηλαδή διάμεσος και IQR, που δεν επηρεάζονται εύκολα από ακραίες τιμές (Komorowski et al., 2016). Όπως φαίνεται παρακάτω στην Εικόνα 5, τα δεδομένα μας δεν έχουν ακραίες τιμές και έτσι δεν θα επηρεάσουν δυσανάλογα στατιστικές εκτιμήσεις διαστρεβλώνοντας τα συμπεράσματα.



Εικόνα 5 Boxplot αριθμητικών μεταβλητών

7. Ανάλυση πίνακα συσχέτισης Pearson correlation: Για πολυδιάστατα δεδομένα με πολλαπλές αριθμητικές μεταβλητές, ένα στάδιο της EDA είναι ο υπολογισμός του πίνακα συσχέτισης – συνήθως με τον συντελεστή Pearson για κάθε ζεύγος μεταβλητών. Ο συντελεστής συσχέτισης Pearson (r) ποσοτικοποιεί τον βαθμό γραμμικής σχέσης μεταξύ δύο μεταβλητών, ουσιαστικά δίνοντας μια κανονικοποιημένη εκδοχή της συνδιακύμανσης που κυμαίνεται από -1, δηλαδή ισχυρή αρνητική γραμμική σχέση, έως +1, δηλαδή ισχυρή θετική γραμμική σχέση. Ο υπολογισμός συνδιακύμανσης και συσχέτισης μάς δείχνει πόσο δυο μεταβλητές μεταβάλλονται μαζί και μπορεί να αναδείξει υποκείμενες συνδέσεις (Komorowski et al., 2016). Στον παρακάτω πίνακα φαίνονται οι σχέσεις μεταξύ των αριθμητικών μεταβλητών και του Default. Οι πολύ μικροί συντελεστές υποδηλώνουν ασθενείς γραμμικές σχέσεις με το Default. Ο μεγαλύτερος θετικός συσχετισμός έχει το InterestRate (0.131), δείχνοντας ότι υψηλότερα επιτόκια σχετίζονται με αυξημένο ρίσκο Default.

Πίνακας 2 Pearson correlation μεταξύ Default και αριθμητικών μεταβλητών

Μεταβλητή	**Pearson r**
Age	–0.1678
Income	–0.0991
LoanAmount	0.0867
CreditScore	–0.0342

MonthsEmployed	–0.0974
NumCreditLines	0.0283
InterestRate	0.1313
LoanTerm	0.0005
DTIRatio	0.0192

8. Ανάλυση μεταβλητής-στόχου και αντιμετώπιση ανισορροπίας τάξεων: Τέλος, αν το πρόβλημα που εξετάζουμε είναι εποπτευόμενο (supervised learning), αφιερώνουμε προσοχή στην μεταβλητή-στόχο Default. Για προβλήματα ταξινόμησης (classification), αυτό σημαίνει εξέταση της κατανομής των κλάσεων: πόσες περιπτώσεις ανήκουν σε κάθε κλάση. Συχνά τα πραγματικά δεδομένα είναι μη ισορροπημένα (imbalanced), δηλαδή μία κατηγορία (συνήθως η φυσιολογική) υπερεκπροσωπείται, ενώ η ενδιαφέρουσα κατηγορία (η μη αποπληρωμή δανείου) εμφανίζεται σπάνια. Η ανισορροπία στη μεταβλητή-στόχο μπορεί να δημιουργήσει σοβαρά προβλήματα στην εκπαίδευση μοντέλων, καθώς αυτά τείνουν να μεροληπτούν προς την πολυπληθέστερη κλάση, αγνοώντας τη μειοψηφία (Siguna et al., 2025). Εφόσον η EDA εντοπίσει έντονη ανισορροπία, προδιαγράφεται η ανάγκη για τεχνικές αντιμετώπισης. Μια ευρέως χρησιμοποιούμενη μέθοδος είναι το SMOTE (Synthetic Minority Over-sampling Technique), που προτάθηκε από τους Chawla et al. (2002). Το SMOTE πραγματοποιεί υπερδειγματοληψία της μειοψηφικής κλάσης συνθετικά: δημιουργεί τεχνητά δείγματα της μειοψηφίας, συνδυάζοντας χαρακτηριστικά γειτονικών πραγματικών δειγμάτων, αντί να αντιγράφει απλώς τυχαία υπάρχοντα δείγματα. Με αυτόν τον τρόπο, το σύνολο δεδομένων γίνεται πιο ισορροπημένο χωρίς να εισάγονται απλώς αντίγραφα (που θα μπορούσαν να οδηγήσουν σε overfitting).

Πίνακας 3 Μεταβλητή-στόχος Default

Default	
0 (αποπληρωμή δανείου)	225694
1 (μη αποπληρωμή δανείου)	29653

Ακολουθώντας αυτά τα βήματα, από την πρώτη επισκόπηση και τον έλεγχο ποιότητας των δεδομένων, μέχρι την εξερεύνηση σχέσεων και την αντιμετώπιση ειδικών προβλημάτων όπως η ανισορροπία τάξεων, εξασφαλίζουμε ότι η ανάλυσή μας θα βασίζεται σε έγκυρα και αξιόπιστα δεδομένα, μεγιστοποιώντας την πιθανότητα να εξαγάγουμε σωστά και ουσιαστικά συμπεράσματα.

3.3.2 Προεπεξεργασία Δεδομένων και Μετασχηματισμοί

Παρά την απουσία πραγματικών ελλείψεων στο dataset (όπως διαπιστώθηκε από την κλήση df.isnull().sum()), η μεθοδολογία προβλέπει ολοκληρωμένο pipeline διαχείρισης τυχόν κενών τιμών, ώστε να μπορεί να εφαρμοστεί γενικώς και σε δεδομένα με σημαντικά κενά. Συγκεκριμένα στην συνάρτηση preprocess(df):

1. Αριθμητικές στήλες (όπως LoanAmount, Income, CreditScore, MonthsEmployed): τα κενά συμπληρώνονται με τον μέσο όρο της αντίστοιχης μεταβλητής. Η τεχνική αυτή, γνωστή ως mean imputation, διατηρεί την κεντρική τάση των δεδομένων χωρίς να εισάγει νέες ακραίες τιμές, αλλά έχει την τάση να υποτιμά την πραγματική διασπορά (Zhang, 2016).

2. Κατηγορικές στήλες (όπως Education, LoanPurpose): τα κενά αντικαθίστανται με την επικρατέστερη τιμή (mode imputation), μειώνοντας την πιθανότητα στρέβλωσης από σπάνιες κατηγορίες (Zhang, 2016).

3. Αφαίρεση μη-προγνωστικών στηλών όπως η LoanID που χρησιμεύουν μόνο ως αναγνωριστικά και δεν φέρουν πληροφορία για την πρόβλεψη. Έτσι αποφεύγουμε να φορτώσουμε στο μοντέλο ανεπιθύμητο «θόρυβο» ή δείκτες που δεν συμβάλλουν στον υπολογισμό.

4. Διαχωρισμός χαρακτηριστικών (X) και ετικέτας (y): η στήλη Default θεωρείται μεταβλητή-στόχος, όλες οι υπόλοιπες στήλες συνθέτουν το σύνολο X. Με αυτόν τον τρόπο τα χαρακτηριστικά και η ετικέτα είναι διακριτά αντικείμενα για την εκπαίδευση.

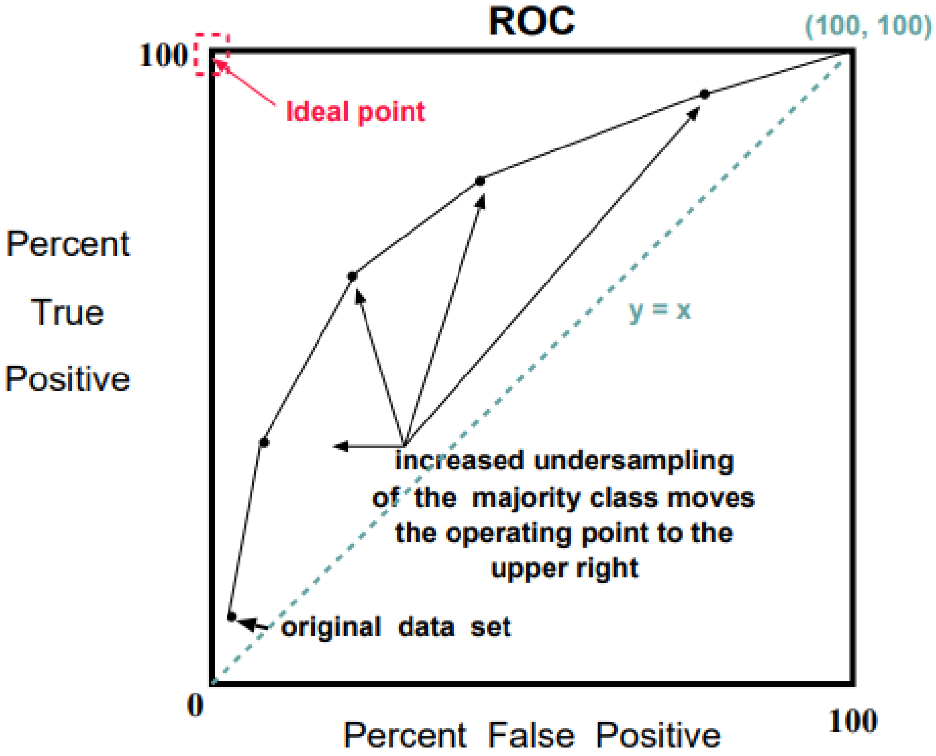
5. Οι κατηγορικές μεταβλητές (όπως είδος επαγγέλματος, κατηγορία πιστωτικής διαβάθμισης) μετατρέπονται σε δυαδικές μεταβλητές (one-hot encoding), ώστε να μπορούν να χρησιμοποιηθούν από αλγορίθμους που απαιτούν αριθμητικά δεδομένα. Η τεχνική αυτή δημιουργεί "dummy" γνωρίσματα που παίρνουν τιμές 0/1 για κάθε κατηγορία μιας ονομαστικής μεταβλητής, επιτρέποντας στο μοντέλο να μάθει ξεχωριστά επιδράσεις για κάθε κατηγορία. (Seger, 2018)

Πριν την εκπαίδευση, η συνάρτηση split_and_scale πραγματοποιεί stratified διαχωρισμό σετ εκπαίδευσης και δοκιμής (80/20) με random_state=42 για επαναληπτότητα, διασφαλίζοντας διατήρηση της αναλογίας των κλάσεων. Στη συνέχεια, εφαρμόζεται τυποποίηση (standardization) των συνεχή αριθμητικών χαρακτηριστικών μέσω του μετασχηματισμού StandardScaler: κάθε χαρακτηριστικό κεντράρεται (μηδενικός μέσος) και κλιμακάρεται σε μοναδιαία τυπική απόκλιση. Αυτή η διαδικασία διασφαλίζει ότι όλα τα χαρακτηριστικά βρίσκονται στην ίδια αριθμητική κλίμακα, κάτι που είναι σημαντικό για μοντέλα όπως η λογιστική παλινδρόμηση (βελτιστοποίηση με βαθμίδωση) και διευκολύνει τη σύγκλιση του αλγορίθμου εκπαίδευσης. Τα δέντρα αποφάσεων είναι δενδροειδείς μοντέλα διαχωρισμού που βασίζονται σε κανονικοποιημένες συνθήκες και θεωρητικά είναι ανθεκτικά στις διαφορές κλίμακας, όμως η κλιμάκωση των χαρακτηριστικών δεν επηρεάζει αρνητικά την απόδοσή τους. (Adeyemo et al.,2018)

Η χρήση αυτού του μηχανισμού εξασφαλίζει ότι ακόμη και σε περιπτώσεις με σημαντικά κενά δεδομένα, η ανάλυση δεν θα υποστεί σοβαρή μεροληψία ή απώλεια πληροφορίας.

3.3.3 Αντιμετώπιση Ανισορροπίας Τάξεων με SMOTE

Όπως αποκάλυψε η Εξερευνητική Ανάλυση Δεδομένων (EDA), υπάρχει έντονη ανισορροπία στην μεταβλητή Default. Η αντιμετώπιση της ανισορροπίας θα γίνει με την μέθοδο του SMOTE (Synthetic Minority Over-sampling Technique) μέσω της συνάρτησης apply_smote(X_train, y_train). Το SMOTE δημιουργεί τεχνητά δείγματα της μειοψηφικής κλάσης, στην περίπτωσή μας το ενδεχόμενο αθέτησης δανείου, μέσω παρεμβολής μεταξύ υπαρκτών δειγμάτων της μειοψηφίας. Συγκεκριμένα, για κάθε δείγμα μειοψηφίας, βρίσκονται οι πλησιέστεροι γείτονές του και παράγονται νέα συνθετικά δείγματα στο διάστημα του χώρου χαρακτηριστικών που ορίζεται από το δείγμα και τους γείτονες. Με αυτόν τον τρόπο, αντί να αντιγραφούν απλώς οι μειοψηφικές περιπτώσεις, που θα μπορούσε να οδηγήσει σε υπερεκμάθηση, overfitting, δημιουργούνται νέες, ποικίλες περιπτώσεις που αυξάνουν την πυκνότητα της μειοψηφικής τάξης στον χώρο των χαρακτηριστικών. Συνδυάζοντας το υπερ-δείγμα μειοψηφίας μέσω SMOTE με ενδεχόμενο υπο-δείγμα της πλειοψηφίας, μπορούμε να πετύχουμε καλύτερη απόδοση του ταξινομητή, ιδιαίτερα ως προς την ικανότητα ανίχνευσης της μειοψηφικής τάξης, σε σχέση με μεθόδους που βασίζονται αποκλειστικά στην υποδειγματοληψία. Ο συνδυασμός συνθετικής υπερδειγματοληψίας της μειοψηφίας και λελογισμένης υποδειγματοληψίας της πλειοψηφίας αποδίδει καλύτερα ως προς την καμπύλη ROC από την απλή υποδειγματοληψία ή την ρύθμιση κοστολογικών παραμέτρων σε άλλους αλγόριθμους (Chawla et al. 2002).



Εικόνα 6 Απεικόνιση καμπύλης ROC μέσω υποδειγματοληψίας. Πηγή: Chawla et al. (2002)

Όπως αναφέρουν οι Chawla et al. (2002), η αυξημένη υποδειγματοληψία της πλειοψηφικής (αρνητικής) κατηγορίας θα μετακινήσει την απόδοση από το κάτω αριστερό σημείο στο πάνω δεξί.

Κατά την εκπαίδευση, εφαρμόζεται SMOTE μόνο στο σύνολο εκπαίδευσης. Δηλαδή, δημιουργούνται συνθετικά δείγματα μειοψηφίας για την εκπαίδευση, ενώ το σύνολο δοκιμής παραμένει με την αρχική κατανομή, ώστε η αξιολόγηση να αντικατοπτρίζει ρεαλιστικές συνθήκες. Είναι σημαντικό να αποφευχθεί η διαρροή πληροφορίας (information leakage): το SMOTE πρέπει να εφαρμόζεται μετά τον διαχωρισμό train-test, ώστε οι συνθετικές παρατηρήσεις να μην αντιγράφουν πληροφορία από τα δεδομένα δοκιμής.

3.4 Μοντέλα Ταξινόμησης: Λογιστική Παλινδρόμηση και Δέντρο Αποφάσεων

Για την πρόβλεψη της αθέτησης δανείου εκπαιδεύονται δύο τύποι μοντέλων: ένα στατιστικό γραμμικό μοντέλο Λογιστικής Παλινδρόμησης και ένα μη-παραμετρικό Δέντρο Αποφάσεων. Κάθε μοντέλο βασίζεται σε διαφορετικές υποθέσεις και μηχανισμούς εκμάθησης, ενώ οι επιλογές παραμετροποίησης και οι ιδιαιτερότητές τους αντανακλούν τον τρόπο με τον οποίο «μαθαίνουν» από τα δεδομένα. Πριν την εκπαίδευση, το σύνολο δεδομένων χωρίζεται σε σύνολο εκπαίδευσης και σύνολο δοκιμής με μια τυπική αναλογία 80/20. Ο διαχωρισμός αυτός εξασφαλίζει ότι η αξιολόγηση της απόδοσης γίνεται σε δεδομένα που το μοντέλο δεν έχει "δει" κατά την εκπαίδευση, δίνοντας μια ρεαλιστική ένδειξη της ικανότητας γενίκευσης. Στην συνέχεια, με τις συναρτήσεις train_logistic(X_train, y_train) και train_tree(X_train, y_train) εκπαιδεύονται τα δύο αυτά μοντέλα.

Η λογιστική παλινδρόμηση είναι ιδανικό μοντέλο για δυαδικά προβλήματα ταξινόμησης. Η βασική υπόθεση του μοντέλου είναι ότι ο λογάριθμος των πιθανοτήτων εμφάνισης του γεγονότος (log-odds της αθέτησης) εκφράζεται ως γραμμικός συνδυασμός των εισαγόμενων χαρακτηριστικών. Δηλαδή, κάθε μεταβλητή συμβάλλει θετικά ή αρνητικά στην πιθανότητα αθέτησης, με το βάρος της να εκτιμάται κατά την εκπαίδευση (Victor & Raheem, 2021).

Επιπλέον, στην υλοποίηση με scikit-learn ενσωματώνεται κανονικοποίηση L2 regularization, δηλαδή προστίθεται ο όρος ποινής $\lambda \sum \omega^2$ στην συνάρτηση κόστους. Αυτό γίνεται για να αποτραπεί η υπερεκμάθηση, ένα φαινόμενο όπου το μοντέλο «προσαρμόζεται» υπερβολικά στις ιδιαιτερότητες του συνόλου εκπαίδευσης, χάνοντας σε γενικευσιμότητα. Η παράμετρος C=1/λ καθορίζει το επίπεδο της ποινής: μικρό C συνεπάγεται ισχυρή ποινή και προτιμάται όταν τα δεδομένα περιέχουν θόρυβο ή πολλές μεταβλητές, ενώ μεγάλο C επιτρέπει στο μοντέλο να προσαρμοστεί περισσότερο (Conroy & Sajda, 2012).

Ο αλγόριθμος εκπαίδευσης χρησιμοποιεί το solver 'lbfgs', ο οποίος είναι αποδοτικός για σχετικά μεγάλα datasets και κατάλληλος για L2-regularization. Η παράμετρος max_iter=1000 διασφαλίζει ότι η διαδικασία μεγιστοποίησης της πιθανοφάνειας έχει αρκετές επαναλήψεις για να συγκλίνει σε βέλτιστη λύση, αποφεύγοντας σφάλματα μη σύγκλισης που συχνά εμφανίζονται σε πιο σύνθετα ή εκτεταμένα σύνολα.

Στην παραγωγή αποτελεσμάτων, το μοντέλο παρέχει τόσο πιθανότητες εκδήλωσης αθέτησης, χρήσιμες για την καμπύλη ROC/AUC, όσο και δυαδικές προβλέψεις (default/no-default) μέσω ενός τυπικού κατωφλίου (threshold) 0.5. Το κατώφλι αυτό, ωστόσο, μπορεί να προσαρμοστεί αναλόγως του κόστους των ψευδών θετικών ή αρνητικών προβλέψεων, ειδικά σε προβλήματα με άνιση σημαντικότητα μεταξύ των κατηγοριών.

Το δέντρο αποφάσεων αποτελεί μια μη-παραμετρική μέθοδο ταξινόμησης που δομεί τη λήψη αποφάσεων ως διαδοχικές ερωτήσεις (διαχωρισμούς) στα χαρακτηριστικά των δειγμάτων. Κάθε εσωτερικός κόμβος του δέντρου αντιστοιχεί σε ένα τεστ πάνω σε μία μεταβλητή, με το αποτέλεσμα της σύγκρισης να καθορίζει την περαιτέρω διαδρομή στο δέντρο. Οι τερματικοί κόμβοι (φύλλα) αντιστοιχούν στην τελική απόφαση του μοντέλου: πρόβλεψη αθέτησης ή μη. Η επιλογή του βέλτιστου χαρακτηριστικού για κάθε διαχωρισμό στηρίζεται σε κάποιο κριτήριο καθαρότητας, με το συνηθέστερο να είναι η εντροπία (entropy) (Lee et al., 2014).

Για να προληφθεί η υπερεκμάθηση, η οποία είναι ιδιαίτερα σύνηθες πρόβλημα στα δέντρα αποφάσεων, υιοθετούνται συγκεκριμένοι περιορισμοί στην πολυπλοκότητα του δέντρου: περιορισμός στο μέγιστο βάθος (max_depth=5) και ελάχιστος αριθμός δειγμάτων ανά εσωτερικό κόμβο (min_samples_split=10). Η επιλογή των τιμών τους μπορεί να γίνει με πειραματισμό: ένα πολύ βαθύ δέντρο μπορεί να υπερπροσαρμοστεί, ενώ ένα υπερβολικά ρηχό δέντρο ίσως δεν συλλαμβάνει όλη την χρήσιμη πληροφορία. Οι παράμετροι αυτοί διασφαλίζουν ότι το δέντρο δεν αναπτύσσεται υπερβολικά, αποφεύγοντας να μοντελοποιήσει το θόρυβο ή σπάνιες περιπτώσεις του συνόλου εκπαίδευσης (Lee et al., 2014).

Η οπτικοποίηση του δέντρου με την συνάρτηση plot_decision_tree_model, ιδιαίτερα των πρώτων επιπέδων, παρέχει ουσιαστική πληροφορία για το ποια χαρακτηριστικά παίζουν σημαντικό ρόλο στις προβλέψεις. Έτσι, τα δέντρα αποφάσεων είναι εκτός από αποτελεσματικά ταξινομητές και ιδιαίτερα χρήσιμα ως εργαλεία ερμηνείας.

3.5 Αξιολόγηση Μοντέλων

Η απόδοση των δύο μοντέλων της λογιστικής παλινδρόμησης και δέντρου αποφάσεων αξιολογείται με διάφορα μετρικά ταξινόμησης, που προσφέρουν συμπληρωματικές οπτικές. Τα κύρια μετρικά που αναφέρονται είναι: Ακρίβεια (Accuracy), Μέτρο F1 (F1-score) και Εμβαδό κάτω από την καμπύλη ROC (ROC-AUC).

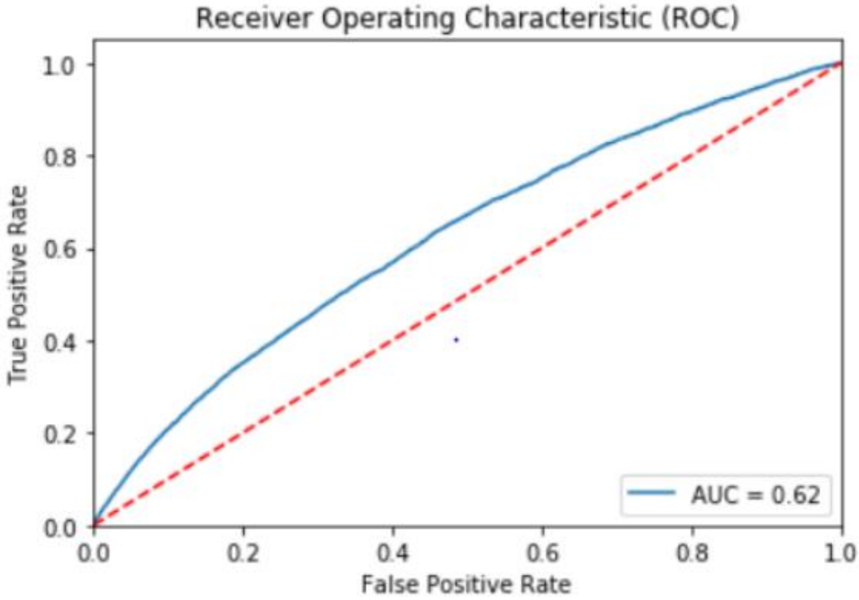
1. Ακρίβεια: Ορίζεται ως το ποσοστό των περιπτώσεων που ταξινομούνται σωστά, δηλαδή $\frac{TP+TN}{TP+TN+FP+FN}$ όπου TP, TN οι σωστές θετικές/αρνητικές προβλέψεις και FP, FN οι λανθασμένες. Η ακρίβεια είναι εύληπτο μέτρο, όμως μπορεί να είναι παραπλανητική σε μη ισορροπημένα δεδομένα. Ένα πολύ υψηλό ποσοστό ακρίβειας δεν εγγυάται ότι το μοντέλο έχει μάθει να ανιχνεύει τις σπάνιες περιπτώσεις αθέτησης. Στο παρόν πρόβλημα, η ακρίβεια από μόνη της θα μπορούσε να ωραιοποιεί την απόδοση προβλέποντας σχεδόν πάντα «καμία αθέτηση». Επομένως, απαιτούνται και άλλα κριτήρια (Kulkarni et al, 2020).

2. Precision και Recall: Αυτά τα δύο μετρικά αποτελούν τη βάση για το F1-score. Το Precision (θετική προγνωστική αξία) είναι το ποσοστό των προβλέψεων αθέτησης που είναι πραγματικά σωστές, δηλαδή πόσο ξεκάθαρες είναι οι θετικές προβλέψεις. Το Recall (ευαισθησία) είναι το ποσοστό των πραγματικών αθετήσεων που το μοντέλο πέτυχε να εντοπίσει, δηλαδή πόσο καλά προβλέπει η ταξινόμηση τις περιπτώσεις αθέτησης. Στόχος στο πλαίσιο πιστωτικού κινδύνου είναι συχνά η μέγιστη ανάκληση, δηλαδή να μην ξεφύγει καμία περίπτωση υψηλού κινδύνου, χωρίς όμως η precision να πέφτει δραματικά διότι αυτό θα σήμαινε πολλές ψευδοσυναγερμούς. Υφίσταται μια γνωστή αντιστάθμιση μεταξύ precision και recall: ένα πολύ συντηρητικό μοντέλο

μπορεί να έχει υψηλό precision, δηλαδή λίγα false positives, αλλά χαμηλό recall, ενώ ένα πιο ευαίσθητο μοντέλο μπορεί να εντοπίζει τους περισσότερους κινδύνους ,δηλαδή υψηλό recall, εις βάρος της precision (Kulkarni et al, 2020).

3. F1-score: Το μέτρο F1 συνδυάζει τα παραπάνω σε μια ενιαία τιμή, είναι ο αρμονικός μέσος του precision και του recall. Το F1 δίνει μια ισορροπημένη εικόνα της απόδοσης ως προς την ανάκληση και την ακρίβεια των θετικών προβλέψεων, και δείχνει έντονα περιπτώσεις όπου το ένα από τα δύο είναι πολύ χαμηλό (Hicks et al, 2022). Εφόσον η αθέτηση δανείου είναι το ενδιαφέρον συμβάν (θετική κλάση), το F1-score εδώ αποτιμά κατά πόσο το μοντέλο τα καταφέρνει συνολικά στο να ανιχνεύσει τις αθετήσεις με σχετική ακρίβεια. Σε σοβαρά ανισοβαρή σύνολα, το F1 είναι πολύ πιο αξιόπιστο από την ακρίβεια, διότι απαιτεί τόσο ικανοποιητική κάλυψη (recall) της μειοψηφίας όσο και λογική ακρίβεια (precision) αυτών των προβλέψεων (Kulkarni et al, 2020).

4. Καμπύλη ROC και AUC: Για πιο συνολική αξιολόγηση, χρησιμοποιείται η καμπύλη ROC (Receiver Operating Characteristic) και το εμβαδόν κάτω από αυτήν (AUC). Η καμπύλη ROC απεικονίζει τη συμπεριφορά του ταξινομητή καθώς μεταβάλλεται το κατώφλι διάκρισης, σχεδιάζοντας τον True Positive Rate (TPR) ως προς τον False Positive Rate (FPR) για όλα τα δυνατά κατώφλια. Ο TPR είναι ουσιαστικά το recall (ευαισθησία), ενώ ο FPR ισούται με το ποσοστό των μη-αθετήσεων που εσφαλμένα σημάνθηκαν ως αθετήσεις. Ένα μοντέλο που δίνει απευθείας πιθανότητες παράγει ένα συνεχές σκορ, η καμπύλη ROC μάς δείχνει το trade-off ευαισθησίας για κάθε πιθανό κατώφλι ταξινόμησης. Η ιδανική καμπύλη ROC θα πήγαινε απότομα επάνω αριστερά (TPR=1 με FPR κοντά στο 0). Αυτό σημαίνει τέλεια ανίχνευση των θετικών με μηδενικούς ψευδείς συναγερμούς. Η διαγώνιος γραμμή αντιστοιχεί σε τυχαιότητα (TPR=FPR). Το εμβαδό - Area Under the Curve (AUC) συνοψίζει την καμπύλη σε έναν αριθμό: κυμαίνεται από 0 έως 1, με 1 για τέλειο ταξινομητή, 0.5 για τυχαίο μοντέλο (διάγωνος) και κάτω από 0.5 για χειρότερο του τυχαίου. Το AUC είναι ιδιαίτερα χρήσιμο σε ανισόρροπα σύνολα δεδομένων διότι δεν επηρεάζεται από την απόλυτη κλίμακα πιθανοτήτων ή το συγκεκριμένο κατώφλι, αξιολογεί την ικανότητα διάκρισης του μοντέλου σε όλες τις βαθμίδες ευαισθησίας (Kulkarni et al, 2020).



Εικόνα 7 Καμπύλη ROC και AUC Πηγή: Kulkarni et al, 2020

Σε περιπτώσεις σοβαρής ανισορροπίας πρέπει να προτιμώνται μέτρα όπως το F1, η καμπύλη Precision-Recall ή η AUC αντί της απλής ακρίβειας. Είναι αξιοσημείωτο ότι το F1-score και η καμπύλη ROC δίνουν διαφορετικές οπτικές: το F1 εστιάζει σε ένα συγκεκριμένο κατώφλι, σε εκείνο που μεγιστοποιεί τον αρμονικό μέσο των Precision/Recall, ενώ η χρήση της AUC τα δύο μοντέλα ανεξαρτήτως κατωφλίου: ποιο, δηλαδή, έχει συνολικά καλύτερη διακριτική ικανότητα.

ΚΕΦΑΛΑΙΟ 4: ΑΝΑΛΥΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Στο παρόν κεφάλαιο παρουσιάζεται η πλήρης ανάλυση των αποτελεσμάτων που προέκυψαν από την εφαρμογή και αξιολόγηση δύο βασικών αλγορίθμων ταξινόμησης, της λογιστικής παλινδρόμησης και του δέντρου αποφάσεων, στο πρόβλημα πρόβλεψης αθέτησης δανείου. Η διαδικασία αξιολόγησης περιλαμβάνει τόσο ποσοτικά μέτρα (accuracy, precision, recall, F1-score, ROC/AUC), όσο και ποιοτική ερμηνεία της συμπεριφοράς των μοντέλων με βάση τα δεδομένα και τα χαρακτηριστικά του προβλήματος.

Ακολουθεί ανάλυση της συμπεριφοράς κάθε αλγορίθμου με βάση τις μετρικές, γραφικές απεικονίσεις, σύγκριση των αποτελεσμάτων, συζήτηση για την καταλληλόλητα κάθε τεχνικής, και τέλος προτάσεις για περαιτέρω βελτιώσεις με έμφαση σε σύγχρονες προσεγγίσεις ensemble learning.

4.1 Εκπαίδευση και Αξιολόγηση Λογιστικής Παλινδρόμησης

4.1.1 Περιγραφή και Εκπαίδευση

Η λογιστική παλινδρόμηση, ως γραμμικό μοντέλο, προσφέρει την απαραίτητη ερμηνευσιμότητα, στοιχείο που τη διατηρεί σταθερά δημοφιλή σε τραπεζικές και χρηματοοικονομικές εφαρμογές. Η ερμηνεία των συντελεστών του μοντέλου, δηλαδή πόσο επηρεάζει κάθε χαρακτηριστικό (όπως το εισόδημα, το επιτόκιο, ο αριθμός των πιστωτικών γραμμών) την πιθανότητα αθέτησης, αποτελεί σημαντικό πλεονέκτημα, καθώς συμβάλλει τόσο στη λήψη επιχειρηματικών αποφάσεων όσο και στη συμμόρφωση με κανονιστικές απαιτήσεις (Lessmann et al., 2015).

Κατά την προετοιμασία του dataset, εφαρμόστηκε κανονικοποίηση (standardization), ώστε όλα τα χαρακτηριστικά να έχουν το ίδιο εύρος τιμών και να μην επικρατεί η επίδραση μεταβλητών με μεγάλη διασπορά (Conroy & Sajda, 2012). Ακολούθως, για την αντιμετώπιση της έντονης ανισορροπίας των κλάσεων, αξιοποιήθηκε η μέθοδος SMOTE (Chawla et al., 2002), που συντέλεσε στη δημιουργία συνθετικών δειγμάτων της μειοψηφικής κλάσης, οδηγώντας σε ισορροπημένη εκπαίδευση και καλύτερη αξιοποίηση των δεδομένων.

Η εκπαίδευση της λογιστικής παλινδρόμησης πραγματοποιήθηκε με μέγιστο αριθμό επαναλήψεων (max_iter=1000), διασφαλίζοντας τη σύγκλιση ακόμη και σε σύνολα με πολυδιάστατα χαρακτηριστικά. Η επιλογή του solver 'lbfgs' θεωρείται state-of-the-art, καθώς συνδυάζει ταχύτητα και σταθερότητα. Τέλος, το μοντέλο αξιολογήθηκε αποκλειστικά στο test set, το οποίο δεν είχε χρησιμοποιηθεί σε κανένα βήμα εκπαίδευσης, εξασφαλίζοντας αντικειμενικότητα στη μέτρηση της απόδοσης.

4.1.2 Ποσοτικά Αποτελέσματα

Ο κώδικας του παραρτήματος Α έδωσε τα εξής αποτελέσματα:

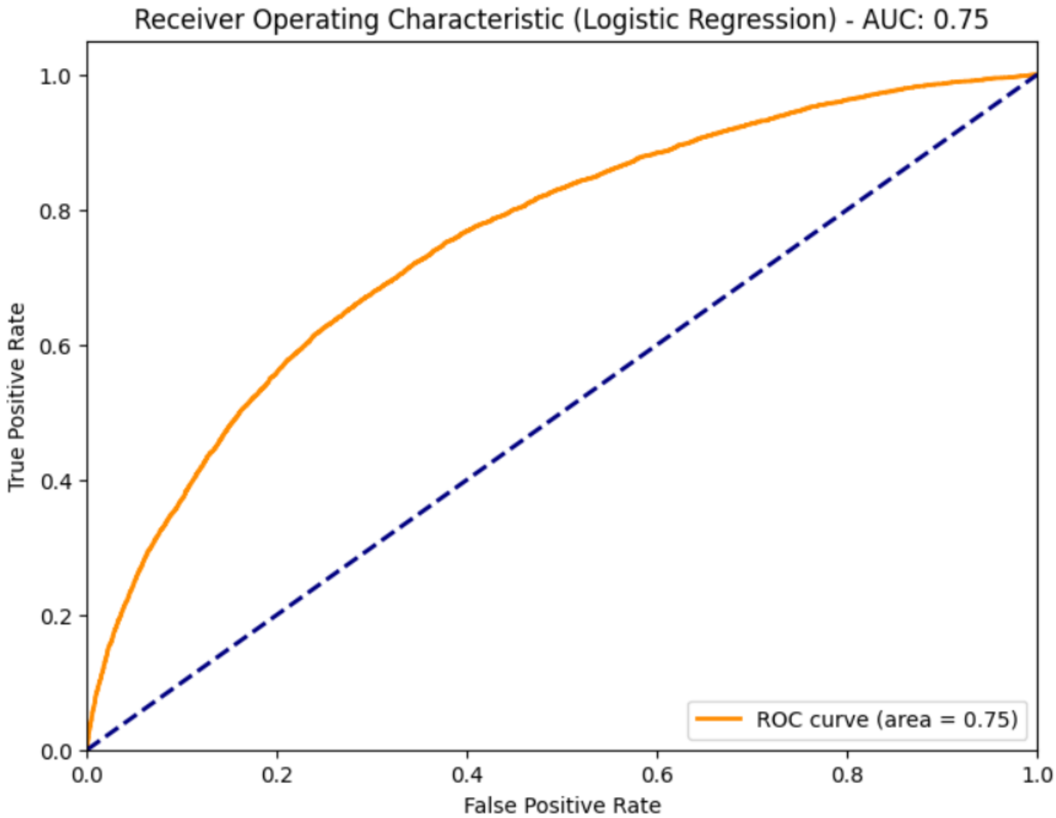
Πίνακας 4 Ποσοτικά αποτελέσματα Λογιστικής Παλινδρόμησης

Μετρική	Κλάση 0	Κλάση 1	Macro Avg	Weighted Avg
Precision	0.94	0.22	0.58	0.86
Recall	0.68	0.69	0.69	0.68
F1-score	0.79	0.34	0.57	0.74
Υποστήριξη	45139	5931		
Accuracy	0.6843			

Η λογιστική παλινδρόμηση πέτυχε accuracy 68.43%. Το υψηλό recall (0.69) στην πρόβλεψη της αθέτησης σημαίνει ότι η πλειοψηφία των πελατών που τελικά δεν αποπλήρωσαν το δάνειο εντοπίζονται σωστά. Ωστόσο, το χαμηλό precision (0.22) υποδηλώνει ότι ένας σημαντικός αριθμός πελατών που χαρακτηρίζονται από το μοντέλο ως "default", στην πραγματικότητα δεν αθετούν, το φαινόμενο γνωστό ως "false positive". Αυτό αποτελεί σημαντικό επιχειρηματικό trade-off: προτιμάται συχνά να θυσιαστεί το precision προς όφελος του recall, ώστε να μειωθεί ο κίνδυνος χρηματοοικονομικής ζημίας από αθέτηση (Lessmann et al., 2015).

Η τιμή του F1-score για την κλάση της αθέτησης (0.34) αποτελεί αντιπροσωπευτικό μέτρο συνολικής απόδοσης, καθώς συνδυάζει το precision και το recall σε μία ενιαία τιμή. Το F1-score άνω του 0.3 θεωρείται ικανοποιητικό για προβλήματα όπου η μειοψηφική κλάση είναι τόσο σπάνια (Brown & Mues, 2012).

Η ROC καμπύλη και το AUC αποτελούν βασικά εργαλεία για την εκτίμηση της διακριτικής ικανότητας των ταξινομητών. Η τιμή AUC = 0.75 (Εικόνα 8) υποδηλώνει ότι το μοντέλο μπορεί να διαχωρίσει αποτελεσματικά τους πελάτες που θα αθετήσουν από εκείνους που δεν θα το πράξουν. Το AUC άνω του 0.7 είναι επιθυμητό όριο για credit scoring (Moscato et al, 2021).



Εικόνα 8 ROC – AUC Λογιστικής Παλινδρόμησης

Στο confusion matrix, παρατηρούμε 4103 σωστές προβλέψεις αθέτησης (true positives), έναντι 1828 false negatives, δηλαδή πελάτες που αθέτησαν αλλά το μοντέλο δεν το προέβλεψε και 14295 false positives, δηλαδή πελάτες που θεωρήθηκαν αθετητές αλλά δεν ήταν. Η παρουσία σημαντικού αριθμού false positives είναι συνηθισμένη στα προβλήματα αυτά, και η διαχείρισή τους αποτελεί αντικείμενο περαιτέρω ανάλυσης.

Πίνακας 5 Confusion Matrix Λογιστικής Παλινδρόμησης

	Προβλέπει 0	Προβλέπει 1
Πραγματικό 0	30844	14295
Πραγματικό 1	1828	4103

4.1.3 Ερμηνεία και Σχολιασμός

Η λογιστική παλινδρόμηση αποδίδει καλύτερα σε σύνολα όπου τα χαρακτηριστικά έχουν ουσιαστική συνεισφορά στη διαφοροποίηση μεταξύ των κλάσεων (Gopinath et al., 2021). Η γραμμικότητα του μοντέλου επιτρέπει την απευθείας αξιολόγηση του βάρους κάθε χαρακτηριστικού στην πρόβλεψη, γεγονός που βοηθά και στη διαδικασία επιλογής μεταβλητών, μια επιπλέον ενέργεια που θα μπορούσε να βελτιώσει τη σταθερότητα και την απλοποίηση του μοντέλου.

Είναι σημαντικό να τονιστεί ότι τα όρια recall και precision είναι αποτέλεσμα τόσο της φύσης των δεδομένων όσο και των παραμέτρων του μοντέλου, όπως το threshold της πιθανότητας. Σε επιχειρησιακά περιβάλλοντα, η τιμή του threshold μπορεί να ρυθμιστεί ανάλογα με το αποδεκτό ρίσκο και τα κόστη των ψευδών προβλέψεων.

Τέλος, η λογιστική παλινδρόμηση, λόγω της στατιστικής της βάσης και της απλότητάς της, συχνά λειτουργεί ως benchmark για τη σύγκριση πιο περίπλοκων αλγορίθμων, διατηρώντας υψηλή εμπιστοσύνη από τις ρυθμιστικές αρχές. Παρά τα όρια στην αντιμετώπιση μη γραμμικών σχέσεων ή σύνθετων αλληλεπιδράσεων, το μοντέλο κρίνεται κατάλληλο για τα περισσότερα τυπικά προβλήματα credit scoring.

4.2 Εκπαίδευση και Αξιολόγηση Δέντρου Αποφάσεων

4.2.1 Περιγραφή και Εκπαίδευση του Μοντέλου

Το δέντρο αποφάσεων ανήκει στην κατηγορία των μη παραμετρικών, εποπτευόμενων αλγορίθμων ταξινόμησης και διακρίνεται για τη διαφάνειά του και την ευκολία ερμηνείας των αποτελεσμάτων. Κατά την εκπαίδευση, ο αλγόριθμος κατασκευάζει ένα διαδοχικό σύνολο ερωτήσεων που οδηγούν σε προβλέψεις, ομαδοποιώντας τα δείγματα με βάση χαρακτηριστικά όπως η ηλικία, το εισόδημα, το επιτόκιο ή ο αριθμός πιστωτικών γραμμών (Kotu & Deshpande, 2019). Αυτό το δομικό πλεονέκτημα επιτρέπει σε χρηματοπιστωτικούς οργανισμούς να κατανοούν με σαφήνεια ποιοι παράγοντες συμβάλλουν περισσότερο στην απόφαση για αθέτηση δανείου. Για την αποφυγή υπερεκμάθησης (overfitting), που αποτελεί κοινό πρόβλημα στα δέντρα αποφάσεων λόγω της υψηλής ευελιξίας τους, επιλέχθηκε μέγιστο βάθος max_depth=5 και ελάχιστος αριθμός δειγμάτων για διαχωρισμό min_samples_split=10. Αυτοί οι περιορισμοί είναι σημαντικοί διότι το βάθος του δέντρου και το μέγεθος των φύλλων επηρεάζουν σημαντικά τη γενικευσιμότητα και τη σταθερότητα του αλγορίθμου (Halabaku & Bytyçi, 2024). Σε πιο προηγμένες εφαρμογές, μπορεί να εφαρμοστεί και κλάδεμα (pruning), είτε κατά τη διάρκεια της εκπαίδευσης (pre-pruning) είτε εκ των υστέρων (post-pruning), ώστε να απομακρύνονται διαχωρισμοί που δεν συνεισφέρουν ουσιαστικά στη βελτίωση της απόδοσης σε σετ επικύρωσης (validation set) (Halabaku & Bytyçi, 2024).

Η εκπαιδευτική διαδικασία του δέντρου πραγματοποιήθηκε με το ίδιο ισορροπημένο training set που δημιουργήθηκε με SMOTE, διασφαλίζοντας δίκαιη σύγκριση με τη λογιστική παλινδρόμηση. Επιπρόσθετα, το δέντρο έχει το πλεονέκτημα ότι μπορεί να χειρίζεται τόσο κατηγορικά όσο και αριθμητικά χαρακτηριστικά χωρίς να απαιτείται ειδική κωδικοποίηση ή κανονικοποίηση, αν και στην παρούσα ανάλυση εφαρμόστηκε ενιαία προεπεξεργασία για λόγους συγκρισιμότητας.

4.2.2 Ποσοτικά Αποτελέσματα

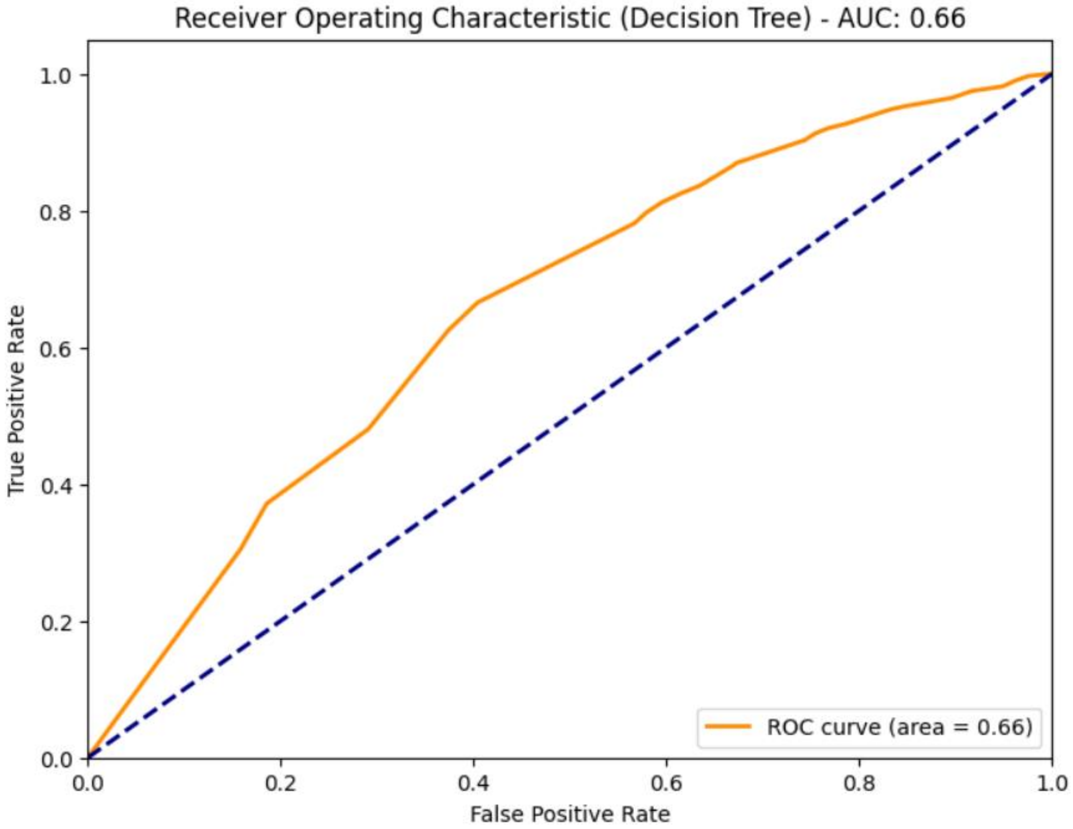
Ο κώδικας του παραρτήματος Α έδωσε τα εξής αποτελέσματα:

Πίνακας 6 Ποσοτικά αποτελέσματα Δέντρου Αποφάσεων

Μετρική	Κλάση 0	Κλάση 1	Macro Avg	Weighted Avg
Precision	0.91	0.18	0.55	0.83
Recall	0.71	0.48	0.59	0.68
F1-score	0.80	0.26	0.53	0.74
Υποστήριξη	45139	5931		
Accuracy	0.6825			

Αναλύοντας τα αριθμητικά αποτελέσματα, το decision tree πέτυχε accuracy 68,25% και F1-score 0,26 για την κλάση αθέτησης, τιμή κατώτερη από εκείνη της λογιστικής παλινδρόμησης. Το recall για τα περιστατικά αθέτησης ανήλθε σε 0,48, τιμή σημαντικά χαμηλότερη από το αντίστοιχο αποτέλεσμα της λογιστικής παλινδρόμησης, ενώ το precision διαμορφώθηκε στο 0,18. Αυτή η απόδοση ερμηνεύεται από τη φύση των δέντρων αποφάσεων: οι αποφάσεις βασίζονται σε τοπικές διαχωριστικές γραμμές στον χώρο χαρακτηριστικών, ενώ τα σπάνια περιστατικά (μειοψηφική κλάση) συχνά συγχωνεύονται σε κόμβους που ευνοούν την πλειοψηφία. Επιπλέον, λόγω της περιορισμένης πολυπλοκότητας (μικρό βάθος), το μοντέλο δυσκολεύεται να ανιχνεύσει λεπτές διαφοροποιήσεις, οδηγώντας σε χαμηλότερο recall και precision για την κλάση της αθέτησης (Lee et al., 2014).

Η τιμή ROC-AUC ήταν 0.66 (Εικόνα 9) ενισχύει την παραπάνω παρατήρηση, υποδηλώνοντας ότι το δέντρο αποφάσεων έχει μέτρια ικανότητα διάκρισης μεταξύ αθετητών και μη αθετητών.



Εικόνα 9 ROC - AUC Δέντρου Αποφάσεων

Στο confusion matrix, παρατηρούμε 2850 σωστές προβλέψεις αθέτησης (true positives), έναντι 3081 false negatives, δηλαδή πελάτες που αθέτησαν αλλά το μοντέλο δεν το προέβλεψε, και 13135 false positives, δηλαδή πελάτες που θεωρήθηκαν αθετητές αλλά δεν ήταν.

Πίνακας 7 Confusion matrix Δέντρου Αποφάσεων

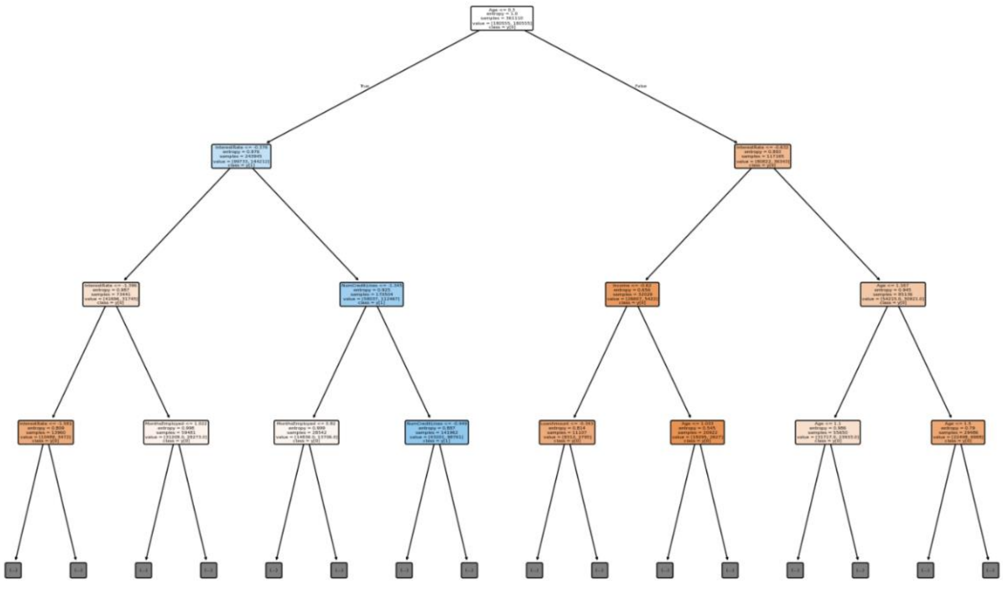
	Προβλέπει 0	Προβλέπει 1
Πραγματικό 0	32004	13135

Πραγματικό 1	3081	2850

4.2.3 Ερμηνεία και Σχολιασμός

Το decision tree εμφάνισε βελτιωμένο recall (0.48) σε σχέση με το precision (0.18) για την κλάση αθέτησης, ενώ το συνολικό F1-score για την ίδια τάξη διαμορφώθηκε σε 0.26. Το ποσοστό σωστών προβλέψεων αθέτησης (true positives) είναι σημαντικό, αλλά το μοντέλο παραμένει ευάλωτο σε ψευδώς θετικές προβλέψεις, γεγονός που αντικατοπτρίζει το γνωστό bias-variance trade-off σε δέντρα χαμηλού βάθους (Guan & Burton, 2022).

Η ερμηνευσιμότητα των δέντρων αποφάσεων παραμένει αδιαμφισβήτητο πλεονέκτημα. Όπως απεικονίζεται παρακάτω στο διάγραμμα του δέντρου (Εικόνα 10), οι βασικές μεταβλητές που χρησιμοποιούνται στους πρώτους διαχωρισμούς, δηλαδή Age, InterestRate, Income, NumCreditLines και MonthsEmployed, αντανακλούν χαρακτηριστικά που είναι ευρέως αναγνωρισμένα ως δείκτες πιστωτικού κινδύνου (Saunders & Cornett 2011). Η δομή του δέντρου επιτρέπει τη γρήγορη εξαγωγή απλών επιχειρηματικών κανόνων (π.χ. "οι νεότεροι δανειολήπτες με υψηλό επιτόκιο και χαμηλό εισόδημα εμφανίζουν αυξημένη πιθανότητα αθέτησης").



Εικόνα 10 Οπτική αναπαράσταση των πρώτων 3 επιπέδων του δέντρου αποφάσεων

4.3 Σύγκριση Μοντέλων

4.3.1 Ποσοτική Σύγκριση

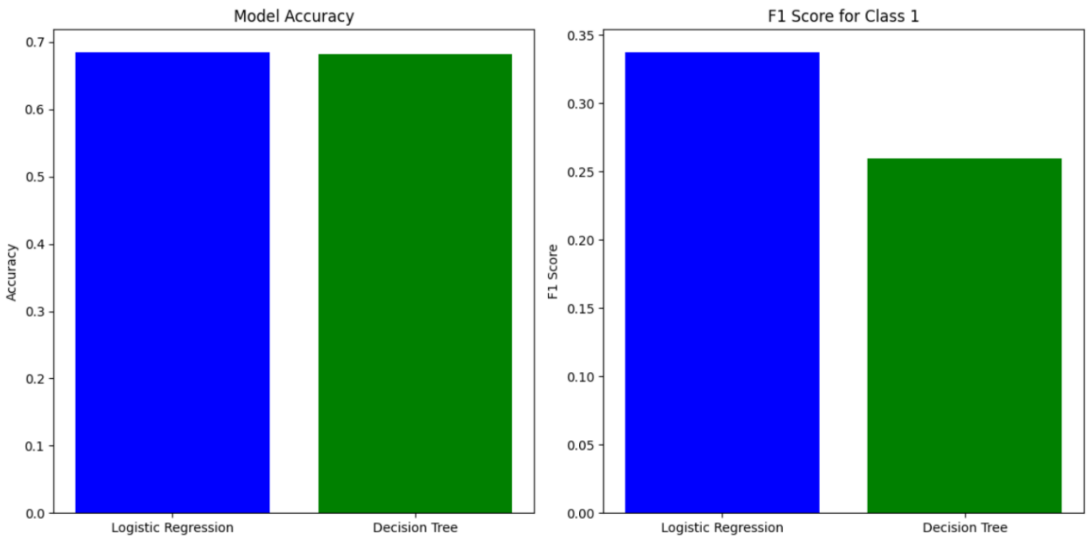
Η συγκριτική αξιολόγηση της λογιστικής παλινδρόμησης και του δέντρου αποφάσεων είναι θεμελιώδης για τη διαμόρφωση τεκμηριωμένων συμπερασμάτων σχετικά με την καταλληλόλητά τους στη διαχείριση πιστωτικού κινδύνου και την πρόβλεψη αθέτησης δανείου.

Πίνακας 8 Σύγκριση βασικών μετρικών μεταξύ Λογιστικής Παλινδρόμησης και Δέντρο Αποφάσεων

Μοντέλο	Accuracy	Recall (κλ. 1)	F1-score (κλ. 1)	ROC-AUC	Precision (κλ. 0)
Logistic Regression	0.6843	0.69	0.34	0.75	0.94
Decision Tree	0.6825	0.48	0.26	0.66	0.91

Η λογιστική παλινδρόμηση υπερτερεί ξεκάθαρα σε F1-score, recall και ROC-AUC στην πρόβλεψη αθέτησης, ενώ το δέντρο αποφάσεων εμφανίζει ελαφρώς καλύτερη ικανότητα στην ταξινόμηση της πλειοψηφικής τάξης αλλά υστερεί στη μειοψηφική.

Η συνολική ακρίβεια (accuracy) των δύο μοντέλων είναι πολύ κοντά, επομένως, η λεπτομερέστερη ανάλυση των κρίσιμων μετρικών ανά κατηγορία αποκαλύπτει σημαντικές διαφορές. Η λογιστική παλινδρόμηση παρουσιάζει αισθητά υψηλότερο recall (0,69 έναντι 0,48) και F1-score (0,34 έναντι 0,26) για την κλάση αθέτησης, ενώ το decision tree εμφανίζει ελαφρώς χαμηλότερη precision για την κλάση μη αθέτησης (0,91 έναντι 0,94). Η υπεροχή της λογιστικής παλινδρόμησης στον δείκτη AUC (0,75 έναντι 0,66) επιβεβαιώνει τη μεγαλύτερη διακριτική της ικανότητα, στοιχείο που έχει ιδιαίτερη αξία σε περιβάλλοντα με έντονη ανισορροπία κλάσεων (Fernández et al., 2018).



Εικόνα 11 Bar plots για Accuracy και F1-score

Η χρήση bar plots και ROC curves προσφέρει οπτική σύνοψη των διαφορών (Εικόνα 11). Τα διαγράμματα καταδεικνύουν πως, μολονότι τα δύο μοντέλα έχουν παρόμοιο συνολικό accuracy, η λογιστική παλινδρόμηση πλεονεκτεί ξεκάθαρα στη δυνατότητα ανίχνευσης περιστατικών αθέτησης, με πολύ μικρότερη απώλεια precision σε σχέση με το δέντρο αποφάσεων.

4.3.2 Ποιοτική Σύγκριση

Η λογιστική παλινδρόμηση διακρίνεται για την υψηλή της σταθερότητα και ερμηνευσιμότητα. Οι συντελεστές (weights) που εκτιμώνται κατά την εκπαίδευση του μοντέλου επιτρέπουν στον αναλυτή να κατανοήσει σε βάθος τον τρόπο που κάθε χαρακτηριστικό επηρεάζει το τελικό αποτέλεσμα, στοιχείο καθοριστικό για τη συμμόρφωση με κανονιστικά πλαίσια και για την επιχειρησιακή αξιολόγηση των κινδύνων. Επιπλέον, το γραμμικό της υπόβαθρο μειώνει το variance και προσφέρει γενικά σταθερές επιδόσεις, ακόμα και σε παρουσία θορύβου ή μεταβαλλόμενων δεδομένων (Rudd & Priestley, 2017).

Αντίθετα, το δέντρο αποφάσεων πλεονεκτεί στην ενστικτώδη κατανόηση του τρόπου λήψης αποφάσεων, καθώς κάθε διαδρομή από τη ρίζα σε ένα φύλλο αντιστοιχεί σε έναν ρητό κανόνα. Επομένως, οι κανόνες αυτοί ενδέχεται να είναι υπεραπλουστευμένοι, ειδικά όταν το βάθος του δέντρου είναι περιορισμένο, και να παραβλέπουν πολύπλοκες συσχετίσεις μεταξύ χαρακτηριστικών. Παράλληλα, η υψηλή διακύμανση (variance) των δέντρων αποφάσεων τα καθιστά ευάλωτα σε μικρές αλλαγές στα δεδομένα, οδηγώντας σε σημαντικές μεταβολές στη δομή του δέντρου. Αυτό μειώνει τη σταθερότητα των προβλέψεων και δυσχεραίνει την αναπαραγωγή των αποτελεσμάτων (Dumitrescu et al, 2022).

4.3.3 Ανάλυση Σφαλμάτων: Κόστος false positives και false negatives

Ένα από τα πλέον κρίσιμα ζητήματα στη διαχείριση πιστωτικού κινδύνου είναι η ισορροπία μεταξύ false positives, δηλαδή πελάτες που προβλέπεται ότι θα αθετήσουν, αλλά τελικά δεν το κάνουν και false negatives, δηλαδή πελάτες που δεν ανιχνεύονται ως επικίνδυνοι, αλλά τελικά αθετούν.

Όπως προκύπτει από το confusion matrix, η λογιστική παλινδρόμηση καταφέρνει να εντοπίσει περισσότερες πραγματικές αθετήσεις true positives, θυσιάζοντας ωστόσο την ακρίβεια της πρόβλεψης (precision), εμφανίζοντας, δηλαδή, μεγαλύτερο αριθμό false positives. Το decision tree εμφανίζει μεγαλύτερο αριθμό false negatives, γεγονός που μπορεί να έχει ιδιαίτερα αρνητικές επιπτώσεις για τον οργανισμό, αφού κάθε μη ανιχνευμένη αθέτηση δύναται να προκαλέσει σημαντική χρηματοοικονομική ζημιά.

Αυτό το trade-off μεταξύ recall και precision, όπως αποτυπώνεται στον δείκτη F1-score, πρέπει να λαμβάνεται υπόψη κατά την επιλογή μοντέλου, με βάση τα ιδιαίτερα χαρακτηριστικά και τις ανάγκες της εφαρμογής. Σε περιβάλλοντα όπου το κόστος μιας αθέτησης είναι πολύ υψηλό, προτιμάται η μεγιστοποίηση του recall ακόμη και αν αυτό οδηγεί σε αύξηση των false positives, πρακτική που συνάδει με τη στρατηγική προληπτικής διαχείρισης κινδύνου (Chang et al, 2024).

4.3.4 Περιορισμοί και Πηγές Μεροληψίας

Είναι απαραίτητο να σημειωθούν οι περιορισμοί των δύο μεθόδων. Η λογιστική παλινδρόμηση, ως γραμμικό μοντέλο, ενδέχεται να υποεκτιμήσει τη συνεισφορά μη γραμμικών σχέσεων ή αλληλεπιδράσεων μεταξύ μεταβλητών. Παράλληλα, τα δέντρα αποφάσεων, κυρίως όταν έχουν περιορισμένο βάθος, μπορεί να αδυνατούν να εντοπίσουν λεπτές διαφοροποιήσεις στη συμπεριφορά των δανειοληπτών. Επιπλέον, το sampling με SMOTE, παρότι βελτιώνει το recall, μπορεί να εισάγει συνθετική μεροληψία και να αλλοιώσει, σε κάποιο βαθμό, τη φυσική κατανομή των δεδομένων (Chawla et al., 2002).

Η ερμηνεία των αποτελεσμάτων πρέπει να γίνεται λαμβάνοντας υπόψη το συγκεκριμένο dataset, τα ποσοστά αθέτησης, τη φύση των χαρακτηριστικών και τη δυνατότητα γενίκευσης των συμπερασμάτων.

4.4 Καταλληλότητα Τεχνικών

Η επιλογή της κατάλληλης ταξινομητικής τεχνικής στην πρόβλεψη αθέτησης δανείου δεν αποτελεί απλώς ζήτημα αριθμητικής απόδοσης σε τυπικές μετρικές, αλλά σχετίζεται άμεσα με τη στρατηγική του οργανισμού, τη φύση των διαθέσιμων δεδομένων και τις επιχειρησιακές ανάγκες για ερμηνεία, κανονιστική συμμόρφωση και διαχείριση ρίσκου.

Η λογιστική παλινδρόμηση αναδεικνύεται σε βασικό εργαλείο για τον τραπεζικό τομέα λόγω της σαφούς ερμηνείας των αποτελεσμάτων της. Οι συντελεστές του μοντέλου μπορούν να μεταφραστούν σε ποσοτικές εκτιμήσεις για τον τρόπο που κάθε χαρακτηριστικό επηρεάζει τον πιστωτικό κίνδυνο, στοιχείο θεμελιώδες για τη λήψη επιχειρησιακών αποφάσεων και την ικανοποίηση των απαιτήσεων διαφάνειας από τις ρυθμιστικές αρχές. Η σταθερότητα των προβλέψεων της λογιστικής παλινδρόμησης είναι σημαντική, ιδίως σε περιβάλλοντα όπου οι διακυμάνσεις στα δεδομένα είναι συχνές ή όπου απαιτείται αξιοπιστία και συνέπεια στις αποφάσεις πιστοδότησης. Επιπλέον, το μοντέλο είναι ιδιαίτερα αποτελεσματικό όταν τα χαρακτηριστικά έχουν γραμμική συσχέτιση με την πιθανότητα αθέτησης (Costa-e-Silva & Lopes, 2020).

Το δέντρο αποφάσεων ξεχωρίζει για την «ανθρωποκεντρική» του ερμηνευσιμότητα και τη δυνατότητα να απεικονίζει την απόφαση με τη μορφή σαφών κανόνων τύπου «if-then-else». Αυτό το χαρακτηριστικό το καθιστά ιδιαίτερα χρήσιμο στις αρχικές φάσεις ανάλυσης δεδομένων (exploratory data analysis) ή όταν υπάρχει ανάγκη για γρήγορη ανάπτυξη κανόνων σε πραγματικό χρόνο. Τα δέντρα αποφάσεων έχουν επίσης τη δυνατότητα να μοντελοποιούν μη γραμμικές σχέσεις και να εντοπίζουν αλληλεπιδράσεις μεταξύ χαρακτηριστικών που δεν μπορούν να ανιχνευτούν από ένα γραμμικό μοντέλο, ειδικά αν αυξηθεί το μέγιστο επιτρεπτό βάθος τους (Vidovic & Yue, 2020). Ωστόσο, όπως αποδείχθηκε και στα αποτελέσματα του παρόντος πειράματος, τα δέντρα περιορισμένου βάθους υπολείπονται στην πρόβλεψη σπάνιων τάξεων και υποφέρουν από υψηλό variance.

Σε οργανισμούς όπου το κόστος της αθέτησης (false negative) είναι εξαιρετικά υψηλό, όπως τράπεζες με χαρτοφυλάκια υψηλού κινδύνου, η προτεραιότητα δίνεται στην επίτευξη υψηλού recall, ακόμη και εις βάρος του precision. Σε αυτές τις περιπτώσεις, η λογιστική παλινδρόμηση εμφανίζει συγκριτικό πλεονέκτημα, ιδίως μετά από εφαρμογή τεχνικών balancing όπως SMOTE (Chawla et al. 2002).

Αντίθετα, σε εφαρμογές όπου η ανάγκη για απλότητα, αυτοματισμό και δημιουργία απλών επιχειρησιακών κανόνων υπερισχύει (π.χ. προ-έγκριση μικρών ποσών, real-time rules σε e-banking), το δέντρο αποφάσεων αποτελούν πρακτική λύση.

Παρά τις αρετές τους, και οι δύο προσεγγίσεις έχουν περιορισμούς:

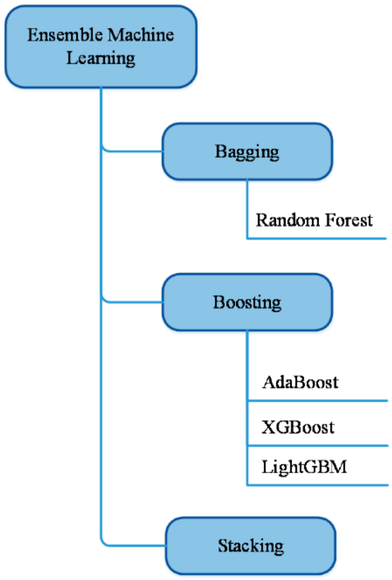
1. Η λογιστική παλινδρόμηση ενδέχεται να υποεκτιμά μη γραμμικές σχέσεις ή ισχυρές αλληλεπιδράσεις, ενώ προϋποθέτει σωστή προεπεξεργασία των χαρακτηριστικών (standardization, dummy encoding).

2. Το decision tree έχει χαμηλή σταθερότητα, ιδίως σε μικρά και θορυβώδη δείγματα, τάση προς overfitting χωρίς έλεγχο βάθους και αδυναμία σε datasets με πολλές αλληλεπιδράσεις ή υπερβολικά ανισόρροπες τάξεις.

Συνοψίζοντας, η επιλογή μεταξύ των δύο ή η υβριδική χρήση τους πρέπει να βασίζεται σε ενδελεχή αξιολόγηση τόσο ποσοτικών μετρικών όσο και επιχειρησιακών αναγκών, με προσοχή στους περιορισμούς και στις δυνητικές πηγές σφάλματος ή μεροληψίας

4.5 Προτάσεις για Βελτιώσεις: Ensemble Learning

Η πρόβλεψη πιστωτικού κινδύνου αποτελεί κρίσιμο ζήτημα για τα χρηματοπιστωτικά ιδρύματα. Παραδοσιακά, τεχνικές όπως η βαθμολόγηση πιστοληπτικής ικανότητας (credit scoring) βασίζονταν σε στατιστικά μοντέλα (π.χ. λογιστική παλινδρόμηση), όμως οι σύγχρονες μέθοδοι μηχανικής μάθησης έχουν βελτιώσει σημαντικά την ακρίβεια των προβλέψεων. Ιδιαίτερα, οι συνδυαστικές μέθοδοι μάθησης (ensemble learning), όπου συνδυάζονται πολλαπλά μοντέλα, αναδεικνύονται ως μια πολλά υποσχόμενη προσέγγιση για την ενίσχυση των επιδόσεων πρόβλεψης σε προβλήματα πιστωτικού κινδύνου. Η χρήση ensemble μοντέλων σε εφαρμογές όπως το credit scoring, την πρόβλεψη αθέτησης δανείων και τη διαχείριση πιστωτικών χαρτοφυλακίων, βρίσκοντας ότι τα ensemble συχνά υπερτερούν των μεμονωμένων ταξινομητών σε όρους ακρίβειας και αξιοπιστίας (Li & Chen, 2020).



Εικόνα 12 Ensemble machine learning framework. Πηγή: Li & Chen, (2020)

Στον χώρο του credit scoring, όπου το ζητούμενο είναι η κατάταξη νέων αιτούντων δανείου σε «καλούς» ή «κακούς» πελάτες, τα ensemble μοντέλα έχουν παρουσιάσει ανώτερη απόδοση συγκριτικά με κλασικές μεθόδους. Σε μια συγκριτική αξιολόγηση πέντε ensemble τεχνικών, συμπεριλαμβανομένων των Random Forest, AdaBoost, XGBoost, LightGBM και stacking, βρέθηκε ότι συνολικά τα ensemble επιτυγχάνουν υψηλότερη διακριτική ικανότητα και ακρίβεια από τους παραδοσιακούς ταξινομητές. (Li & Chen, 2020)

Συγκεκριμένα, ο Random Forest παρουσίασε την καλύτερη επίδοση (ACC ~81%, AUC ~86%), με τα ενισχυτικά μοντέλα XGBoost και LightGBM να ακολουθούν. Τα αποτελέσματα αυτά υπογραμμίζουν ότι η συνδυαστική μάθηση μπορεί να υπερβεί τα όρια ενός μεμονωμένου μοντέλου: για παράδειγμα, το μοντέλο Random Forest πέτυχε σημαντικά υψηλότερη ακρίβεια και AUC από ένα απλό δέντρο αποφάσεων, επιβεβαιώνοντας πως η ενσωμάτωση πολλαπλών δέντρων μειώνει την διακύμανση και βελτιώνει τη γενίκευση. (Li & Chen, 2020)

Ενδεικτικό είναι ότι, βάσει πειραματικών αποτελεσμάτων, όλα τα ensemble μοντέλα στην μελέτη των Li & Chen (2020) υπερκέρασαν τους μεμονωμένους μαθηματικούς ταξινομητές, όπως τη λογιστική παλινδρόμηση, τα νευρωνικά δίκτυα, τα δέντρα, σε κρίσιμους δείκτες απόδοσης. Μοναδική εξαίρεση αποτέλεσε το AdaBoost, το οποίο εμφάνισε απόδοση συγκρίσιμη ή και κατώτερη από ορισμένα απλά μοντέλα, πιθανώς λόγω ευαισθησίας σε θορυβώδη δεδομένα που οδήγησε σε υπερεκμάθηση

Παρά ταύτα, η γενική τάση είναι σαφής: οι μέθοδοι ensemble βελτιώνουν την προβλεπτική ακρίβεια στο credit scoring και συστήνονται ως βέλτιστη πρακτική για τα χρηματοπιστωτικά ιδρύματα. Αξίζει να σημειωθεί ότι ακόμα και αν παραδοσιακά η λογιστική παλινδρόμηση παραμένει ισχυρή βάση (συχνά λόγω ερμηνευσιμότητας), τα σύγχρονα ensembles όπως τα δασικά μοντέλα και τα συστήματα boosting προσφέρουν μια σαφή βελτίωση στην ικανότητα εντοπισμού υψηλού κινδύνου δανειοληπτών (Li & Chen, 2020).

ΚΕΦΑΛΑΙΟ 5: ΜΕΛΕΤΕΣ ΠΕΡΙΠΤΩΣΗΣ

Η εκτίμηση του πιστωτικού κινδύνου και η πρόβλεψη αθέτησης δανείων αποτελούν βασικές προκλήσεις για τις τράπεζες και τα χρηματοπιστωτικά ιδρύματα. Παραδοσιακά, χρησιμοποιούνται στατιστικά μοντέλα όπως η λογιστική παλινδρόμηση για την πρόβλεψη της πιθανότητας αθέτησης, λόγω της απλότητας και της ερμηνευσιμότητάς τους. Παράλληλα, σύγχρονες μέθοδοι μηχανικής μάθησης όπως τα δέντρα αποφάσεων έχουν επίσης υιοθετηθεί για τον ίδιο σκοπό. Πολλές μελέτες έχουν συγκρίνει διάφορα μοντέλα στο πρόβλημα του credit scoring και έχουν καταδείξει ότι πέρα από τα παραδοσιακά μοντέλα, όπως γραμμική ή λογιστική παλινδρόμηση, διαχωριστική ανάλυση κ.α, τεχνικές μηχανικής μάθησης όπως τα δέντρα αποφάσεων, τα νευρωνικά δίκτυα και οι μέθοδοι ενίσχυσης (ensembles) μπορούν να βελτιώσουν την ακρίβεια της πρόβλεψης.

5.1 Μελέτη Περίπτωσης: Credins Bank

Ένα χαρακτηριστικό case study προέρχεται από την εμπορική τράπεζα Credins Bank στην Αλβανία. Σύμφωνα με τους Fejza et al. (2022), η μελέτη αυτή πραγματοποιήθηκε με κίνητρο την ανάγκη της τράπεζας για ένα αξιόπιστο εργαλείο πιστωτικής βαθμολόγησης προσαρμοσμένο στα δεδομένα των πελατών της. Τα δεδομένα της Credins Bank παρουσίαζαν δύο σημαντικές προκλήσεις: (1) έντονη ανισορροπία μεταξύ «καλών» και «κακών» δανειοληπτών, και (2) μεροληψία και ελλιπή πληροφορία σε αρκετά πεδία, λόγω του σχετικά νεότερου χαρακτήρα του τραπεζικού συστήματος στη χώρα. Τέτοιου είδους προβλήματα σπανίζουν σε καθιερωμένα datasets της βιβλιογραφίας, συνεπώς η μελέτη των Fejza et al. (2022) αποτελεί μια από τις πρώτες λεπτομερείς αναλύσεις με πραγματικά δεδομένα από αλβανική τράπεζα.

Οι ερευνητές εφάρμοσαν και αξιολόγησαν πληθώρα μεθόδων μηχανικής μάθησης για την εκτίμηση του πιστωτικού κινδύνου σε αυτό το dataset. Μεταξύ των μοντέλων που συγκρίθηκαν ήταν η λογιστική παλινδρόμηση (ως εκπρόσωπος των γραμμικών στατιστικών μοντέλων) και ένα απλό δέντρο αποφάσεων, καθώς και πιο σύνθετοι αλγόριθμοι όπως το τυχαίο δάσος (Random Forest) και επιτηρούμενης μηχανικής μάθησης (supervised machine learning). Σημαντικό ζήτημα ήταν η αντιμετώπιση του έντονου class imbalance (ανισορροπίας κατηγοριών). Για τον σκοπό αυτό δοκιμάστηκαν τεχνικές όπως η δημιουργία «εξισορροπημένου» Random Forest, όπου το σύνολο εκπαίδευσης τροποποιείται ώστε να μειωθεί η επίδραση της ανισορροπίας των τάξεων (Fejza et al. 2022).

Η συγκριτική αξιολόγηση ανέδειξε ενδιαφέροντα ευρήματα. Το απλό δέντρο αποφάσεων πέτυχε 77.8% συνολική ακρίβεια με AUC = 0.698, επίδοση παρόμοια με εκείνη της λογιστικής παλινδρόμησης με AUC = 0.69, στο ίδιο σύνολο δεδομένων. Αντίθετα, το Random Forest ανέβασε την ακρίβεια στο 88.7% και AUC = 0.765. Συνολικά, το εξισορροπημένο Random Forest αναδείχθηκε ως η βέλτιστη προσέγγιση για τα δεδομένα της Credins Bank, ξεπερνώντας σε απόδοση τόσο τα δέντρα αποφάσεων όσο και τα λοιπά μοντέλα που δοκιμάστηκαν (Fejza et al. 2022).

Η ανάλυση της μελέτης περίπτωσης της Credins Bank δείχνει ότι, αν και τα βασικά μοντέλα όπως η λογιστική παλινδρόμηση και τα δέντρα αποφάσεων μπορούν να προσφέρουν μια ικανοποιητική αφετηρία για την αξιολόγηση πιστωτικού κινδύνου, συχνά δεν αρκούν για να αποδώσουν τα μέγιστα σε πραγματικά, δύσκολα datasets με έντονη ανισορροπία κατηγοριών και ελλιπείς πληροφορίες. Τα αποτελέσματα της μελέτης αποδεικνύουν ότι η υιοθέτηση πιο εξελιγμένων μεθόδων θα ήταν χρήσιμη για να επιτευχθεί σημαντική βελτίωση στην ακρίβεια και την αξιοπιστία των προβλέψεων.

5.2 Μελέτη Περίπτωσης: Lending Club

Το Lending Club είναι μια από τις μεγαλύτερες πλατφόρμες peer-to-peer δανεισμού παγκοσμίως, όπου επενδυτές και δανειολήπτες συνδέονται χωρίς τη μεσολάβηση παραδοσιακών τραπεζών. Λόγω του μεγάλου όγκου και της ποικιλίας των αιτήσεων δανείων, η ανάγκη για ακριβή πρόβλεψη του πιστωτικού κινδύνου είναι ιδιαίτερα αυξημένη. Στη μελέτη των Ko et al. (2022), εξετάστηκε ένα μεγάλο σύνολο πραγματικών δεδομένων δανείων του Lending Club, με στόχο τη σύγκριση της αποδοτικότητας διαφορετικών μοντέλων για την πρόβλεψη αθέτησης δανείων (default). Οι ερευνητές εφάρμοσαν τόσο κλασικά στατιστικά μοντέλα όπως η λογιστική παλινδρόμηση όσο και αλγορίθμους μηχανικής μάθησης, μεταξύ των οποίων δέντρα αποφάσεων, τυχαία δάση (Random Forest), XGBoost και LightGBM.

Η λογιστική παλινδρόμηση χρησιμοποιήθηκε ως «baseline» (σημείο αναφοράς), λόγω της διαχρονικής αξιοπιστίας της και της διαφάνειάς της στα χρηματοοικονομικά ιδρύματα. Τα δέντρα αποφάσεων και ειδικά οι μέθοδοι ενισχυμένης μάθησης (όπως το LightGBM) επιλέχθηκαν για τη δυνατότητά τους να μοντελοποιούν μη γραμμικές και πολύπλοκες αλληλεπιδράσεις μεταξύ μεταβλητών (Ko et al. 2022).

Η λογιστική παλινδρόμηση και τα απλά δέντρα αποφάσεων είχαν ικανοποιητική επίδοση, με σχετικά υψηλή ακρίβεια και ερμηνευσιμότητα. Ωστόσο, παρουσίασαν περιορισμούς ως προς την ικανότητα γενίκευσης σε πιο σύνθετα ή ανισόρροπα δεδομένα.

Η χρήση ensemble models, όπως το LightGBM το οποίο είναι μια εξελιγμένη εκδοχή δέντρων αποφάσεων με gradient boosting, οδήγησε σε σημαντική βελτίωση της απόδοσης, με 2.9% υψηλότερη ακρίβεια σε σχέση με τα απλά μοντέλα, και AUC που προσέγγισε το 0.80. Αυτό μεταφράστηκε σε πρακτικούς όρους σε δυνητική αύξηση εσόδων άνω των 24 εκατομμυρίων δολαρίων, λόγω της μείωσης των μη εξυπηρετούμενων δανείων (Ko et al. 2022).

Τα αποτελέσματα αυτά επιβεβαιώνουν ότι ενώ η λογιστική παλινδρόμηση και τα δέντρα αποφάσεων αποτελούν βασικά εργαλεία για την εκτίμηση πιστωτικού κινδύνου, η αξιοποίηση πιο προηγμένων μεθόδων μπορεί να βελτιστοποιήσει σημαντικά την προβλεπτική ακρίβεια και την επιχειρησιακή αξία στο πλαίσιο του credit risk management.

Στο παράδειγμα του Lending Club, η σύγκριση ανάμεσα στη λογιστική παλινδρόμηση και τα δέντρα αποφάσεων αναδεικνύει τη σημασία της επιλογής μεθοδολογίας ανάλογα με την πολυπλοκότητα και το μέγεθος των δεδομένων. Και οι δύο μέθοδοι προσφέρουν διαφάνεια και ερμηνευσιμότητα, στοιχεία σημαντικά για τον χρηματοπιστωτικό τομέα. Ωστόσο, σε μεγάλα και σύνθετα datasets, τα δέντρα αποφάσεων και οι παραλλαγές τους, όπως το LightGBM, επιτυγχάνουν σημαντικά υψηλότερη ακρίβεια και δυνατότητα ανίχνευσης μη γραμμικών συσχετίσεων. Η μελέτη αυτή υποστηρίζει το βασικό συμπέρασμα ότι η επιλογή ανάμεσα σε λογιστική παλινδρόμηση και δέντρα αποφάσεων εξαρτάται από το εκάστοτε επιχειρηματικό πλαίσιο και τα χαρακτηριστικά των δεδομένων, με τα δέντρα να υπερέχουν σε περιβάλλοντα υψηλής πολυπλοκότητας, ενώ η λογιστική παλινδρόμηση διατηρεί την αξία της για τη διαφάνεια και την απλότητα των αποφάσεων.

ΚΕΦΑΛΑΙΟ 6: ΣΥΜΠΕΡΑΣΜΑΤΑ

6.1 Ανακεφαλαίωση

Η διαχείριση και η ορθή αξιολόγηση του χρηματοοικονομικού κινδύνου αποτελούν θεμελιώδη ζητήματα για τη βιωσιμότητα και τη στρατηγική ανάπτυξη των σύγχρονων επιχειρήσεων. Στο παρόν έργο εξετάστηκαν εις βάθος δύο από τις πιο διαδεδομένες και επιστημονικά τεκμηριωμένες τεχνικές ανάλυσης δεδομένων, η Λογιστική Παλινδρόμηση και τα Δέντρα Αποφάσεων, σε πραγματικά χρηματοοικονομικά δεδομένα. Η μελέτη στόχευσε αφενός στην αποτύπωση των δυνατοτήτων, των περιορισμών και των εφαρμογών των μεθόδων αυτών, αφετέρου στη διατύπωση πρακτικών προτάσεων για την επιλογή του κατάλληλου εργαλείου σε διαφορετικά επιχειρηματικά πλαίσια.

Η ενδελεχής θεωρητική προσέγγιση και η συστηματική εμπειρική ανάλυση που πραγματοποιήθηκαν ανέδειξαν σημαντικά συμπεράσματα τόσο σε μεθοδολογικό όσο και σε πρακτικό επίπεδο. Η σημασία της ανάλυσης δεδομένων στη σύγχρονη διαχείριση χρηματοοικονομικών κινδύνων είναι αναμφισβήτητη, καθώς οι εξελιγμένες τεχνικές επιτρέπουν την αξιοποίηση πολύπλοκων πληροφοριών που παραδοσιακά θα παρέμεναν ανεκμετάλλευτες ή θα οδηγούσαν σε εσφαλμένες εκτιμήσεις. Η σύγκριση μεταξύ λογιστικής παλινδρόμησης και δέντρων αποφάσεων, υπό το πρίσμα πραγματικών δεδομένων, ανέδειξε τη σχετική υπεροχή της κάθε μεθόδου υπό συγκεκριμένες συνθήκες, επιβεβαιώνοντας τη σημασία της πολυκριτηριακής αξιολόγησης στην επιλογή εργαλείων για τη διαχείριση κινδύνου.

Αρχικά, η λογιστική παλινδρόμηση απέδειξε την αξία της ως ένα απλό, διαφανές και ιδιαίτερα ερμηνεύσιμο μοντέλο. Τα πλεονεκτήματά της έγκεινται στην ευκολία εφαρμογής, τη σαφήνεια των αποτελεσμάτων και την αποδοχή της από ρυθμιστικές αρχές και επαγγελματίες του χώρου. Η ικανότητά της να παρέχει ποσοτικές εκτιμήσεις πιθανοτήτων, μέσω ενός γραμμικού πλαισίου, καθιστά το μοντέλο ιδανικό για περιβάλλοντα όπου η ερμηνευσιμότητα και η διαφάνεια αποτελούν προτεραιότητα. Ωστόσο, ορισμένοι εγγενείς περιορισμοί, όπως η απαίτηση γραμμικότητας μεταξύ των μεταβλητών και η αδυναμία αποτύπωσης περίπλοκων αλληλεπιδράσεων, περιορίζουν την απόδοσή της σε ιδιαίτερα ετερογενή και μη γραμμικά δεδομένα.

Αντίστοιχα, τα δέντρα αποφάσεων ανέδειξαν την ευελιξία και τη δύναμή τους στην αποκάλυψη σύνθετων, μη γραμμικών συσχετίσεων και αλληλεπιδράσεων μεταξύ των μεταβλητών. Η οπτικοποιήσιμη και δομικά κατανοητή φύση τους καθιστά τα δέντρα αποφάσεων ιδανικά για επιχειρησιακά περιβάλλοντα όπου η εξήγηση των αποφάσεων προς μη τεχνικά στελέχη ή ρυθμιστικές αρχές είναι απαραίτητη. Επιπλέον, η ικανότητα των δέντρων να αντιμετωπίζουν φυσικά τόσο κατηγορικές όσο και συνεχείς μεταβλητές, καθώς και η ανθεκτικότητά τους σε δεδομένα με ανισορροπία μεταξύ των κατηγοριών, αποτελεί ιδιαίτερο πλεονέκτημα. Παρ’ όλα αυτά, παρατηρήθηκε μια σαφής τάση προς το φαινόμενο της υπερπροσαρμογής (overfitting), ειδικά σε μεγάλα και πολύπλοκα δέντρα, γεγονός που απαιτεί προσεκτική ρύθμιση παραμέτρων ή χρήση τεχνικών όπως το pruning και τα σύνολα δέντρων (ensembles).

Από την ανάλυση των αποτελεσμάτων, διαπιστώθηκε ότι καμία μέθοδος δεν αποτελεί πανάκεια για όλα τα επιχειρηματικά προβλήματα. Η λογιστική παλινδρόμηση αποδίδει εξαιρετικά όταν οι συσχετίσεις μεταξύ μεταβλητών και της εξαρτημένης μεταβλητής είναι κατά βάση γραμμικές και όταν το ζητούμενο είναι η εξαγωγή ερμηνεύσιμων συμπερασμάτων για τη λήψη αποφάσεων. Από την άλλη πλευρά, τα δέντρα αποφάσεων υπερτερούν σε περιπτώσεις όπου υφίστανται πολύπλοκες αλληλεπιδράσεις, μη γραμμικότητες ή η ανάγκη για εύκολη επικοινωνία των αποφάσεων προς μη εξειδικευμένα κοινά. Σημαντική είναι και η διαπίστωση ότι η συνδυαστική χρήση των δύο μεθόδων, στο πλαίσιο υβριδικών ή ensemble τεχνικών (π.χ. Random Forests, Gradient Boosting), μπορεί να ενισχύσει την προβλεπτική ισχύ, μειώνοντας ταυτόχρονα τον κίνδυνο υπερπροσαρμογής και μεροληψίας.

Στο πρακτικό πεδίο, η εφαρμογή των παραπάνω τεχνικών σε πραγματικά χρηματοοικονομικά δεδομένα κατέδειξε τη σημασία της διεξοδικής προεπεξεργασίας των δεδομένων (data preprocessing), της αντιμετώπισης ανισορροπίας κατηγοριών μέσω τεχνικών όπως το SMOTE, και της αξιολόγησης των μοντέλων με πολλαπλές μετρικές (accuracy, precision, recall, F1-score, AUC-ROC). Η ανάγκη για ολιστική και πολυδιάστατη αξιολόγηση καθίσταται επιτακτική, ιδιαίτερα σε προβλήματα με ασύμμετρο κόστος λαθών, όπως το πιστωτικό ρίσκο, όπου το κόστος ενός false negative (μη ανίχνευση κινδύνου) μπορεί να αποβεί καταστροφικό για τον οργανισμό. Παράλληλα, η ανάλυση σφαλμάτων και η μελέτη των περιορισμών των μοντέλων αποτελούν ουσιώδη συστατικά για την περαιτέρω βελτιστοποίηση και αποτελεσματική εφαρμογή των τεχνικών αυτών στην πράξη.

Πέραν της συγκριτικής ανάλυσης των δύο τεχνικών, η εργασία ανέδειξε τη σημασία των δεδομένων ως στρατηγικού πλεονεκτήματος. Η δυνατότητα συλλογής, επεξεργασίας και αξιοποίησης μεγάλων όγκων ετερογενούς πληροφορίας μετασχηματίζει τη διαδικασία διαχείρισης κινδύνων από στατική και αντιδραστική σε δυναμική και προληπτική. Σε αυτό το πλαίσιο, η ανάλυση δεδομένων λειτουργεί ως πολλαπλασιαστής ισχύος, προσφέροντας εργαλεία για τον εντοπισμό κρυφών μοτίβων, την αναγνώριση νέων απειλών και τη βελτιστοποίηση των στρατηγικών αποφάσεων. Ωστόσο, ο μετασχηματισμός αυτός απαιτεί καινοτόμες μεθόδους ανάλυσης, τεχνολογική επάρκεια και συνεχή εκπαίδευση του ανθρώπινου δυναμικού.

6.2 Μελλοντική Έρευνα

Παρά τη σημαντική πρόοδο που έχει σημειωθεί στην ανάλυση και αξιολόγηση χρηματοοικονομικών κινδύνων με τη χρήση μεθόδων ανάλυσης δεδομένων, το πεδίο εξακολουθεί να προσφέρει πλούσιες ευκαιρίες για περαιτέρω διερεύνηση και ανάπτυξη. Τα αποτελέσματα της παρούσας μελέτης ανέδειξαν τόσο τις δυνατότητες όσο και τους περιορισμούς των παραδοσιακών και σύγχρονων τεχνικών, γεγονός που υποδεικνύει την ανάγκη για στοχευμένη ερευνητική προσπάθεια σε συγκεκριμένες κατευθύνσεις.

Μία βασική κατεύθυνση αφορά τη διεύρυνση του φάσματος των μεθόδων μηχανικής μάθησης που εφαρμόζονται στην πρόβλεψη και διαχείριση χρηματοοικονομικών κινδύνων. Ενώ η παρούσα εργασία επικεντρώθηκε στη λογιστική παλινδρόμηση και στα δέντρα αποφάσεων, μέσω των μελετών περίπτωσης αναδεικνύεται η δυναμική πιο σύνθετων αλγορίθμων, όπως τα τυχαία δάση (Random Forests), οι αλγόριθμοι ενίσχυσης βαθμίδωσης (Gradient Boosting Machines), τα βαθιά νευρωνικά δίκτυα και τα σύγχρονα υβριδικά μοντέλα (ensemble methods). Η συγκριτική ανάλυση αυτών των τεχνικών, με στόχο τόσο τη βελτιστοποίηση της ακρίβειας όσο και τη διασφάλιση της ερμηνευσιμότητας, αποτελεί κρίσιμο ερευνητικό ζητούμενο.

Επιπλέον, αξίζει να διερευνηθεί η εφαρμογή τεχνικών ανάλυσης δεδομένων σε πιο ετερογενή και πολυδιάστατα σύνολα δεδομένων, που προέρχονται από διαφορετικά γεωγραφικά και θεσμικά περιβάλλοντα. Η χρήση πολυετών και διακρατικών δεδομένων μπορεί να αποκαλύψει νέα μοτίβα και παραμέτρους κινδύνου, αναδεικνύοντας τη σημασία του μακροοικονομικού, θεσμικού και πολιτισμικού πλαισίου στη διαμόρφωση του χρηματοοικονομικού ρίσκου. Η ανάλυση της διαχρονικής σταθερότητας των μοντέλων και η αξιολόγηση της μεταφερσιμότητάς τους (transferability) μεταξύ διαφορετικών αγορών και περιόδων κρίσης αποτελούν πεδία με ιδιαίτερο ερευνητικό ενδιαφέρον.

Ένα ακόμη σημείο προς διερεύνηση είναι η ενσωμάτωση μη δομημένων δεδομένων (π.χ. κειμένων από οικονομικές εκθέσεις, κοινωνικών δικτύων ή ειδήσεων) στις μεθοδολογίες πρόβλεψης κινδύνου. Οι τεχνικές επεξεργασίας φυσικής γλώσσας (NLP) και τα μοντέλα βαθιάς μάθησης ανοίγουν νέες δυνατότητες για τον εντοπισμό πρόδρομων ενδείξεων κινδύνου πέρα από τα παραδοσιακά αριθμητικά δεδομένα.

Τέλος, η μελέτη του κόστους σφαλμάτων και της αλληλεπίδρασης με τις επιχειρησιακές διαδικασίες αποτελεί μια κρίσιμη κατεύθυνση για το μέλλον. Η ανάπτυξη cost-sensitive ή risk-adjusted μοντέλων, τα οποία λαμβάνουν υπόψη το πραγματικό οικονομικό αντίκτυπο των λαθών πρόβλεψης, μπορεί να οδηγήσει σε πιο αποτελεσματικά και ρεαλιστικά εργαλεία διαχείρισης κινδύνου.

Συνοψίζοντας, το παρόν έργο κατέδειξε ότι η αξιολόγηση χρηματοοικονομικού κινδύνου μέσω ανάλυσης δεδομένων συνιστά ένα πολύπλοκο, δυναμικό και διαρκώς εξελισσόμενο πεδίο, όπου η επιτυχία δεν εξαρτάται μόνο από την τεχνολογική υπεροχή αλλά και από τη στρατηγική ευελιξία, την ερμηνευσιμότητα και τη συνεχή αναζήτηση ισορροπίας μεταξύ ακρίβειας και λογοδοσίας. Τα αποτελέσματα υπογραμμίζουν τη σημασία της πολυδιάστατης προσέγγισης, της καινοτομίας και της διεπιστημονικότητας στη διαχείριση χρηματοοικονομικών κινδύνων, προσφέροντας ένα γόνιμο έδαφος για περαιτέρω ακαδημαϊκή και επαγγελματική διερεύνηση.

Βιβλιογραφία

Adeyemo, A., Wimmer, H., & Powell, L. (2018). Effects of normalization techniques on logistic regression in data science. In Proceedings of the Conference on Information Systems Applied Research ISSN (Vol. 2167, p. 1508).

Biecek, P., Chlebus, M., Gajda, J., Gosiewska, A., Kozak, A., Ogonowski, D., Sztachelski, J., & Wojewnik, P. (2021). Enabling Machine Learning Algorithms for Credit Scoring -- Explainable Artificial Intelligence (XAI) methods for clear understanding complex predictive models.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446–3453.

Chang, V., Sivakulasingam, S., Wang, H., Wong, S. T., Ganatra, M. A., & Luo, J. (2024). Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers. Risks, 12(11), 174.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Chen, N., Ribeiro, B., & Chen, A. (2016). Financial credit risk assessment: a recent review. Artificial Intelligence Review, 45, 1-23.

Conroy, B., & Sajda, P. (2012). Fast, exact model selection and permutation testing for l2-regularized logistic regression. In Artificial Intelligence and Statistics (pp. 246-254). PMLR.

Cornwell, N., Bilson, C. M., Gepp, A., Stern, S., & Vanstone, B. J. (2023). The role of data analytics within operational risk management: A systematic review from the financial services and energy sectors. Journal of the Operational Research Society, Bangor University.

Costa-e-Silva, E., & Lopes, I. C. (2020). A logistic regression model for consumer default risk. Journal of Applied Statistics.

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research, 297(3), 1178-1192.

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). Learning from imbalanced data sets (Vol. 10, No. 2018). Cham: Springer.

Fejza, D., Nace, D., & Kulla, O. (2022). The Credit Risk Problem—A Developing Country Case Study. Risks, 10(8), 146.

Foglia, A. (2009). Stress testing credit risk: A survey of authorities' approaches. Banking and Financial Supervision, Bank of Italy.

Gopinath, M., Maheep, S. S., & Sethuraman, R. (2021). Customer loan approval prediction using logistic regression. Advances in Parallel Computing, 38, 563-569.

Guan, X., & Burton, H. (2022, December). Bias-variance tradeoff in machine learning: Theoretical formulation and implications to structural engineering applications. In Structures (Vol. 46, pp. 17-30). Elsevier.

Halabaku, E., & Bytyçi, E. (2024). Overfitting in Machine Learning: A Comparative Analysis of Decision Trees and Random Forests. Intelligent Automation & Soft Computing, 39(6).

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific reports, 12(1), 5979.

Hull, J. C. (2018). Risk management and financial institutions (5th ed.). Wiley

Jorion, P. (2020). Financial risk manager handbook (7th ed.). Wiley

Jun, Y., & Liu, H. (2022). A decision tree algorithm for financial risk data of small and medium-sized enterprises. International Journal of Economics and Statistics.

Ko, P. C., Lin, P. C., Do, H. T., & Huang, Y. F. (2022). P2P Lending Default Prediction Based on AI and Statistical Models. Entropy (Basel, Switzerland), 24(6), 801.

Koc, O., Ugur, O., & Kestel, A. S. (2023). The Impact of Feature Selection and Transformation on Machine Learning Methods in Determining the Credit Scoring. arXiv.

Komorowski, M., Marshall, D.C., Salciccioli, J.D., Crutain, Y. (2016). Exploratory Data Analysis. Secondary Analysis of Electronic Health Records. Springer, Cham.

Kotu, V. & Deshpande, B., (2019). Data Science Concepts and Practice. 2nd ed. Cambridge: Morgan Kaufmann

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. In Data democracy (pp. 83-106). Academic Press.

Kwak S.K., Kim J.H. (2017). “Statistical data preparation: management of missing values and outliers.” Korean Journal of Anesthesiology.

Lee, V. E., Liu, L., & Jin, R. (2014). Decision trees: Theory and algorithms. In Data Classification (pp. 115-148). Chapman and Hall/CRC.

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124-136.

Li, Y., & Chen, W. (2020). A Comparative Performance Assessment of Ensemble Learning for Credit Scoring. Mathematics, 8(10), 1756.

Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. Expert Systems with Applications, 165, 113986.

Rockfeller, T., & Uryasev, S. (2002). Conditional value-at-risk for general loss distribution. Journal of BANKING and FINANCE

Roy, J., & Vasa, L. (2025). Transforming credit risk assessment: A systematic review of AI and machine learning applications. Journal of Infrastructure, Policy and Development

Rudd, J. M., & Priestley, J. L. (2017). A Comparison of Decision Tree with Logistic Regression Model for Prediction of Worst Non-Financial Payment Status in Commercial Credit. Kennesaw State University.

Saunders, A. and Cornett, M.M. (2011) Financial Institution Management—A Risk Management Approach. 9th Edition, McGraw Hill Irwin, New York

Sen, H., Wu, H., Xu, X., & Xiong, W. (2022). Early warning of enterprise financial risk based on decision tree algorithm. Computational Intelligence and Neuroscience, 2022, Article 9182099.

Seger, C. (2018). An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing.

Suguna, R., Suriya Prakash, J., Aditya Pai, H., Mahesh, T. R., Vinoth Kumar, V., & Yimer, T. E. (2025). Mitigating class imbalance in churn prediction with ensemble methods and SMOTE. Scientific Reports, 15(1), 1-20.

Victor, L., & Raheem, M. (2021). Loan default prediction using Genetic Algorithm: A study within peer-to-peer lending communities. Int. J. Innovative Sci. Res. Technol, 6(3), 1195-1205.

Vidovic, L., & Yue, L. (2020). Machine learning and credit risk modelling (S&P Global Market Intelligence White Paper). Vidovic & Yue, S&P.

Zhang, Z. (2016). Missing data imputation: focusing on single imputation. Annals of translational medicine, 4(1), 9.

Παράρτημα Α: Πλήρης κώδικας σε Python

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_csv(r'C:\Users\USER\Desktop\Loan_default.csv')

# 1. Προβολή των πρώτων γραμμών
print("Πρώτες 5 γραμμές του DataFrame:")
print(df.head())

# 2. Στατιστικά στοιχεία για αριθμητικές στήλες
print("\nΣτατιστικά Στοιχεία (describe) για αριθμητικές στήλες:")
print(df.describe())

# 3. Έλεγχος για NaN τιμές (Missing Values)
print("\nΑριθμός NaN τιμών ανά στήλη:")
print(df.isnull().sum())

# 4. Ποσοστό NaN τιμών
print("\nΠοσοστό NaN τιμών ανά στήλη:")
print(df.isnull().mean())

# 5. Διανομή κατηγορικών μεταβλητών
print("\nΔιανομή των κατηγορικών στηλών:")
for col in df.select_dtypes(include=['object', 'category']).columns:
    print(f"\nΣτήλη: {col}")
    print(df[col].value_counts())

# 6. Εξέταση της κατανομής των αριθμητικών δεδομένων (ιστογράμματα)
print("\nΚατανομή των αριθμητικών στηλών (ιστογράμματα):")
df.select_dtypes(include=[np.number]).hist(figsize=(12, 8), bins=30)
plt.tight_layout()
plt.show()
```

```python
# 8. Συσχέτιση των αριθμητικών στηλών (Εξαίρεση μη αριθμητικών στηλών όπως LoanID)
numeric_cols = df.select_dtypes(include=[np.number]).columns
print("\nΠίνακας συσχέτισης των αριθμητικών στηλών:")
print(df[numeric_cols].corr())

# 9. Εξέταση της κατανομής του target
print("\nΚατανομή του στόχου 'Default':")
print(df['Default'].value_counts())
```

```python
import pandas as pd
import numpy as np
import sys
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
import matplotlib.pyplot as plt
from imblearn.over_sampling import SMOTE
from sklearn.metrics import roc_curve, auc
from sklearn.metrics import roc_auc_score


def preprocess(df):
    # 1. συμπλήρωση numeric με mean
    num_cols = df.select_dtypes(include=[np.number]).columns
    df[num_cols] = df[num_cols].fillna(df[num_cols].mean())

    # 2. συμπλήρωση categorical με mode
    cat_cols = df.select_dtypes(include=['object', 'category']).columns
    for col in cat_cols:
        df[col] = df[col].fillna(df[col].mode()[0])

    # 3. αφαίρεση ID column
    if 'LoanID' in df.columns:
        df.drop('LoanID', axis=1, inplace=True)

    # 4. διαχωρισμός δεδομένων X/y
    X = df.drop('Default', axis=1)
    y = df['Default']

    # 5. One-hot encode categoricals
    X = pd.get_dummies(X, drop_first=True)
    return X, y

def split_and_scale(X, y, scale=True):
    X_train, X_test, y_train, y_test = train_test_split(
        X, y, test_size=0.2, random_state=42, stratify=y
    )
    if scale:
        scaler = StandardScaler()
        X_train = scaler.fit_transform(X_train)
        X_test  = scaler.transform(X_test)
    return X_train, X_test, y_train, y_test

def apply_smote(X_train, y_train):
    # Εφαρμογή SMOTE για oversampling της κλάσης 1
    smote = SMOTE(random_state=42)
    X_train_smote, y_train_smote = smote.fit_resample(X_train, y_train)
    print(f"Resampled dataset shape: {y_train_smote.value_counts()}")
    return X_train_smote, y_train_smote

def train_logistic(X_train, y_train):
    model = LogisticRegression(max_iter=1000, solver='lbfgs', random_state=42)
    model.fit(X_train, y_train)
    return model
```

```python
def train_tree(X_train, y_train):
    model = DecisionTreeClassifier(
        criterion='entropy',
        max_depth=5,
        min_samples_split=10,
        random_state=42
    )
    model.fit(X_train, y_train)
    return model

def evaluate(model, X_test, y_test, name):
    y_pred = model.predict(X_test)
    print(f"\n=== {name} ===")
    print(f"Accuracy: {accuracy_score(y_test, y_pred):.4f}")
    print("Classification Report:")
    print(classification_report(y_test, y_pred))
    print("Confusion Matrix:")
    print(confusion_matrix(y_test, y_pred))
```

```python
def plot_decision_tree_model(model, feature_names):
    plt.figure(figsize=(18, 12))
    plot_tree(
        model,
        feature_names=feature_names,
        class_names=True,
        filled=True,
        rounded=True,
        max_depth=3
    )
    plt.title("Decision Tree (partial view up to depth=3)")
    plt.show()

def plot_performance(lr_accuracy, dt_accuracy, lr_f1, dt_f1):
    models = ['Logistic Regression', 'Decision Tree']
    accuracies = [lr_accuracy, dt_accuracy]
    f1_scores = [lr_f1, dt_f1]

    fig, ax = plt.subplots(1, 2, figsize=(12, 6))

    # Bar plot για accuracy
    ax[0].bar(models, accuracies, color=['blue', 'green'])
    ax[0].set_title('Model Accuracy')
    ax[0].set_ylabel('Accuracy')

    # Bar plot για F1 Score (για κλάση 1)
    ax[1].bar(models, f1_scores, color=['blue', 'green'])
    ax[1].set_title('F1 Score for Class 1')
    ax[1].set_ylabel('F1 Score')

    plt.tight_layout()
    plt.show()
```

```python
def plot_roc_curve_with_auc(model, X_test, y_test, model_name):
    # Υπολογισμός των πιθανοτήτων για την κλάση 1 (default)
    y_prob = model.predict_proba(X_test)[:, 1]

    # Υπολογισμός της ROC Curve
    fpr, tpr, _ = roc_curve(y_test, y_prob)
    roc_auc = auc(fpr, tpr)

    # Οπτικοποίηση της ROC Curve με AUC
    plt.figure(figsize=(8, 6))
    plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC curve (area = %0.2f)' % roc_auc)
    plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
    plt.xlim([0.0, 1.0])
    plt.ylim([0.0, 1.05])
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title(f'Receiver Operating Characteristic ({model_name}) - AUC: {roc_auc:.2f}')
    plt.legend(loc='lower right')
    plt.show()
```

```python
def main():
    file_path = r'C:\Users\USER\Desktop\Loan_default.csv'

    print("1) Loading data...")
    df = pd.read_csv(file_path)
    print("   Data shape:", df.shape)
    print(df.head(), "\n")

    print("2) Preprocessing...")
    X, y = preprocess(df)
    print("   Features after encoding:", X.shape[1])

    print("3) Splitting & scaling...")
    X_train, X_test, y_train, y_test = split_and_scale(X, y)

    print("4) Applying SMOTE...")
    X_train_smote, y_train_smote = apply_smote(X_train, y_train)

    print("5) Training Logistic Regression...")
    lr_model = train_logistic(X_train_smote, y_train_smote)

    print("6) Training Decision Tree...")
    dt_model = train_tree(X_train_smote, y_train_smote)

    print("7) Evaluating models...")
    evaluate(lr_model, X_test, y_test, "Logistic Regression")
    evaluate(dt_model, X_test, y_test, "Decision Tree")

    print("8) Plotting Decision Tree...")
    plot_decision_tree_model(dt_model, feature_names=X.columns)
```

```python
# Εξαγωγή αποτελεσμάτων για bar chart και ROC curve
lr_accuracy = accuracy_score(y_test, lr_model.predict(X_test))
lr_f1 = classification_report(y_test, lr_model.predict(X_test), output_dict=True)['1']['f1-score']

dt_accuracy = accuracy_score(y_test, dt_model.predict(X_test))
dt_f1 = classification_report(y_test, dt_model.predict(X_test), output_dict=True)['1']['f1-score']

# Κλήση της συνάρτησης για bar chart
plot_performance(lr_accuracy, dt_accuracy, lr_f1, dt_f1)

# Κλήση της συνάρτησης για ROC Curve με AUC
plot_roc_curve_with_auc(lr_model, X_test, y_test, "Logistic Regression")
plot_roc_curve_with_auc(dt_model, X_test, y_test, "Decision Tree")
```