



ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ
ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΤΖΕΣΙ ΖΝΤΡΑΒΑ

**Ανάλυση Πιστωτικού Κινδύνου με Αλγορίθμους
Μηχανικής Μάθησης και Τεχνικές Explainable AI.
Έμφαση στη Διαφάνεια, Δικαιοσύνη και Μεροληψία**

Επιβλέπων : Αντρέας Ζάρας – Γεώργιος Λεκάκος

Υποβληθείσα ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος (MSc) στη Διοικητική Επιστήμη και Τεχνολογία

Αθήνα, Ιανουάριος 2026



Η σελίδα αυτή είναι σκόπιμα λευκή.





ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ
ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΤΖΕΣΙ ΖΝΤΡΑΒΑ

**Ανάλυση Πιστωτικού Κινδύνου με Αλγορίθμους
Μηχανικής Μάθησης και Τεχνικές Explainable AI.
Έμφαση στη Διαφάνεια, Δικαιοσύνη και Μεροληψία**

Επιβλέπων : Ανδρέας Ζάρας – Γεώργιος Λεκάκος

Υποβληθείσα ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος (MSc) στη Διοικητική Επιστήμη και Τεχνολογία

Αθήνα, Ιανουάριος 2026



Βεβαίωση εκπόνησης Διπλωματικής εργασίας

«Δηλώνω υπεύθυνα ότι η συγκεκριμένη μεταπτυχιακή εργασία για τη λήψη του μεταπτυχιακού τίτλου σπουδών του ΠΜΣ στη Διοικητική Επιστήμη και Τεχνολογία του Τμήματος Διοικητικής Επιστήμης και Τεχνολογίας του Οικονομικού Πανεπιστημίου Αθηνών έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει εγκριθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών στην Ελλάδα ή το εξωτερικό. Η εργασία αυτή έχοντας εκπονηθεί από εμένα, αντιπροσωπεύει τις προσωπικές μου απόψεις επί του θέματος. Οι πηγές στις οποίες ανέτρεξα για την εκπόνηση της συγκεκριμένης διπλωματικής αναφέρονται στο σύνολό τους, δίνοντας πλήρεις αναφορές στους συγγραφείς, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο».

(Υπογραφή)



.....
ZNPABA TZESI

Φοιτητής MSc στη Διοικητική Επιστήμη και Τεχνολογία



Περίληψη

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη και εφαρμογή μιας μεθοδολογίας ανάλυσης πιστωτικού κινδύνου με τη χρήση αλγορίθμων μηχανικής μάθησης και τεχνικών Explainable AI, με έμφαση στη διαφάνεια, τη δικαιοσύνη και την μεροληψία στις αλγοριθμικές αποφάσεις. Η ανάλυση πιστωτικού κινδύνου είναι μια κρίσιμη διαδικασία των πιστωτικών ιδρυμάτων, καθώς επηρεάζει άμεσα τη λήψη αποφάσεων δανεισμού και συνδέεται με αυξημένη απαίτηση λογοδοσίας και συμμόρφωσης με το κανονιστικό πλαίσιο.

Η μεθοδολογία εφαρμόστηκε στο German Credit Dataset και περιλαμβάνει τη συστηματική προεπεξεργασία των δεδομένων, την εκπαίδευση και σύγκριση πολλαπλών μοντέλων ταξινόμησης, καθώς και τη ρύθμιση υπερπαραμέτρων και την τελική αξιολόγηση σε ανεξάρτητο σύνολο ελέγχου. Η προβλεπτική απόδοση των μοντέλων αξιολογήθηκε με διάφορες μετρικές για την ανάλυση πιστωτικού κινδύνου. Επιπλέον πραγματοποιήθηκε και ανάλυση κατάταξης μέσω cumulative lift, ώστε να ελεγχθεί η ικανότητά τους να ιεραρχούνται οι δανειολήπτες με βάση τον εκτιμώμενο κίνδυνο.

Εφαρμόστηκαν τεχνικές Explainable AI, όπως οι SHAP και LIME για να εξεταστεί γιατί το μοντέλο πήρε μια απόφαση και για να κατανοήσουμε τους παράγοντες που επηρεάζουν τις αποφάσεις του. Ζητήματα δικαιοσύνης και πιθανής μεροληψίας εξετάστηκαν με ερμηνευτικές τεχνικές και με ποσοτικές μετρικές fairness, δίνοντας έμφαση στον ρόλο ευαίσθητων δημογραφικών χαρακτηριστικών και στη σχέση τους με τις προβλέψεις των μοντέλων.

Αυτή η έρευνα είναι σημαντική διότι προσφέρει μια συνοπτική αλλά ολοκληρωμένη εικόνα των τεχνικών και προκλήσεων που σχετίζονται με την ανάλυση πιστωτικού κινδύνου, ζήτημα που είναι ιδιαίτερα επίκαιρο στον χρηματοπιστωτικό τομέα. Στην εργασία συνδυάζονται η προβλεπτική απόδοση με τη διαφάνεια και την ερμηνευσιμότητα των μοντέλων, με σκοπό να καλυφθεί η αυξανόμενη ανάγκη για υπεύθυνη χρήση αλγοριθμικών συστημάτων από τα πιστωτικά ιδρύματα. Ακόμη, οι πρόσφατες κανονιστικές εξελίξεις, όπως η εισαγωγή του AI Act και του GDPR, έχουν αναδείξει την κρίσιμη ανάγκη επεξηγησιμότητας και λογοδοσίας στην εφαρμογή τεχνικών μηχανικής μάθησης στην πιστοληπτική αξιολόγηση.

Λέξεις Κλειδιά: << Ανάλυση Πιστωτικού Κινδύνου, Μηχανική Μάθηση, Explainable AI, Διαφάνεια, Δικαιοσύνη, Μεροληψία >>



Η σελίδα αυτή είναι σκόπιμα λευκή.



Abstract

The scope of this thesis is the development and application of a credit risk analysis methodology using Machine Learning algorithms and Explainable AI techniques, with an emphasis on transparency, fairness, and bias in algorithmic decision-making. Credit risk analysis is a critical process for financial institutions, as it directly affects lending decisions and is associated with increased requirements for accountability and compliance with the regulatory framework.

The proposed methodology was applied to the German Credit Dataset and includes systematic data preprocessing, the training and comparison of multiple classification models, as well as hyperparameter tuning and final evaluation on an independent test set. The predictive performance of the models was assessed using various metrics appropriate for credit risk analysis. In addition, a ranking analysis using cumulative lift was conducted in order to evaluate the models' ability to rank borrowers according to their estimated risk.

Explainable AI techniques, such as SHAP and LIME, are employed to investigate the rationale behind model decisions and to gain insights into the factors that influence those decisions. Issues of fairness and potential bias were examined using both interpretability techniques and quantitative fairness metrics, with emphasis on the role of sensitive demographic attributes and their relationship to the models' predictions.

This research is significant as it provides a concise yet comprehensive overview of the techniques and challenges associated with credit risk analysis, a topic of growing importance in the financial sector. The study combines predictive performance with model transparency and interpretability, addressing the increasing demand for the responsible use of algorithmic systems by credit institutions. Furthermore, recent regulatory developments, such as the introduction of the AI Act and the GDPR, have highlighted the critical need for explainability and accountability in the application of machine learning techniques to credit scoring.

Keywords: << Credit Risk Analysis, Machine Learning, Explainable AI, Transparency, Fairness, Bias>>

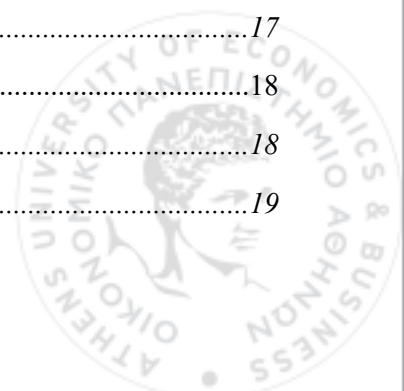


Η σελίδα αυτή είναι σκόπιμα λευκή.

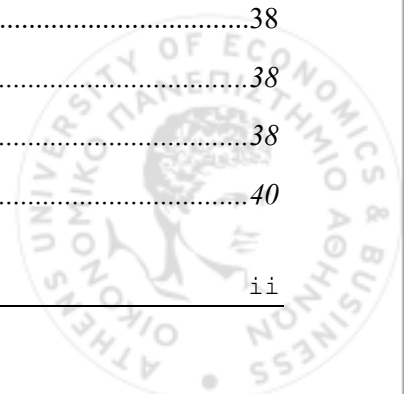


Πίνακας περιεχομένων

1	Εισαγωγή.....	1
1.1	Παρουσίαση του προβλήματος του πιστωτικού κινδύνου και της σημασίας του στον χρηματοπιστωτικό τομέα.....	1
1.2	Αντικείμενο διπλωματικής	2
1.2.1	Συνεισφορά.....	2
1.3	Οργάνωση κειμένου.....	3
2	Βιβλιογραφική Ανασκόπηση και Σχετικές εργασίες	4
2.1	Κανονιστικό και Θεσμικό Πλαίσιο Πιστωτικού Κινδύνου και Διαφάνειας.....	4
2.1.1	Χρονική επισκόπηση των συμφωνιών της Βασιλείας	4
2.1.2	Οι υποχρεώσεις των πιστωτικών ιδρυμάτων για διαφάνεια και επεξηγησιμότητα όπως προκύπτουν από το GDPR και το AI Act.....	6
2.2	Σχετικές εργασίες σε Μοντέλα Μηχανικής Μάθησης για Πιστωτικό Κίνδυνο: Δικαιοσύνη και Επεξηγησιμότητα	9
2.2.1	Μεροληψία και Δικαιοσύνη.....	9
2.2.2	Εξηγησιμότητα (XAI)	11
2.2.3	Προχωρημένες Τεχνικές Τοπικών Εξηγήσεων και Counterfactuals	13
3	Θεωρητικό υπόβαθρο.....	14
3.1	Machine Learning για Credit Scoring.....	14
3.2	Μοντέλα Machine Learning	15
3.2.1	Logistic regression	15
3.2.2	Decision Tree	15
3.2.3	Random Forest.....	16
3.2.4	Extreme Gradient Boosting (XGBoost)	16
3.2.5	Support Vector Machines (SVM).....	17
3.2.6	Artificial Neural Network (ANN).....	17
3.3	Αξιολόγηση Μοντέλων.....	18
3.3.1	Confusion Matrix.....	18
3.3.2	Accuracy.....	19



3.3.3	<i>Recall</i>	19
3.3.4	<i>Precision</i>	19
3.3.5	<i>F1-score</i>	19
3.3.6	<i>ROC (Receiver Operating Characteristic) Curves και AUC</i>	19
3.4	Μετρικές Δικαιοσύνης και Διαφάνειας.....	20
3.4.1	<i>Disparate Impact Ratio (DIR)</i>	20
3.4.2	<i>Equal Opportunity Difference (EOD)</i>	20
3.4.3	<i>Equalized odds (EOdds)</i>	21
3.4.4	<i>Statistical Parity Difference (SPD)</i>	21
3.5	Αντιμετώπιση Μεροληψίας (Bias mitigation).....	21
3.5.1	<i>Pre-processing τεχνικές</i>	23
3.5.2	<i>In-processing τεχνικές</i>	23
3.5.3	<i>Post-processing τεχνικές</i>	24
3.6	Proxy μεταβλητές.....	25
3.7	Explainable AI (XAI).....	26
3.7.1	<i>Partial Dependence Plots (PDP)</i>	27
3.7.2	<i>Individual Conditional Expectation (ICE)</i>	28
3.7.3	<i>Shapley Additive exPlanations (SHAP)</i>	28
3.7.4	<i>Local Interpretable Model-agnostic Explanations (LIME)</i>	29
3.7.5	<i>Surrogate models</i>	29
3.7.6	<i>Counterfactual explanations</i>	30
4	Μεθοδολογία.....	31
4.1	Δεδομένα.....	31
4.2	Προεπεξεργασία Δεδομένων.....	31
5	Αξιολόγηση.....	36
5.1	Παράμετροι αξιολόγησης.....	36
5.2	Σύστημα αξιολόγησης.....	36
5.3	Αποτελέσματα.....	38
5.3.1	<i>Fairness analysis στα δεδομένα</i>	38
5.3.2	<i>Συγκριτική αξιολόγηση απόδοσης μοντέλων</i>	38
5.3.3	<i>Καμπύλες ROC</i>	40



5.3.4	<i>Cumulative Lift Analysis</i>	41
5.3.5	<i>Explainability</i>	42
5.3.6	<i>Fairness Analysis στο test set</i>	50
5.3.7	<i>Post-processing Bias Mitigation για την ηλικία</i>	51
6	Επίλογος	53
6.1	Σύνοψη και συμπεράσματα	53
6.2	Μελλοντικές επεκτάσεις.....	55
7	Βιβλιογραφία	56



1 *Εισαγωγή*

1.1 Παρουσίαση του προβλήματος του πιστωτικού κινδύνου και της σημασίας του στον χρηματοπιστωτικό τομέα

Η χορήγηση πιστώσεων αποτελεί βασική λειτουργία του τραπεζικού συστήματος. Η δραστηριότητα των χρηματοοικονομικών ιδρυμάτων στηρίζεται στη χορήγηση δανείων σε επιχειρήσεις και ιδιώτες, οι οποίοι υποχρεούνται να αποπληρώσουν το δάνειο και τους τόκους. Όμως, η παροχή δανείων συνεπάγεται τον κίνδυνο αθέτησης πληρωμών από τον οφειλέτη. Ο κίνδυνος αυτός, γνωστός ως πιστωτικός κίνδυνος (credit risk) αποτελεί έναν από τους σημαντικότερους κινδύνους στον τραπεζικό τομέα (Munkhdalai et al.). Δεδομένου ότι ο πιστωτικός κίνδυνος μπορεί να απειλήσει τη χρηματοπιστωτική σταθερότητα, απαιτείται η συνεχής παρακολούθηση και αποτελεσματική διαχείρισή του στο πλαίσιο της ευρύτερης διαχείρισης χρηματοοικονομικών κινδύνων (André Aoun Montevechi et al.).

Η αποτελεσματική διαχείριση του πιστωτικού κινδύνου επιτρέπει στα πιστωτικά ιδρύματα να διατηρούν τη χρηματοοικονομική τους σταθερότητα και τη μακροχρόνια βιωσιμότητά τους, συμβάλλοντας παράλληλα στη συνολική σταθερότητα του πιστωτικού συστήματος. Επομένως, η ακριβής αξιολόγηση του πιστωτικού κινδύνου αποτελεί κρίσιμο στοιχείο της διαχείρισης χρηματοοικονομικών κινδύνων (André Aoun Montevechi et al.).

Η αξιολόγηση του πιστωτικού κινδύνου στις χρηματοοικονομικές δραστηριότητες συνήθως προσεγγίζεται ως ένα δυαδικό πρόβλημα ταξινόμησης, με βάση το αν ο δανειολήπτης αποπληρώνει ή όχι το χρέος του. Η μεταβλητή της αποπληρωμής θεωρείται διχοτομική: τα δάνεια που έχουν εξοφληθεί πλήρως κωδικοποιούνται ως «0», ενώ εκείνα που βρίσκονται σε αθέτηση ως «1» (Moscatto et al.). Άτομα, και επιχειρήσεις μπορούν να ζητήσουν πίστωση για διάφορους σκοπούς, όπως η αγορά εξοπλισμού, ακινήτων ή καταναλωτικών αγαθών, χρησιμοποιώντας μέσα όπως πιστωτικές κάρτες,

δάνεια ή προγράμματα πληρωμής με πίστωση (Chang et al.). Από την πλευρά τους οι τράπεζες παρέχουν τη χρηματοδότηση αναμένοντας την εμπρόθεσμη αποπληρωμή του ποσού, με το εκάστοτε επιτόκιο που έχει οριστεί σε κάθε περίπτωση και αντανακλά την πιθανότητα μη αποπληρωμής.

1.2 Αντικείμενο διπλωματικής

Η παρούσα διπλωματική εξετάζει την ανάλυση πιστωτικού κινδύνου χρησιμοποιώντας αλγόριθμους μηχανικής μάθησης σε συνδυασμό με τεχνικές Explainable AI (XAI), δίνοντας έμφαση στη διαφάνεια, τη δικαιοσύνη και την μεροληψία στις αλγοριθμικές αποφάσεις. Η ανάλυση πιστωτικού κινδύνου είναι σημαντική στον χρηματοπιστωτικό τομέα καθώς επηρεάζει τις αποφάσεις σχετικά με τον δανεισμό, και υπόκειται σε έλεγχο και κανονιστικές ρυθμίσεις.

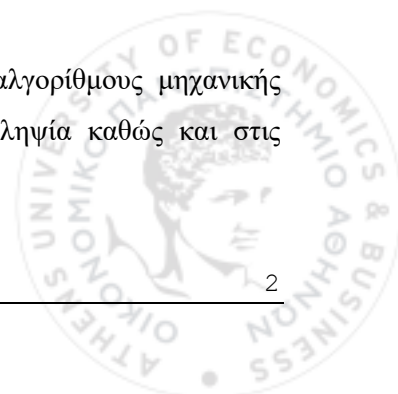
Στόχος της διπλωματικής είναι να εξεταστεί με ποιον τρόπο τα σύγχρονα μοντέλα μηχανικής μάθησης είναι σε θέση να προβλέψουν τον πιστωτικό κίνδυνο χωρίς οι αποφάσεις των μοντέλων να είναι αδιαφανείς και δύσκολα ερμηνεύσιμες. Διερευνάται επίσης εάν αυτά τα μοντέλα ενδέχεται να οδηγήσουν σε μεροληπτικές προβλέψεις ή σε άνιση μεταχείριση διαφορετικών ομάδων δανειοληπτών, ένα πρόβλημα που συνδέεται άμεσα με την έννοια της δικαιοσύνης στις αυτοματοποιημένες αποφάσεις.

Για την επίλυση των παραπάνω ζητημάτων, εκτελούνται και συγκρίνεται η απόδοση διάφορων αλγορίθμων ταξινόμησης που μπορούν να εφαρμοστούν σε τέτοια προβλήματα. Τα μοντέλα εκπαιδεύονται και αξιολογούνται σε ένα κοινό σύνολο δεδομένων, με στόχο τη σύγκριση της προβλεπτικής τους ικανότητας. Δίνεται ιδιαίτερη έμφαση στην εφαρμογή μεθόδων Explainable AI, όπως το SHAP και το LIME, οι οποίες δείχνουν ποιες μεταβλητές και πως συνέβαλαν στην απόφαση που πήρε το μοντέλο. Παραδοσιακά μοντέλα όπως η Logistic Regression είναι από την φύση τους ερμηνεύσιμα, ωστόσο άλλα μοντέλα μηχανικής μάθησης, όπως για παράδειγμα το XGBoost, θεωρούνται black box διότι είναι περίπλοκα και δύσκολο να κατανοηθεί με ποιον τρόπο πήραν την απόφαση και να επαληθευτεί αυτή η απόφαση. Επιπλέον πραγματοποιείται μια ανάλυση δικαιοσύνης με στόχο να διαπιστωθεί αν υπάρχουν ενδείξεις μεροληψίας ήδη στα δεδομένα καθώς και αν αυτές εμφανίζονται στις προβλέψεις των μοντέλων στο σύνολο ελέγχου μέσα από την εξέταση βασικών μετρικών fairness που χρησιμοποιούνται στην πράξη.

1.2.1 Συνεισφορά

Η συνεισφορά της διπλωματικής συνοψίζεται ως εξής:

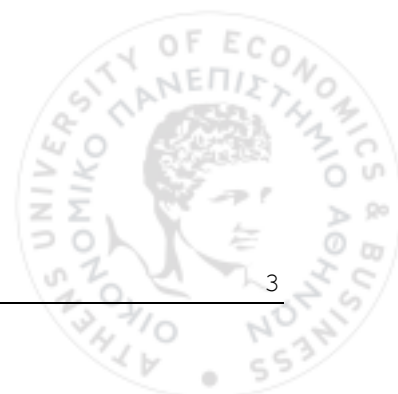
1. Εξετάστηκε το πρόβλημα της ανάλυσης πιστωτικού κινδύνου με αλγόριθμους μηχανικής μάθησης με έμφαση στη διαφάνεια, τη δικαιοσύνη και την μεροληψία καθώς και στις κανονιστικές απαιτήσεις που ελέγχουν τη χρήση τέτοιων εφαρμογών.



2. Έγινε σύγκριση διάφορων μοντέλων ταξινόμησης, τα οποία περιλαμβάνουν παραδοσιακά μοντέλα (Logistic Regression) και σύγχρονα μοντέλα μηχανικής μάθησης (Decision Tree, Random Forest, XGBoost, SVM και ANN) εφαρμόζοντας μια κοινή διαδικασία εκπαίδευσης και αξιολόγησης.
3. Η προετοιμασία των δεδομένων πραγματοποιήθηκε συστηματικά, ξεκινώντας με την κωδικοποίηση των κατηγορικών μεταβλητών, τα δεδομένα χωρίστηκαν σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου, και έγινε ρύθμιση υπερπαραμέτρων, για την δίκαιη σύγκριση των μοντέλων.
4. Τα μοντέλα αξιολογήθηκαν με βάση την απόδοση τους σε κατάλληλες μετρικές για την ανάλυση πιστωτικού κινδύνου, όπως η ROC–AUC και το Recall της κατηγορίας Bad, ώστε να αποτυπώνεται τόσο η διακριτική ικανότητα όσο και η πρακτική χρησιμότητα των μοντέλων. Επιπλέον έγινε ανάλυση κατάταξης μέσω cumulative lift για επιλεγμένα μοντέλα.
5. Οι μέθοδοι Explainable AI (XAI) αναπτύχθηκαν για να διερευνήσουν και να παρέχουν ερμηνευσιμότητα στις προβλέψεις σε global και local επίπεδο.
6. Πραγματοποιήθηκε ανάλυση δικαιοσύνης στα ευαίσθητα χαρακτηριστικά και σε συνδυασμό με την ανάλυση ερμηνευσιμότητας διερευνήθηκε η επίδραση τους στις αποφάσεις των μοντέλων

1.3 Οργάνωση κειμένου

Το Κεφάλαιο 1 της διπλωματικής παρέχει μια εισαγωγή στο θέμα, το αντικείμενο και την συνεισφορά. Το Κεφάλαιο 2 κάνει ανασκόπηση της σχετικής βιβλιογραφίας για τα θεσμικά και κανονιστικά πλαίσια στα οποία υπόκεινται τα χρηματοπιστωτικά ιδρύματα για την ανάλυση πιστωτικού κινδύνου και παρουσιάζει σχετικές εργασίες που χρησιμοποιούνται μοντέλα μηχανικής μάθησης στα οποία αναλύονται και ζητήματα δικαιοσύνης, διαφάνειας και μεροληψίας. Το Κεφάλαιο 3 αναπτύσσει το θεωρητικό υπόβαθρο των μοντέλων, παρουσιάζει τις μετρικές αξιολόγησης απόδοσης, τα fairness metrics, τεχνικές αντιμετώπισης μεροληψίας και τις μεθόδους Explainable AI. Το Κεφάλαιο 4 παρέχει λεπτομέρειες για την μεθοδολογία της εργασίας, ξεκινώντας από την προεπεξεργασία των δεδομένων, την εκπαίδευση και την αξιολόγησή τους. Το Κεφάλαιο 5 παρουσιάζει τα αποτελέσματα εξετάζοντας την προβλεπτική απόδοση, την ερμηνευσιμότητα των μοντέλων και την ανάλυση δικαιοσύνης. Το Κεφάλαιο 6 συνοψίζει τα συμπεράσματα της διπλωματικής και πού θα μπορούσε να κατευθυνθεί η μελλοντική έρευνα.



2 *Βιβλιογραφική Ανασκόπηση και Σχετικές εργασίες*

2.1 Κανονιστικό και Θεσμικό Πλαίσιο Πιστωτικού Κινδύνου και Διαφάνειας

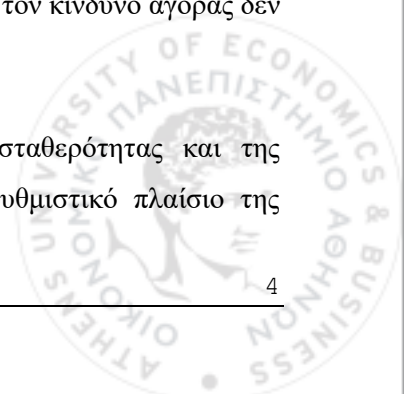
Η ανάλυση πιστωτικού κινδύνου πρέπει να εξεταστεί και από το θεσμικό και κανονιστικό πλαίσιο που τη διέπει καθώς αυτό καθορίζει τόσο τις πρακτικές διαχείρισης κινδύνου όσο και τις απαιτήσεις διαφάνειας των πιστωτικών ιδρυμάτων. Οπότε, παρουσιάζεται η εξέλιξη των Συμφωνιών της Βασιλείας και το σύγχρονο νομικό περιβάλλον που διαμορφώνεται μέσω του GDPR και του AI Act, με έμφαση στην επεξηγησιμότητα και την προστασία των δικαιωμάτων των υποκειμένων.

2.1.1 Χρονική επισκόπηση των συμφωνιών της Βασιλείας

Η Επιτροπή της Βασιλείας για την Τραπεζική Εποπτεία (BCBS) ιδρύθηκε το 1974, με στόχο την ενδυνάμωση του διεθνούς χρηματοπιστωτικού συστήματος. Έως τώρα έχουν εκπονηθεί τρεις διαδοχικές συμφωνίες της Βασιλείας (Basel Accords) για τη ρύθμιση του τραπεζικού τομέα (Βασιλεία I, II και III).

Η Βασιλεία I θεσπίστηκε το 1988 και εισήγαγε ένα ενιαίο ρυθμιστικό πλαίσιο για τις διεθνώς ενεργές τράπεζες, με στόχο να μειωθεί ο πιστωτικός κίνδυνος και να ενισχυθεί η χρηματοπιστωτική σταθερότητα. Προέβλεπε ελάχιστο δείκτη κεφαλαιακής επάρκειας 8% στα σταθμισμένα στοιχεία ενεργητικού και καθόριζε δύο επίπεδα κεφαλαίου (Tier 1 και Tier 2). Ωστόσο, τον κίνδυνο αγοράς δεν τον κάλυπτε (Barra et al.).

Η Βασιλεία II θεσπίστηκε το 2004 και στόχευε στην ενίσχυση της σταθερότητας και της ανθεκτικότητας του διεθνούς τραπεζικού συστήματος, επεκτείνοντας το ρυθμιστικό πλαίσιο της



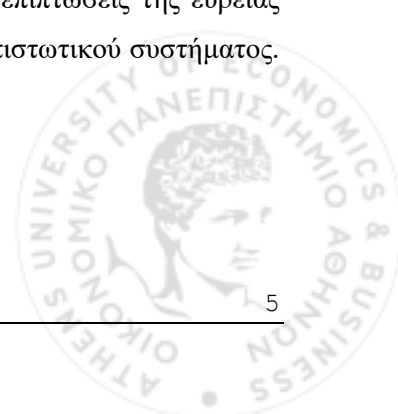
Βασιλείας I και καλύπτοντας τρεις βασικούς κινδύνους: πιστωτικό, αγοράς και λειτουργικό. Επιπροσθέτως βασίζεται σε τρεις θεμελιώδεις πυλώνες. Ο πρώτος αφορά τις ελάχιστες κεφαλαιακές απαιτήσεις που καθορίζουν το ύψος του απαιτούμενου κεφαλαίου ανάλογα με το επίπεδο κινδύνου που αναλαμβάνει κάθε τράπεζα. Ο δεύτερος πυλώνας είναι ο εποπτικός έλεγχος. Μέσω αυτού οι αρμόδιες αρχές αξιολογούν την κεφαλαιακή επάρκεια και τη συνολική διαχείριση κινδύνου των τραπεζικών ιδρυμάτων και όταν αυτό κρίνεται αναγκαίο λαμβάνουν διορθωτικά μέτρα. Ο τρίτος πυλώνας αφορά την πειθαρχία της αγοράς η οποία ενισχύεται μέσω της υποχρεωτικής δημοσιοποίησης πληροφοριών ώστε οι επενδυτές και οι συμμετέχοντες στην αγορά να μπορούν να αξιολογούν τη φερεγγυότητα και τη διαφάνεια των τραπεζών. Επιπλέον με την Βασιλεία II τα μοντέλα για την αξιολόγηση πιστωτικού κινδύνου πρέπει να είναι διαφανή που σημαίνει ότι οι προβλέψεις τους να μπορούν να ερμηνευτούν (Dastile, Celik, & Potsane, 2020)

Το 2007 η χρηματοπιστωτική κρίση ανέδειξε τις αδυναμίες του κανονιστικού πλαισίου της Βασιλείας II. Ως επακόλουθο το 2010 εισήχθη η Βασιλεία III με στόχο να ενισχύσει τη σταθερότητα και την ανθεκτικότητα του χρηματοπιστωτικού συστήματος απέναντι σε αρνητικές οικονομικές εξελίξεις (Barra et al.). Εισηγήθηκε παγκόσμια ρυθμιστικά πρότυπα για την κεφαλαιακή επάρκεια και τον κίνδυνο ρευστότητας ο οποίος δεν καλυπτόταν από τη Βασιλεία II.

Η Βασιλεία II επικεντρωνόταν στην αθέτηση υποχρεώσεων από τους πελάτες, ενώ η Βασιλεία III δίνει έμφαση στη συμμόρφωση των ίδιων των τραπεζών απέναντι στις υποχρεώσεις τους προς τους πελάτες.

Οι μεταρρυθμίσεις της Βασιλείας III το 2017 στοχεύουν στην ενίσχυση της αξιοπιστίας των υπολογισμών των σταθμισμένων στοιχείων ενεργητικού (RWA), στη βελτίωση της συγκρισιμότητας των κεφαλαιακών δεικτών και στον περιορισμό της ευελιξίας των εσωτερικών μοντέλων που χρησιμοποιούν οι τράπεζες για ρυθμιστικούς σκοπούς. Η εφαρμογή της πραγματοποιείται σταδιακά, με τις κεντρικές τροποποιήσεις στη διαχείριση του πιστωτικού, του λειτουργικού και του κινδύνου αγοράς να τίθενται σε ισχύ από 1 Ιανουαρίου 2023.

Τον Μάρτιο του 2022, η Επιτροπή της Βασιλείας (BCBS) ανακοίνωσε ότι θα συνεχίσει το έργο της εστιάζοντας σε τρία κύρια ζητήματα: (i) το βάθος κατανόησης και τη δυνατότητα εξήγησης των αποτελεσμάτων που παράγουν τα μοντέλα (ii) τις δομές διακυβέρνησης που αφορούν τη χρήση τεχνητής νοημοσύνης και μηχανικής μάθησης (AI/ML) και (iii) τις πιθανές επιπτώσεις της ευρείας υιοθέτησής τους στη σταθερότητα των τραπεζών και του συνολικού χρηματοπιστωτικού συστήματος. (Ridzuan et al.).



2.1.2 Οι υποχρεώσεις των πιστωτικών ιδρυμάτων για διαφάνεια και επεξηγησιμότητα όπως προκύπτουν από το GDPR και το AI Act

2.1.2.1 GDPR

Η χρήση αλγορίθμων μηχανικής μάθησης (ML) και συστημάτων τεχνητής νοημοσύνης (AI) έχει αυξηθεί στην αξιολόγηση πιστωτικού κινδύνου. Το γεγονός αυτό έχει προκαλέσει ανησυχίες σχετικά με τη διαφάνεια, τη λογοδοσία και την προστασία των θεμελιωδών δικαιωμάτων των δανειοληπτών. Υπάρχουν ωστόσο υποχρεώσεις που απορρέουν από τον Γενικό Κανονισμό για την Προστασία Δεδομένων (GDPR) στον οποίο υπάγονται τα πιστωτικά ιδρύματα. Αυτές αφορούν κυρίως όταν η απόφαση χορήγησης ή απόρριψης πίστωσης βασίζεται ολοκληρωτικά ή σε σημαντικό βαθμό σε αυτοματοποιημένη επεξεργασία και profiling (Άρθρο 22 GDPR). Το GDPR συνιστά το κεντρικό ρυθμιστικό θεμέλιο για τη διασφάλιση της διαφάνειας και της δικαιοσύνης στα συστήματα credit scoring, αναγνωρίζοντας ρητά το δικαίωμα του δανειολήπτη σε εξήγηση όταν η απόφαση βασίζεται σε αυτοματοποιημένη επεξεργασία. Ειδικότερα το άρθρο 22 παρέχει στον δανειολήπτη τη δυνατότητα να ζητήσει ανθρώπινη παρέμβαση, να εκφράσει την άποψη του και να αμφισβητήσει την απόφαση (Škocijanec, 2025). Σε πρόσφατη νομολογία του το Δικαστήριο της Ευρωπαϊκής Ένωσης στην υπόθεση SCHUFA επικύρωσε ότι και ένα probabilistic credit score μπορεί να αποτελεί «απόφαση» κατά την έννοια του άρθρου 22 όταν το πιστωτικό ίδρυμα βασίζεται σε αυτό για να παράγει έννομα ή σημαντικά αποτελέσματα για το υποκείμενο των δεδομένων, όπως είναι η απόρριψη δανείου (Aza, 2024).

Τα άρθρα 13(2)(f) και 14(2)(g) του GDPR αναφέρουν ότι τα πιστωτικά ιδρύματα οφείλουν να ενημερώνουν εκ των προτέρων το υποκείμενο των δεδομένων εάν θα υπάρξει αυτοματοποιημένη λήψη αποφάσεων και να τους παρέχει ουσιαστικές πληροφορίες για τη λογική που εμπλέκεται. Επιπλέον οφείλουν να ενημερώσουν για τη σημασία και τις προβλεπόμενες συνέπειες της επεξεργασίας. Επιπροσθέτως, με το άρθρο 15(1)(h) ο πολίτης έχει το δικαίωμα να ζητήσει εξατομικευμένη πληροφόρηση σχετικά με τον τρόπο που επηρέασε η αλγοριθμική διαδικασία την περίπτωση του. Ο Γενικός Εισαγγελέας Pikamäe στην υπόθεση SCHUFA ενίσχυσε την υποχρέωση αυτή με την ερμηνεία του τονίζοντας ότι οι πληροφορίες που παρέχονται πρέπει να περιλαμβάνουν αναλυτική εξήγηση της μεθόδου υπολογισμού του credit score και των παραγόντων που οδήγησαν στο συγκεκριμένο αποτέλεσμα. Αυτό πρέπει να ισχύει και στην περίπτωση που ο εκάστοτε υπεύθυνος επεξεργασίας δεδομένων (controller) προστατεύει εμπορικά μυστικά, ακόμα και τέτοιες περιπτώσεις δεν δικαιολογούν πλήρη άρνηση πληροφόρησης (Lui et al., 2025).

Η εφαρμογή αλγοριθμικών μοντέλων στη πιστοληπτική αξιολόγηση συχνά ακολουθείται από το φαινόμενο του «black box», τα αποτελέσματα παράγονται χωρίς να γνωρίζουμε πως λειτουργεί

εσωτερικά το μοντέλο και πως κατέληξε στην απόφαση. Αυτό έχει ως αποτέλεσμα να υπάρχει αδιαφάνεια στην διαδικασία, έλλειψη επεξηγησιμότητας και κίνδυνος μεροληψίας. Η ανάπτυξη τεχνικών Explainable AI (XAI), επιτρέπουν την ερμηνεία των αποτελεσμάτων σύνθετων μοντέλων καθιστώντας εφικτή τη συμμόρφωση με τις απαιτήσεις διαφάνειας του GDPR.

Η θεωρητική συζήτηση γύρω από την ύπαρξη «δικαιώματος στην εξήγηση» παραμένει ανοιχτή. Κάποιοι μελετητές θεωρούν ότι το άρθρο 22 GDPR παρέχει μόνο περιορισμένη λειτουργική πληροφόρηση ενώ άλλοι τάσσονται υπέρ μιας ολιστικής ερμηνείας. Υποστηρίζουν ότι το άρθρο 22 GDPR σε συνδυασμό με τις διατάξεις των άρθρων 13–15 θεμελιώνουν έναν ουσιαστικό μηχανισμό λογοδοσίας για αλγοριθμικά συστήματα (Lui et al., 2025).

Επομένως η διαφάνεια καθίσταται νομική υποχρέωση και τεχνολογική πρόκληση. Η ενσωμάτωση XAI τεχνικών στα συστήματα μηχανικής μάθησης από τα πιστωτικά ιδρύματα είναι απαραίτητη. Με αυτό τον τρόπο θα παρέχονται κατανοητές και λειτουργικές εξηγήσεις για τις αποφάσεις των μοντέλων, θα ενισχύεται η εμπιστοσύνη και θα περιοριστεί η συστημική μεροληψία. Η επεξηγησιμότητα θα ενισχύει την εμπιστοσύνη των χρηστών συμβάλλοντας στην πιο υπεύθυνη και διαφανή χρήση της τεχνητής νοημοσύνης στην πιστοληπτική αξιολόγηση (Lui et al., 2025).

2.1.2.2 AI Act

Ο τρόπος με τον οποίο αξιολογείται η πιστοληπτική ικανότητα έχει αλλάξει από την διάδοση αλγορίθμων μηχανικής μάθησης (ML) και συστημάτων τεχνητής νοημοσύνης (AI). Η ανθρώπινη κρίση και τα περιορισμένα δεδομένα έχουν αντικατασταθεί από εξελιγμένα μοντέλα που αξιοποιούν μεγάλους όγκους δεδομένων. Η εξέλιξη αυτή όμως εμπεριέχει και σημαντικούς κινδύνους. Οι προκαταλήψεις που υπάρχουν στα δεδομένα ή στους αλγορίθμους ενδέχεται να οδηγήσουν σε άνιση ή άδικη μεταχείριση διαφορετικών κοινωνικών ομάδων (Langenbacher, 2022).

Στο πλαίσιο αυτό η Ευρωπαϊκή Επιτροπή στις 21 Απριλίου 2021 παρουσίασε το AI Act που αποτελεί την πρώτη ολοκληρωμένη ευρωπαϊκή προσπάθεια για θέσπιση ενιαίου, οριζόντιου κανονιστικού πλαισίου για την τεχνητή νοημοσύνη. Τέθηκε σε ισχύ την 1η Αυγούστου 2024 και θα εφαρμοστεί πλήρως στις 2 Αυγούστου 2026, με ορισμένες εξαιρέσεις (European Commission, 2025). Η ευρωπαϊκή ρύθμιση για το AI υιοθετεί μια προσέγγιση βάσει κινδύνου που σημαίνει ότι κατηγοριοποιούνται τα συστήματα σε μη αποδεκτού, υψηλού, περιορισμένου και ελάχιστου κινδύνου. Τα συστήματα αξιολόγησης πιστοληπτικής ικανότητας φυσικών προσώπων κατατάσσονται στην κατηγορία υψηλού κινδύνου (Παράρτημα III, 5(β)). Τα συστήματα υψηλού κινδύνου υπάγονται σε αυστηρές απαιτήσεις συμμόρφωσης όπως είναι η διασφάλιση ποιότητας δεδομένων, η αξιολόγηση και

ο μετριασμός κινδύνων, η διαφάνεια, η ανθρώπινη επίβλεψη και η αυξημένη αξιοπιστία, ώστε να ελαχιστοποιηθεί ο κίνδυνος διακρίσεων και αδικαιολόγητων αποτελεσμάτων (European Commission, 2025).

Οι πάροχοι τώρα των συστημάτων υψηλού κινδύνου πρέπει να τηρούν απαιτήσεις όπως ποιότητα και διακυβέρνηση δεδομένων, τεκμηρίωση και καταγραφή διαδικασιών, ανθρώπινη επίβλεψη και διαφάνεια προς τους οργανισμούς που χρησιμοποιούν τα συστήματα (Langenbacher, 2022).

Στους αλγορίθμους μηχανικής μάθησης για την αξιολόγηση πιστοληπτικής ικανότητας εντοπίζονται κίνδυνοι που σχετίζονται με τη μεροληψία στα δεδομένα, τη στατιστική ανισορροπία μεταξύ ομάδων και την έλλειψη διαφάνειας ως προς τον τρόπο λήψης αποφάσεων. Τα μοντέλα συχνά βασίζονται σε συσχετίσεις και αυτό μπορεί να ενισχύσει υπάρχουσες μεροληψίες όταν η ποιότητα ή η αντιπροσωπευτικότητα των δεδομένων διαφέρει μεταξύ κοινωνικών ομάδων. (Langenbacher, 2022). Ακόμη πολλά μοντέλα είναι «black-box» το οποίο δυσκολεύει τη δυνατότητα ελέγχου και κατανόησης των αποφάσεων τόσο από τους εποπτικούς φορείς όσο και από τους ενδιαφερόμενους πολίτες.

Η Langenbacher (2022), στο κείμενό της που δημοσιεύτηκε στα ECB Legal Conference Proceedings, αναφέρει ότι το AI Act δημιουργεί μια μορφή διπλής ρύθμισης. Από την μια, οι τράπεζες εποπτεύονται από το υφιστάμενο πλαίσιο CRD IV (Capital Requirements Directive) και θεωρείται ότι πληρούν κάποιες απαιτήσεις του AI Act κάτω από τα ήδη λειτουργούντα συστήματα διαχείρισης κινδύνου. Από την άλλη πλευρά, FinTechs και άλλοι μη τραπεζικοί πάροχοι θα εποπτεύονται από διαφορετικές αρχές, οδηγώντας ενδεχομένως σε ανομοιογενείς πρακτικές συμμόρφωσης και άνιση προστασία καταναλωτών. Το γεγονός αυτό μπορεί να έχει επιπτώσεις στον ανταγωνισμό αλλά και στη συνοχή του ρυθμιστικού πλαισίου στο συγκεκριμένο τομέα.

Με το AI Act η προστασία των θεμελιωδών δικαιωμάτων τίθεται ως κεντρικός προβληματισμός. Σε αντίθεση με άλλους κινδύνους, για παράδειγμα η ασφάλεια, οι κίνδυνοι διάκρισης και αδικίας δεν είναι εύκολα ποσοτικοποιήσιμοι και απαιτούν ηθική και κοινωνική στάθμιση, πέρα από τεχνική επάρκεια (Langenbacher, 2022). Ακόμη η ανθρώπινη επίβλεψη δεν είναι επαρκής διότι πολλές φορές συνοδεύεται με την τάση του ανθρώπου να δείχνει πολλή εμπιστοσύνη στην απόφαση που προέρχεται από έναν αλγόριθμο.

Επομένως εφαρμογή του AI Act στην αξιολόγηση πιστοληπτικής ικανότητας αποτελεί κρίσιμη δοκιμασία. Αν και ο κανονισμός αποτελεί ένα φιλόδοξο πλαίσιο, χρειάζεται εξειδίκευση και εμβάθυνση ειδικά στον χρηματοπιστωτικό τομέα ώστε να επιτευχθεί ισορροπία μεταξύ καινοτομίας, προστασίας δικαιωμάτων και χρηματοπιστωτικής σταθερότητας (Langenbacher, 2022).

2.2 Σχετικές εργασίες σε Μοντέλα Μηχανικής Μάθησης για Πιστωτικό Κίνδυνο: Δικαιοσύνη και Επεξηγησιμότητα

2.2.1 Μεροληψία και Δικαιοσύνη

Οι Mehrabi et al. (2021b) κάνουν μια λεπτομερή θεωρητική παρουσίαση αναφέροντας τις βασικές πηγές μεροληψίας στα συστήματα μηχανικής μάθησης και τις διαχωρίζουν σε μεροληψίες που προέρχονται από τα δεδομένα, τους αλγορίθμους και την αλληλεπίδραση με τους χρήστες. Αναφέρουν συνοπτικά βασικούς τύπους data bias, όπως measurement bias, omitted variable bias, representation bias και sampling bias και δηλώνουν ότι ακόμη και χωρίς μεροληπτικά δεδομένα, οι αλγόριθμοι ενδέχεται να εισάγουν μεροληψία λόγω σχεδιαστικών επιλογών. Ακόμη υποστηρίζουν ότι οι αποφάσεις αλγορίθμων επηρεάζουν τη συμπεριφορά των χρηστών και μπορούν να δημιουργήσουν νέα μεροληπτικά δεδομένα, γεγονός που ολοκληρώνει τον κύκλο μεροληψίας. Το άρθρο συγκεντρώνει διαφορετικούς ορισμούς της δικαιοσύνης που υπάρχουν σε διάφορες μελέτες για να δείξει ότι δεν υπάρχει καθολικός ορισμός της. Παρατίθενται κλασικές μετρικές δικαιοσύνης, όπως demographic parity, equal opportunity και equalized odds και άλλοι ορισμοί όπως test fairness (calibration), treatment equality και fairness through awareness. Αναφέρονται και πιο σύνθετοι ορισμοί δικαιοσύνης, όπως counterfactual fairness και causal fairness, οι οποίοι εξετάζουν αποφάσεις σε υποθετικά σενάρια όπου τα άτομα ανήκουν σε ξεχωριστές δημογραφικές ομάδες. Οι μετρικές δικαιοσύνης κατηγοριοποιούνται σε ομαδική δικαιοσύνη (group fairness), ατομική δικαιοσύνη (individual fairness) και δικαιοσύνη υποομάδων (subgroup fairness), και εξηγούν πώς κάθε κατηγορία εφαρμόζει διαφορετικούς περιορισμούς στη συμπεριφορά μοντέλων. Οι ορισμοί της δικαιοσύνης από τους συγγραφείς είναι μαθηματικά ασύμβατοι μεταξύ τους και η επιλογή μιας μετρικής δικαιοσύνης πρέπει να εξαρτάται από το συγκεκριμένο πλαίσιο εφαρμογής.

Στο άρθρο τους οι Zhou et al. (2022) περιγράφουν πώς η δικαιοσύνη μπορεί να ποσοτικοποιηθεί χρησιμοποιώντας διαφορετικές μετρήσεις δικαιοσύνης. Οι πιο συνηθισμένες μετρικές είναι οι demographic parity, predictive rate parity, equalized odds και equal opportunity, και κάθε μια από αυτές καθορίζει διάφορες απαιτήσεις ισότητας μεταξύ προστατευμένων και μη προστατευμένων ομάδων. Οι συγγραφείς εισάγουν επίσης την έννοια conditional parity. Αυτή η έννοια δεν εξετάζει την δικαιοσύνη μεταξύ όλων των ατόμων, αλλά μεταξύ ατόμων με τα ίδια χαρακτηριστικά, όπως η πιστοληπτική ικανότητα. Παρουσιάζουν επίσης την ατομική δικαιοσύνη και το counterfactual fairness που σημαίνει ότι μια απόφαση είναι δίκαιη όταν σε παρόμοια άτομα δίνονται παρόμοιες αποφάσεις και το αποτέλεσμα δεν θα άλλαζε εάν το άτομο είχε διαφορετικό δημογραφικό υπόβαθρο. Πάλι και

εδώ οι συγγραφείς υπογραμμίζουν την έλλειψη ενός κοινού ορισμού της δικαιοσύνης και ότι πολλές μετρικές δικαιοσύνης βρίσκονται σε μαθηματική σύγκρουση. Σύμφωνα με την έρευνα, οι μετρήσεις δικαιοσύνης δεν θα πρέπει να χρησιμοποιούνται μόνο στο στάδιο ολοκλήρωσης ως έλεγχος, αλλά θα πρέπει να ενσωματώνονται σε όλη τη διαδικασία ανάπτυξης του μοντέλου. Συγκεκριμένα, προτείνεται είτε να εφαρμόζονται ως περιορισμοί κατά τη διάρκεια της εκπαίδευσης είτε να αξιολογούνται μετά την εκπαίδευση, προκειμένου να διαπιστωθεί εάν το μοντέλο είναι μεροληπτικό, σε εφαρμογές όπως η αξιολόγηση πιστοληπτικής ικανότητας.

Η μελέτη των Bono et al. (2021) είναι μία από τις πιο ολοκληρωμένες εμπειρικές αναλύσεις για τη δικαιοσύνη των αλγορίθμων μηχανικής μάθησης στην αξιολόγηση πιστωτικού κινδύνου. Στόχος ήταν να συγκρίνουν την ακρίβεια πρόβλεψης και τη στατιστική δικαιοσύνη μεταξύ διαφορετικών υποομάδων του πληθυσμού. Αυτό που έκαναν ήταν να προσομοιώσουν τη μετάβαση από ένα παραδοσιακό λογιστικό μοντέλο (logit) σε ensemble τεχνικές μηχανικής μάθησης, όπως τα Random Forests και το XGBoost χρησιμοποιώντας δεδομένα πιστωτικών αρχείων από 800.000 ενήλικες στο Ηνωμένο Βασίλειο. Στα αποτελέσματα παρουσιάζεται ότι τα ML μοντέλα είναι αρκετά πιο ακριβή χωρίς να σημαίνει ότι είναι λιγότερο δίκαια από τα παραδοσιακά. Οι συγγραφείς για να μετρήσουν την δικαιοσύνη χρησιμοποίησαν τα κριτήρια performance parity, separation και sufficiency και δημιούργησαν proxy μεταβλητές για το φύλο, την εθνικότητα, την υγεία και την κοινωνικοοικονομική αποστέρηση μέσω UK census data και αλγορίθμου k-means clustering. Τελικά στα αποτελέσματα υπήρχαν λίγες αποκλίσεις στην απόδοση μεταξύ ομάδων, όπως για παράδειγμα για το φύλο, οι οποίες όμως εμφανίζονταν και στα παραδοσιακά μοντέλα και η μετάβαση σε πιο σύνθετους αλγορίθμους δεν επιδείνωσε τις ανισότητες.

Επιπροσθέτως στην μελέτη έγινε και μια εισαγωγή ευαίσθητων χαρακτηριστικών στα μοντέλα η οποία όμως δεν βελτίωσε ουσιαστικά την πρόβλεψη που υποδηλώνει ότι οι σχετικές πληροφορίες είναι ήδη έμμεσα ενσωματωμένες στα δεδομένα πίστωσης όπως το ιστορικό πληρωμών ή η γεωγραφική τοποθεσία. Στα συμπεράσματα οι συγγραφείς δηλώνουν ότι τα μοντέλα μηχανικής μάθησης ενισχύουν την ακρίβεια αλλά δεν επιτυγχάνουν την πλήρη στατιστική δικαιοσύνη. Καταλήγουν ότι το “fairness” πρέπει να εξετάζεται στο πλαίσιο των πραγματικών αποφάσεων δανεισμού και των κοινωνικών συνεπειών τους.

Στο άρθρο τους οι Ρανόν Pérez et al. (2023) εξέτασαν συστηματικά τη συμπεριφορά της μεροληψίας σε πιστωτικά συστήματα που βασίζονται στη μηχανική μάθηση τα οποία βασίζονται σε παραδοσιακά και σύγχρονα τραπεζικά συστήματα. Τα ευρήματα της μελέτης αποκαλύπτουν ότι η αφαίρεση ευαίσθητων χαρακτηριστικών όπως το φύλο δεν οδηγεί σε αμερόληπτες αποφάσεις, καθώς η μεροληψία μπορεί ακόμα να ενσωματωθεί μέσω υποκατάστατων χαρακτηριστικών (proxies). Για την αντιμετώπιση της μεροληψίας, οι συγγραφείς προτείνουν μια μέθοδο που βασίζεται στη μελέτη των

σχέσεων μεταξύ μεταβλητών. Στόχος της μεθόδου είναι ο εντοπισμός χαρακτηριστικών που παρέχουν κρυφά πληροφορίες σχετικά με ευαίσθητα χαρακτηριστικά, όπως το φύλο. Η μέθοδος επιτρέπει να προσδιοριστεί ποια χαρακτηριστικά συμβάλλουν στατιστικά στη δημιουργία μεροληψίας. Η μελέτη μέσα από εφαρμογές σε πραγματικά δεδομένα δείχνει ότι η επιλεκτική αφαίρεση τέτοιων μεταβλητών μπορεί να μειώσει σημαντικά την ανισότητα στις αποφάσεις του μοντέλου σύμφωνα με μετρικές δικαιοσύνης όπως το statistical parity και το equalized odds, χωρίς να επηρεάζεται η ακρίβεια της πρόβλεψης. Αυτή η έρευνα είναι σημαντική καθώς προσφέρει με σαφή και οργανωμένο τρόπο ένα μέσο αναζήτησης και αντιμετώπισης μεροληψίας στα συστήματα αξιολόγησης πιστοληπτικής ικανότητας. Δείχνει ότι σωστή ανάλυση δεδομένων αποτελεί σημαντικό βήμα προς τη διασφάλιση δίκαιων και αξιόπιστων αποφάσεων από αλγοριθμικά μοντέλα.

Το άρθρο των Goodness et al. (2025) συμπληρώνει την υπάρχουσα έρευνα με μια συστηματική αξιολόγηση της δικαιοσύνης και της διαφάνειας σε μοντέλα μηχανικής μάθησης για την αξιολόγηση πιστωτικού κινδύνου. Χρησιμοποιούν δείκτες δικαιοσύνης και τεχνικές Explainable AI. Για να ελέγξουν εάν οι αποφάσεις δανεισμού ευνοούν ή αδικούν κάποιες ομάδες, χρησιμοποιούν τρεις βασικές μετρικές οι οποίες είναι οι Disparate Impact Ratio, Equal Opportunity Difference και Statistical Parity Difference. Η ανάλυση δείχνει ότι τα deep learning μοντέλα, παρουσιάζουν τιμές που υποδεικνύουν πιθανή διάκριση. Αντίθετα τα μοντέλα που βασίζονται σε δέντρα και ενσωματώνουν περιορισμούς επιτυγχάνουν καλύτερα αποτελέσματα στην ισότιμη μεταχείριση των ομάδων. Επιπλέον χρησιμοποιούν τεχνικές Explainable AI όπως είναι οι Feature Importance, SHAP, Partial Dependence Plots και LIME, για να δείξουν ποια χαρακτηριστικά του μοντέλου επηρεάζουν τις αποφάσεις και πώς μικρές αλλαγές στις μεταβλητές επηρεάζουν τα αποτελέσματα. Όλα αυτά συνδέονται άμεσα με τα κανονιστικά πλαίσια όπως το GDPR ότι η ερμηνευσιμότητα είναι απαραίτητη για τη συμμόρφωση και την προστασία των ατόμων ειδικά στις χρηματοοικονομικές υπηρεσίες.

2.2.2 Εξηγησιμότητα (XAI)

Οι Ariza-Garzón et al. (2020) έκαναν μια μελέτη η οποία εστιάζει στην ανάπτυξη και την ερμηνεία μοντέλων μηχανικής μάθησης με στόχο την αξιολόγηση του πιστωτικού κινδύνου στις πλατφόρμες peer-to-peer (P2P) δανεισμού. Οι συγγραφείς κάνουν μια σύγκριση ανάμεσα σε λογιστική παλινδρόμηση, που είναι η κλασική, παραδοσιακή προσέγγιση στον πιστωτικό κίνδυνο με πιο εξελιγμένους αλγόριθμους μηχανικής μάθησης όπως τα Decision Tree, τα Random Forest και το XGBoost. Στην ανάλυση τους χρησιμοποίησαν το dataset Lending Club, που είναι διαθέσιμο στο Kaggle, το οποίο περιλαμβάνει πάνω από 1,3 εκατομμύρια εγγραφές δανείων για το διάστημα 2007-2018. Στο μοντέλο οι ανεξάρτητες μεταβλητές που χρησιμοποιήθηκαν είχαν δημογραφικά και οικονομικά χαρακτηριστικά όπως το ύψος του δανείου (loan amount) και το ετήσιο εισόδημα

(revenue), καθώς και κατηγορικές μεταβλητές όπως η ιδιοκατοίκηση (home ownership) και ο σκοπός του δανείου (purpose).

Έδωσαν ιδιαίτερη έμφαση στην εξηγησιμότητα των προβλέψεων, χρησιμοποιώντας τις SHAP (SHapley Additive exPlanations) values, που καθιστούν δυνατή την ανάλυση τόσο της συνολικής (global) όσο και της τοπικής (local) συνεισφοράς κάθε μεταβλητής στο αποτέλεσμα. Οι SHAP values, μετρούν τη συνεισφορά κάθε μεταβλητής στην τελική πρόβλεψη, αποσαφηνίζοντας την αδιαφάνεια των black box της μηχανικής μάθησης. Στην έρευνα τους οι συγγραφείς προσαρμίζουν τις SHAP values και για τις κατηγορικές μεταβλητές, ώστε να ληφθεί υπόψη η αλληλεξάρτηση μεταξύ των διαφόρων κατηγοριών. Επίσης παρουσιάζουν τα διαγράμματα summary και dependence plots στα οποία βλέπουμε τις μη-γραμμικές σχέσεις, δομικές μεταβολές και ετεροσκεδαστικότητα που παραμένουν αόρατες στη λογιστική παλινδρόμηση. Τελικά ο XGBoost παρείχε την πιο ακριβή πρόβλεψη και είχε την πιο υψηλή εξηγησιμότητα. Αυτό δείχνει ότι τα σύγχρονα μοντέλα μηχανικής μάθησης μπορούν να είναι και διαφανή και να έχουν υψηλή απόδοση. Στο άρθρο αναφέρεται ότι η χρήση τεχνικών όπως οι SHAP values ενισχύει την εμπιστοσύνη μεταξύ χρηστών, επενδυτών και ρυθμιστικών αρχών, και καθιστά τα μοντέλα μηχανικής μάθησης πιο αποδεκτά σε εφαρμογές πιστωτικού κινδύνου.

Οι Bussmann et al. (2021) ανέπτυξαν ένα πλαίσιο μηχανικής μάθησης για να καταστήσουν την ανάλυση πιστωτικού κινδύνου ακριβή και ερμηνεύσιμη. Οι συγγραφείς αναφέρουν ότι τα μοντέλα μηχανικής μάθησης black box δεν θεωρούνται κατάλληλα σε ρυθμιζόμενα χρηματοοικονομικά περιβάλλοντα διότι αυτά απαιτούν σαφή αιτιολόγηση στις αποφάσεις. Για να αντιμετωπίσουν αυτό, προτείνουν μια μεταγενέστερη (post-processing), model-agnostic μεθοδολογία, η οποία χρησιμοποιεί τιμές Shapley για την αξιολόγηση των συνεισφορών μεμονωμένων μεταβλητών σε τοπικό και παγκόσμιο επίπεδο. Η έρευνα εισάγει μια νέα προσέγγιση ενσωματώνοντας τιμές Shapley με δίκτυα συσχέτισης για την ομαδοποίηση παρόμοιων προφίλ δανειοληπτών για βελτιωμένη ανάλυση πιστωτικού κινδύνου. Χρησιμοποιήθηκαν δεδομένα μικρομεσαίων επιχειρήσεων για εμπειρική εφαρμογή και η μελέτη έδειξε να διατηρείται υψηλή ακρίβεια σε σύνθετα μοντέλα και να παρέχονται εξηγήσεις που είναι ταυτόχρονα εξατομικευμένες και ολοκληρωμένες, καθώς και να ενισχύεται η δυνατότητα λογοδοσίας των συστημάτων αξιολόγησης πιστοληπτικής ικανότητας.

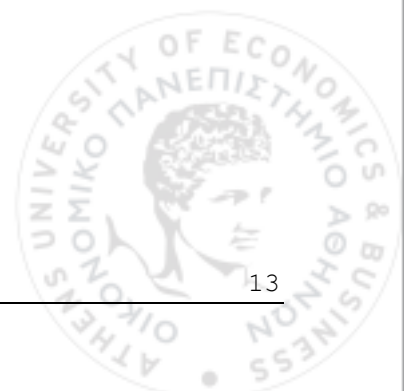
Η διαδικασία λήψης αποφάσεων σχετικά με την ανάλυση πιστωτικού κινδύνου καθίσταται πιο διαφανής και αξιόπιστη μέσω του πλαισίου Explainable AI (XAI), αναφέρουν στο άρθρο τους οι Nallakaruppan et al. (2024). Επιπλέον αναγνωρίζουν ότι αποτελεί πρόκληση στα παραδοσιακά και σύγχρονα μοντέλα μηχανικής μάθησης η εξήγηση των αποφάσεων των μοντέλων επειδή είναι black box, με αποτέλεσμα να δημιουργούν κινδύνους δικαιοσύνης και συμμόρφωσης. Για να αντιμετωπιστεί αυτό συνδυάζονται τεχνικές explainability σε τοπικό αλλά και σε παγκόσμιο επίπεδο, όπως τα LIME,

SHAP και PDP, για να εξηγηθεί ο τρόπος με τον οποίο κάθε στοιχείο επηρεάζει τις αποφάσεις του μοντέλου. Η διαδικασία αξιολόγησης της πιστοληπτικής ικανότητας καθίσταται πιο διαφανής, υπεύθυνη και δίκαιη μέσω της χρήσης εξηγήσεων χωρίς να επηρεάζεται η ακρίβεια του μοντέλου, σύμφωνα με τα πειράματα δεδομένων που διεξήγαγαν οι συγγραφείς.

2.2.3 Προχωρημένες Τεχνικές Τοπικών Εξηγήσεων και Counterfactuals

Στόχος των Dastile και Celik (2024) είναι να δείξουν πώς η αξιολόγηση της πιστοληπτικής ικανότητας μπορεί να γίνει πιο διαφανής και δίκαιη μέσω της χρήσης counterfactual explanations. Προτείνουν μια νέα βελτιστοποιητική διατύπωση για τη δημιουργία counterfactual explanations που είναι ικανές να υποστηρίζουν ταυτόχρονα πολλαπλά χαρακτηριστικά ποιότητας. Η μέθοδός τους περιλαμβάνει ένα πρόβλημα βελτιστοποίησης που δημιουργεί counterfactuals με βάση πέντε χαρακτηριστικά, την εγκυρότητα (validity), αραιότητα αλλαγών (sparsity), ομοιότητα με το αρχικό προφίλ (similarity), δυνατότητα εφαρμογής των αλλαγών από τον χρήστη (actionability) και ρεαλιστικότητα (plausibility). Τα counterfactual explanations ορίζονται ως υποθετικά σενάρια τα οποία δείχνουν πώς η απόφαση του μοντέλου θα μεταβληθεί αλλάζοντας ορισμένες ανεξάρτητες μεταβλητές. Στα συμπεράσματα, οι συγγραφείς αναφέρουν ότι η προτεινόμενη προσέγγιση είναι σε θέση να παράγει εξηγήσεις με counterfactual explanations που επιτυγχάνουν καλύτερη ισορροπία μεταξύ πολλαπλών επιθυμητών ιδιοτήτων σε σύγκριση με τις υπάρχουσες μεθόδους.

Οι Liang et al. (2025) προτείνουν μια νέα μέθοδο για την κατανόηση των προβλέψεων των μοντέλων που βασίζονται σε δέντρα, με έμφαση στις εξηγήσεις σε τοπικό επίπεδο. Παρουσιάζουν τη μέθοδο Local MDI+ (LMDI+), η οποία υπολογίζει την επίδραση κάθε χαρακτηριστικού σε μια συγκεκριμένη πρόβλεψη. Η μέθοδος βασίζεται στη δομή του ίδιου του μοντέλου και όχι σε τυχαίες αλλαγές των δεδομένων, όπως κάνουν άλλες τεχνικές. Για την μέθοδο χρησιμοποιούνται δεδομένα που δεν μπήκαν στην εκπαίδευση και ένα απλό στατιστικό μοντέλο (regularized GLM) για να εκτιμηθεί η σημασία των χαρακτηριστικών. Συγκρίνουν επίσης το LMDI+ με άλλες μεθόδους όπως το LIME και το TreeSHAP. Τα αποτελέσματα δείχνουν ότι η προτεινόμενη μέθοδος παρέχει αξιόπιστες εξηγήσεις. Στα συμπεράσματα, οι συγγραφείς υποστηρίζουν ότι το LMDI+ μπορεί να χρησιμοποιηθεί για την αναζήτηση υποομάδων και για τη δημιουργία βελτιωμένων counterfactual explanations.



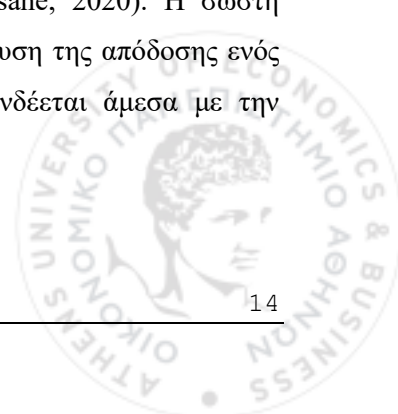
3

Θεωρητικό υπόβαθρο

3.1 Machine Learning για Credit Scoring

Η μηχανική μάθηση (ML) είναι ένας κλάδος της τεχνητής νοημοσύνης που ασχολείται με τη δημιουργία αλγορίθμων ικανών να επιτρέπουν στους υπολογιστές να «μαθαίνουν» από τα δεδομένα και να βελτιώνουν τη συμπεριφορά τους χωρίς ρητό προγραμματισμό. Τα ML συστήματα ψάχνουν για κρυφά πρότυπα και σχέσεις στα σύνολα δεδομένων ώστε να μπορούν να κάνουν προβλέψεις ή να παίρνουν αποφάσεις όταν τους εισάγουμε νέα, άγνωστα δεδομένα. Αρχικά τα δεδομένα συγκεντρώνονται και προεπεξεργάζονται, μετά επιλέγεται και εκπαιδεύεται το κατάλληλο μοντέλο και στο τέλος αξιολογείται η απόδοσή του. Οι κύριες κατηγορίες στη μηχανική μάθηση χωρίζονται σε τρία σύνολα: (i) η Επιβλεπόμενη Μάθηση (Supervised Learning). Το μοντέλο εκπαιδεύεται με δεδομένα που έχουν γνωστές ετικέτες, και χρησιμοποιείται για πρόβλεψη τιμών ή ταξινόμηση. (ii) Η Μη επιβλεπόμενη Μάθηση (Unsupervised Learning). Τα δεδομένα δεν έχουν ετικέτες και ο αλγόριθμος προσπαθεί να εντοπίσει κρυφά πρότυπα ή να σχηματίσει ομάδες (clusters). (iii) Η Ενισχυτική Μάθηση (Reinforcement Learning). Ο αλγόριθμος μαθαίνει μέσω επαναλαμβανόμενης αλληλεπίδρασης με το περιβάλλον, λαμβάνοντας ενίσχυση (επιβράβευση) για τις ενέργειές του, προσαρμόζοντας τη συμπεριφορά του ώστε να βελτιώνεται με το χρόνο.

Η ανάλυση πιστωτικού κινδύνου είναι ένα πρόβλημα Επιβλεπόμενης Μάθησης και πιο συγκεκριμένα, ένα πρόβλημα δυαδικής ταξινόμησης (binary classification problem), όπου ο στόχος είναι να γίνει διάκριση μεταξύ καλών και κακών δανειοληπτών (Dastile, Celik, & Potsane, 2020). Η σωστή διάκριση ανάμεσα στις δύο αυτές ομάδες είναι ιδιαίτερα σημαντική. Η ενίσχυση της απόδοσης ενός μοντέλου, κυρίως στην πρόβλεψη των λιγότερο αξιόπιστων πελατών, συνδέεται άμεσα με την κερδοφορία και τη βιωσιμότητα των χρηματοπιστωτικών ιδρυμάτων.



Τα μοντέλα μηχανικής μάθησης προσφέρουν πολύ πιο ακριβείς προβλέψεις αλλά δεν είναι εύκολο να τις εξηγήσουμε και να τις ερμηνεύσουμε. Ειδικά στα χρηματοπιστωτικά ιδρύματα που υπάρχουν ρυθμιστικά πλαίσια, η διαφάνεια στις αποφάσεις είναι σημαντική.

3.2 Μοντέλα Machine Learning

3.2.1 Logistic regression

Η Logistic Regression είναι ένας αλγόριθμος που χρησιμοποιείται για τη μοντελοποίηση προβλημάτων δυαδικής ταξινόμησης. Εφαρμόζει τη λογιστική συνάρτηση (sigmoid), για να δημιουργήσει μια καμπύλη σχήματος S, και μας δείχνει την πιθανότητα ενός συμβάντος να συμβεί. Πιθανότητες που προβλέπονται βρίσκονται εντός του εύρους 0 και 1. Στην ανάλυση πιστωτικού κινδύνου το μοντέλο με βάση ένα σύνολο χαρακτηριστικών θα εκτιμήσει την πιθανότητα αθέτησης. (James et al., 2021)

Οι συντελεστές του μοντέλου δείχνουν πως αλλάζουν οι λογαριθμικές πιθανότητες (log-odds) της εξαρτημένης μεταβλητής με βάση την αύξηση μίας μονάδας σε ένα χαρακτηριστικό. Αυτό επιτρέπει να κατανοηθεί πόσο επηρεάζει κάθε χαρακτηριστικό το τελικό αποτέλεσμα. Η Logistic Regression έχει όμως και ορισμένα μειονεκτήματα. Παρουσιάζει περιορισμούς, κυρίως λόγω της υπόθεσης γραμμικής σχέσης μεταξύ των χαρακτηριστικών και της λογαριθμικής πιθανότητας του αποτελέσματος. Αυτό συμβαίνει επειδή προϋποθέτει μια απλή σχέση μεταξύ των χαρακτηριστικών και της λογαριθμικής πιθανότητας της πρόβλεψης. Μειώνει επίσης την αποτελεσματικότητα σε καταστάσεις όπου υπάρχει ισχυρή μη γραμμική ή σύνθετη σχέση μεταξύ των μεταβλητών. (James et al., 2021)

3.2.2 Decision Tree

Το Decision Tree είναι αλγόριθμος για προβλήματα ταξινόμησης και παλινδρόμησης. Ο τρόπος που δουλεύει είναι να ξεκινάει με μια διαδοχική διάσπαση των δεδομένων με βάση ένα σύνολο κανόνων της μορφής "αν-τότε" έτσι ώστε οι παρατηρήσεις που καταλήγουν σε κάθε τελικό κόμβο (leaf) να είναι όσο το δυνατόν πιο ομοιογενείς ως προς τη μεταβλητή-στόχο. Στην ταξινόμηση, η επιλογή του split γίνεται συνήθως με βάση μετρικές ακαθαρσίας όπως το Gini index. (James et al., 2023)

Το Decision Tree είναι επίσης εύκολα ερμηνεύσιμο, μας βοηθά να κατανοήσουμε πώς λαμβάνονται οι αποφάσεις. Από την άλλη όμως δεν είναι σταθερό μοντέλο, μικρές αλλαγές στα δεδομένα μπορούν να οδηγήσουν σε άλλο δέντρο και να εμφανίζουν υπερπροσαρμογή (overfitting) όταν επιτρέπεται να μεγαλώσουν χωρίς περιορισμούς. Για να αντιμετωπιστούν αυτά στην πράξη εφαρμόζονται συχνά τεχνικές ελέγχου πολυπλοκότητας όπως περιορισμός βάθους, ελάχιστος αριθμός δειγμάτων ανά

φύλλο ή κλάδεμα (pruning), οι οποίες στοχεύουν στη βελτίωση της γενικευσιμότητας του μοντέλου. (James et al., 2023)

3.2.3 *Random Forest*

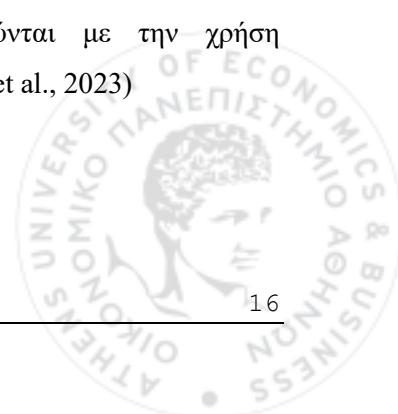
Ο αλγόριθμος Random Forest είναι επίσης για προβλήματα ταξινόμησης και παλινδρόμησης. Είναι μία μέθοδος ensemble μάθησης και συνδυάζει πολλαπλά δέντρα αποφάσεων τα οποία εκπαιδεύονται με ένα τυχαίο υποσύνολο των δεδομένων και μπορούν να χρησιμοποιηθούν συλλογικά για την πραγματοποίηση προβλέψεων. Ουσιαστικά αποτελείται από ένα σύνολο decision trees, οι ξεχωριστές τους προβλέψεις συνδυάζονται μέσω πλειοψηφικής ψήφου ή μέσου όρου, και έχουν ως αποτέλεσμα μικρότερο γενικευμένο σφάλμα από ένα μόνο δέντρο. (Breiman, 2001)

Τα κύρια πλεονεκτήματά του αλγορίθμου είναι ότι μπορεί να διαχειρίζεται πολλές ανεξάρτητες μεταβλητές, έχει ανθεκτικότητα στο overfitting και είναι ταχύτερος από άλλες ensemble μεθόδους. Επιπλέον ο αλγόριθμος Random Forest διαθέτει ενσωματωμένα εργαλεία που δείχνουν την σημασία των μεταβλητών. Η σημασία ενός χαρακτηριστικού μπορεί να αξιολογηθεί παρατηρώντας την αύξηση του σφάλματος πρόβλεψης όταν οι τιμές του αναδιατάσσονται τυχαία στα δείγματα που δεν χρησιμοποιήθηκαν στην εκπαίδευση. (Breiman, 2001)

3.2.4 *Extreme Gradient Boosting (XGBoost)*

Το XGBoost είναι ένας αλγόριθμος της τεχνικής boosting που εφαρμόζεται σε μοντέλα που βασίζονται σε δέντρα και λειτουργεί δημιουργώντας αργά ένα σύνολο ασθενών μοντέλων, που συχνά είναι μικρά δέντρα αποφάσεων, τα οποία εκπαιδεύονται σε μια ακολουθία για να διορθώνουν τα σφάλματα που έγιναν από το προηγούμενο μοντέλο. Η διαφορά με τα decision trees, είναι ότι το boosting δεν εκπαιδεύει τα δέντρα ανεξάρτητα αλλά κάθε νέο δέντρο προσαρμόζεται στα κατάλοιπα (residuals) του τρέχοντος μοντέλου και ακολουθεί τη λογική της βαθμιαίας ελαχιστοποίησης μιας συνάρτησης απώλειας. (James et al., 2023). Το XGBoost επεκτείνει την κλασική προσέγγιση του gradient boosting και φέρνει σαφή κανονικοποίηση, διαχείριση πολυπλοκότητας δέντρων και γρήγορες τεχνικές υπολογισμού, με αποτέλεσμα σημαντικά βελτιωμένη ακρίβεια πρόβλεψης και ανθεκτικότητα στο overfitting (Chen & Guestrin, 2016).

Πλεονεκτήματα του μοντέλου αυτού είναι η πολύ καλή ακρίβεια στην απόδοση και η μοντελοποίηση μη γραμμικών σχέσεων και πολύπλοκων αλληλεπιδράσεων. Ο έλεγχος της πολυπλοκότητας του μοντέλου και η βελτίωση της γενικευσιμότητάς του πραγματοποιούνται με την χρήση υπερπαραμέτρων, όπως το βάθος των δέντρων και ο ρυθμός μάθησης. (James et al., 2023)



3.2.5 *Support Vector Machines (SVM)*

Η μέθοδος Support Vector Machines (SVM) χρησιμοποιείται για προβλήματα ταξινόμησης και παλινδρόμησης. Στην περίπτωση της ταξινόμησης, ο κύριος στόχος ενός SVM είναι να προσδιορίσει ένα όριο απόφασης (hyperplane) που διαχωρίζει τις κλάσεις μεγιστοποιώντας το περιθώριο (margin), δηλαδή την απόσταση μεταξύ του ορίου και των πλησιέστερων σημείων δεδομένων. Τα σημεία που καθορίζουν το περιθώριο και επηρεάζουν το μοντέλο αναφέρονται ως διανύσματα υποστήριξης (support vectors) και το άλλο μέρος της απόφασης βασίζεται σε αυτά.. Στην περίπτωση που οι κλάσεις δεν είναι πλήρως διαχωρίσιμες, εισάγεται μια παράμετρος σφάλματος η οποία τιμωρεί τα σημεία εντός του περιθωρίου ή τα σημεία που έχουν ταξινομηθεί λανθασμένα. Αυτό λειτουργεί ως βασικός μηχανισμός ελέγχου της πολυπλοκότητας του μοντέλου. Επιπλέον, τα SVM μπορούν να δημιουργήσουν μη γραμμικές αποφάσεις χρησιμοποιώντας πυρήνες (kernels) αντί για το γραμμικό εσωτερικό γινόμενο. Η επιλογή του πυρήνα, οι παράμετροί και η τιμή κόστους καθορίζουν εάν το μοντέλο έχει υποπροσαρμογή (underfitting) ή υπερπροσαρμογή (overfitting) και συνήθως επιλέγονται χρησιμοποιώντας τεχνικές αναδειγματοληψίας. Τα SVM είναι ένα υποσύνολο μεθόδων πυρήνα. Αναπτύχθηκαν για να χρησιμοποιηθούν σε αυστηρή ταξινόμηση και είναι επιρρεπείς σε μη πληροφοριακά δεδομένα. (Kuhn & Johnson, 2013, pp. 343–350)

Πλεονεκτήματα του μοντέλου είναι η καλή απόδοση του σε πολλά προβλήματα, επηρεάζεται από τις παρατηρήσεις που βρίσκονται στο περιθώριο ή παραβιάζουν το περιθώριο και πολύ λιγότερο από τις παρατηρήσεις μακριά από hyperplane. Λειτουργεί επίσης καλά σε χώρους μεγάλης διάστασης. Παρ' όλα αυτά, το μοντέλο είναι ευαίσθητο στην παρουσία μη πληροφοριακών μεταβλητών και στη ρύθμιση των παραμέτρων κόστους και πυρήνα, τα οποία εάν δεν γίνουν σωστά μπορεί να οδηγήσει σε overfitting. (Kuhn & Johnson, 2013, pp. 343–350)

3.2.6 *Artificial Neural Network (ANN)*

Το μοντέλο Artificial Neural Networks (ANN) χρησιμοποιείται για προβλήματα ταξινόμησης και παλινδρόμησης. Για την ταξινόμηση, το μοντέλο κωδικοποιεί τις κλάσεις σε δυαδικές μεταβλητές. Τα νευρωνικά δίκτυα αποτελούνται από επίπεδα μονάδων όπου οι γραμμικοί συνδυασμοί των μεταβλητών εισόδου μετασχηματίζονται με μη γραμμικές συναρτήσεις για να μοντελοποιήσουν πιο σύνθετες σχέσεις. Οι έξοδοι αυτών των ενδιάμεσων μονάδων συνδυάζονται για να δημιουργήσουν τις τελικές προβλέψεις και στην περίπτωση πολλαπλών κλάσεων, λαμβάνεται ο μετασχηματισμός softmax για να διατηρούνται οι έξοδοι του μοντέλου συγκρίσιμες. Το νευρωνικό δίκτυο εκπαιδεύεται τροποποιώντας τα βάρη του μοντέλου για να ελαχιστοποιηθεί μια μετρική σφάλματος όπως το άθροισμα των τετραγώνων των σφαλμάτων ή η συνάρτηση της εντροπίας. Οι διαφορές απόδοσης μεταξύ των δύο μεθόδων είναι γενικά χαμηλές. (Kuhn & Johnson, 2013, pp. 333–338)

Πλεονέκτημα του μοντέλου είναι η ικανότητα μοντελοποίησης πολύπλοκων σχέσεων και η καλή του απόδοση. Μειονέκτημα αποτελεί το γεγονός ότι τα νευρωνικά δίκτυα τείνουν να υπερεκπαιδούνται όταν το μοντέλο είναι μεγάλο και περίπλοκο. Η ακρίβεια του ANN εξαρτάται από την επιλογή των παραμέτρων του μοντέλου και των μη απαραίτητων ή σχετικών μεταβλητών. (Kuhn & Johnson, 2013, pp. 333–338)

3.3 Αξιολόγηση Μοντέλων

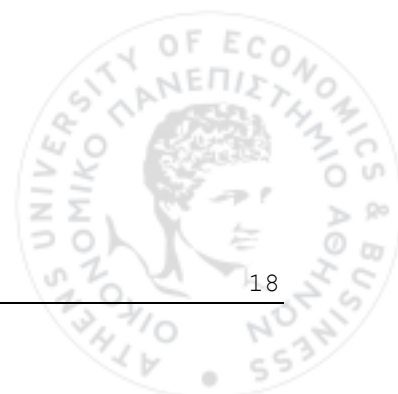
Στη μηχανική μάθηση η αξιολόγηση της απόδοσης ενός μοντέλου είναι απαραίτητη. Διαφορετικές μετρικές απόδοσης μπορεί να οδηγούν σε διαφορετική αξιολόγηση για το πόσο καλό είναι ένα μοντέλο. Η απλή ακρίβεια (accuracy) δεν παρέχει πάντα επαρκείς πληροφορίες (Fawcett, 2006). Έχουν αναπτυχθεί και χρησιμοποιούνται και άλλες μετρικές, όπως η ανάκληση (recall), το F-score και οι καμπύλες ROC με το αντίστοιχο εμβαδό κάτω από την καμπύλη (AUC), τα οποία είναι βασισμένα στον πίνακα σύγχυσης (confusion matrix) και καταγράφουν διάφορες διαστάσεις της απόδοσης ενός μοντέλου ταξινόμησης (classifier).

3.3.1 Confusion Matrix

Όταν έχουμε ένα δυαδικό πρόβλημα ταξινόμησης κάθε παρατήρηση θα λάβει ένα από τα 4 πιθανά αποτελέσματα: πραγματικά θετικές (true positive), ψευδώς αρνητικές (false negative), πραγματικά αρνητικές (true negative) και ψευδώς θετικές (false positive). Αυτά τα αποτελέσματα επιτρέπουν την κατασκευή ενός πίνακα 2 x 2 που ονομάζεται confusion matrix, ο οποίος αντιπροσωπεύει τη σχέση μεταξύ της πραγματικής κλάσης και της προβλεπόμενης κλάσης σε ένα σύνολο δεδομένων. Τα πιο δημοφιλή μέτρα αξιολόγησης, όπως το ποσοστό αληθώς θετικών και ψευδώς θετικών, υπολογίζονται χρησιμοποιώντας αυτόν τον πίνακα. (Fawcett, 2006)

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Εικόνα 1. Confusion Matrix



3.3.2 Accuracy

Το accuracy είναι η αναλογία του σωστά ταξινομημένου ποσοστού των παρατηρήσεων. Το υπολογίζουμε προσθέτοντας τις πραγματικά θετικές και τις πραγματικά αρνητικές προβλέψεις και τις διαιρούμε με το συνολικό αριθμό παρατηρήσεων. Το accuracy είναι ένα γενικό μέτρο της αποτελεσματικότητας ενός classifier αλλά δεν διακρίνει τα σφάλματα στις θετικές και αρνητικές κλάσεις. Αν και είναι το μέτρο που εφαρμόζεται πιο συχνά μπορεί να μην είναι αξιόπιστο, ειδικά στην περίπτωση άνισης κατανομής των κλάσεων καθώς είναι ένα μέτρο που σχετίζεται άμεσα με τον αριθμό των σωστών αρνητικών προβλέψεων. (Sokolova & Lapalme, 2009)

$$\text{Accuracy} = \text{TP} + \text{TN} / \text{TP} + \text{TN} + \text{FP} + \text{FN}$$

3.3.3 Recall

Το recall, το οποίο αναφέρεται και ως sensitivity, δείχνει πόσο καλά ένας classifier εντοπίζει σωστά τις πραγματικές περιπτώσεις. Είναι ο λόγος των true positive προβλέψεων προς τον συνολικό αριθμό των πραγματικών θετικών περιπτώσεων. Το recall εστιάζει αποκλειστικά στη θετική κλάση και δεν λαμβάνει υπόψη τις σωστές ή λανθασμένες αρνητικές προβλέψεις. Δεν μπορεί να χρησιμοποιηθεί από μόνο του διότι αγνοεί και άλλες πτυχές της απόδοσης. (Sokolova & Lapalme, 2009)

$$\text{Recall} = \text{TP} / \text{TP} + \text{FN}$$

3.3.4 Precision

Το precision είναι ο τρόπος με τον οποίο ένας classifier αντιστοιχίζει σωστά την θετική ετικέτα. Υπολογίζεται ως το κλάσμα των true positives προβλέψεων προς το άθροισμα των true positives και false positives. Η μετρική αυτή δεν λαμβάνει υπόψη τις αρνητικές προβλέψεις οπότε από μόνη της δεν επαρκεί για την αξιολόγηση ενός μοντέλου ταξινόμησης. (Sokolova & Lapalme, 2009)

$$\text{Precision} = \text{TP} / \text{TP} + \text{FP}$$

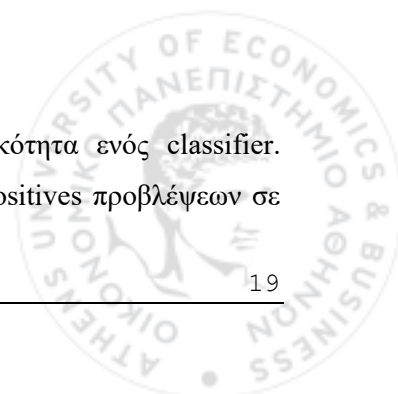
3.3.5 F1-score

Το F-score είναι ο αρμονικός μέσος του precision και του recall και ανήκει στις μετρικές που εστιάζουν στη θετική κλάση. Το F-score εστιάζει κυρίως στη σωστή αναγνώριση των θετικών προβλέψεων (Sokolova & Lapalme, 2009)

$$\text{F-score} = 2 * \text{Precision} * \text{Recall} / \text{Precision} + \text{Recall}$$

3.3.6 ROC (Receiver Operating Characteristic) Curves και AUC

Η καμπύλη ROC χρησιμοποιείται για να απεικονίσει την αποτελεσματικότητα ενός classifier. Παρουσιάζει το ποσοστό true positives προβλέψεων και το ποσοστό false positives προβλέψεων σε



διάφορα σημεία απόφασης. Οι ROC καμπύλες είναι ανεξάρτητες από διάφορες κατανομές κλάσεων και κόστος σφαλμάτων, επιτρέποντας τη σύγκριση διαφορετικών classifiers υπό διαφορετικές συνθήκες. Το εμβαδό κάτω από την καμπύλη ROC (AUC) αντιπροσωπεύει την πιθανότητα ένα μοντέλο να αντιστοιχίσει μια δεδομένη πραγματική περίπτωση υψηλότερα από μια δεδομένη αρνητική περίπτωση. Ένα υψηλό AUC υποδηλώνει γενικά καλύτερη απόδοση, ωστόσο ένας classifier με υψηλότερο AUC μπορεί να υστερεί σε συγκεκριμένες περιοχές της ROC καμπύλης, ανάλογα με το επιλεγμένο σημείο λειτουργίας. (Fawcett, 2006)

3.4 Μετρικές Δικαιοσύνης και Διαφάνειας

3.4.1 *Disparate Impact Ratio (DIR)*

Η μετρική disparate impact ratio (DIR) είναι ο λόγος της πιθανότητας έγκρισης για τις μη προστατευόμενες και τις προστατευόμενες ομάδες. Μια τιμή κοντά στο 1 συνεπάγεται έναν ιδανικό βαθμό δικαιοσύνης, ενώ τιμές μικρότερες από το 1 υποδηλώνουν πλεονέκτημα για την προνομιούχα ομάδα και τιμές μεγαλύτερες από το 1 υποδηλώνουν πλεονέκτημα για τη μη προνομιούχα ομάδα. Σε μια πιο ευέλικτη προσέγγιση το διάστημα (0.8, 1.25) θεωρείται αποδεκτό ώστε ένας ταξινομητής να θεωρείται δίκαιος. (Moldovan, 2023)

$$DIR = P(\hat{y} = |Unprivileged) / P(\hat{y} = |Privileged)$$

Υπάρχουν όμως κάποιοι σημαντικοί περιορισμοί γιατί το κριτήριο στο οποίο βασίζεται δηλαδή η σύγκριση των ποσοστών αποδοχής μεταξύ ομάδων, δεν επαληθεύει την ορθότητα των αποφάσεων του μοντέλου αλλά μετρά μόνο τον αριθμό των ατόμων που λαμβάνουν θετική απόφαση σε κάθε ομάδα. Το σύστημα μπορεί να φαίνεται δίκαιο ως προς την μετρική DIR αποκλείοντας τους πιο καλούς υποψηφίους από την προστατευόμενη ομάδα κάτι που δεν ενισχύει τη δικαιοσύνη. Ακόμη η μετρική σχεδόν πάντα δεν είναι συμβατή με άλλες μετρικές δικαιοσύνης όπως τα equalized odds ειδικά όταν τα ποσοστά εμφάνισης της θετικής κλάσης δεν είναι ίσα μεταξύ αυτών των ομάδων. (Solon Barocas et al., 2019)

3.4.2 *Equal Opportunity Difference (EOD)*

Στο equal opportunity difference το ποσοστό των αληθώς θετικών αποφάσεων πρέπει να είναι ίσο μεταξύ προστατευόμενων και μη προστατευόμενων ομάδων. Στο πλαίσιο της πιστωτικής βαθμολόγησης αυτό σημαίνει ότι ο ταξινομητής θα πρέπει να έχει το ίδιο ποσοστό σφάλματος όταν προτείνει την έγκριση δανείων τόσο στις προστατευόμενες όσο και στις μη προστατευόμενες ομάδες. Η απαίτηση για εξισωμένα σφάλματα ασκεί πίεση στους υπεύθυνους λήψης αποφάσεων να βελτιώσουν τα ποσοστά λανθασμένης ταξινόμησης, βελτιστοποιώντας τα μοντέλα και αυξάνοντας την

ποιότητα των δεδομένων. Το διάστημα δικαιοσύνης που λαμβάνεται υπόψη για αυτή τη μετρική είναι $(-0.1, 0.1)$. (Moldovan, 2023)

$$EOD = TPR_{unprivileged} - TPR_{privileged}$$

3.4.3 Equalized odds (EOdds)

Ένας προβλεπτικός αλγόριθμος ικανοποιεί το κριτήριο των equalized odds όταν το ποσοστό των πραγματικών θετικών προβλέψεων και το ποσοστό των ψευδώς θετικών προβλέψεων είναι ίδιοι μεταξύ των ομάδων. Οι equalized odds επιτρέπουν στο προβλεπόμενο αποτέλεσμα να εξαρτάται από το προστατευόμενο χαρακτηριστικό αλλά μόνο μέσω της μεταβλητής-στόχου. Αυτό σημαίνει ότι τα άτομα με καλό πιστωτικό προφίλ και εκείνα με κακό θα πρέπει να έχουν παρόμοια ταξινόμηση ανεξάρτητα από το αν ανήκουν στην προστατευόμενη ή στη μη προστατευόμενη ομάδα. Έτσι ένα μοντέλο πιστωτικής βαθμολόγησης θεωρείται δίκαιο εάν ο προβλεπτικός μηχανισμός έχει ίσους ρυθμούς αληθώς θετικών (δηλαδή την πιθανότητα ένα πραγματικά θετικό άτομο να αναγνωριστεί ως τέτοιο) και ίσους ρυθμούς ψευδώς θετικών (δηλαδή την πιθανότητα να εγκριθεί εσφαλμένα μια αρνητική περίπτωση). Μια λιγότερο αυστηρή εκδοχή των equalized odds είναι να απαιτείται μη διάκριση μόνο εντός της ομάδας με θετικό αποτέλεσμα. Δηλαδή να απαιτείται τα άτομα με καλό πιστωτικό προφίλ να έχουν ίση ευκαιρία να λάβουν το δάνειο εξαρχής. Αυτή η χαλάρωση συχνά ονομάζεται equal opportunity. (Hurlin et al., 2021)

$$P(\hat{y} = 1 | Y = y, A = a) = P(\hat{y} = 1 | Y = y, A = b) \text{ for all } y \in \{0, 1\}$$

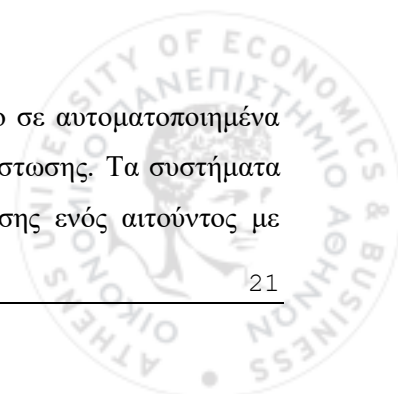
3.4.4 Statistical Parity Difference (SPD)

Η statistical parity difference μετρά τη διαφορά μεταξύ των πιθανοτήτων έγκρισης στις προστατευόμενες και στις μη προστατευόμενες ομάδες. Μια τιμή κοντά στο μηδέν συνεπάγεται το ίδιο ποσοστό έγκρισης και για τις δύο ομάδες. Το εύρος δικαιοσύνης για αυτή τη μετρική θεωρείται ότι βρίσκεται στο διάστημα $(-0.1, 0.1)$ (Moldovan, 2023). Η μετρική είναι μια από τις ευκολότερες και πιο κατανοητές τεχνικές για την ανάλυση της δικαιοσύνης. Όμως δεν κοιτάει τίποτα άλλο για το άτομο εκτός από το σε ποια ομάδα ανήκει, για παράδειγμα άνδρας/γυναίκα. Επομένως αυτός ο τρόπος μέτρησης μπορεί να κρύψει αδικία. (Verma & Rubin, 2018)

$$SPD = P(\hat{y} = 1 | A = Unprivileged) - P(\hat{y} = 1 | A = Privileged)$$

3.5 Αντιμετώπιση Μεροληψίας (Bias mitigation)

Τα σύγχρονα χρηματοπιστωτικά συστήματα βασίζονται όλο και περισσότερο σε αυτοματοποιημένα συστήματα για να πάρουν αποφάσεις σχετικά με την έγκριση ή απόρριψη πίστωσης. Τα συστήματα αξιολόγησης πιστοληπτικής ικανότητας προβλέπουν την πιθανότητα αθέτησης ενός αιτούντος με

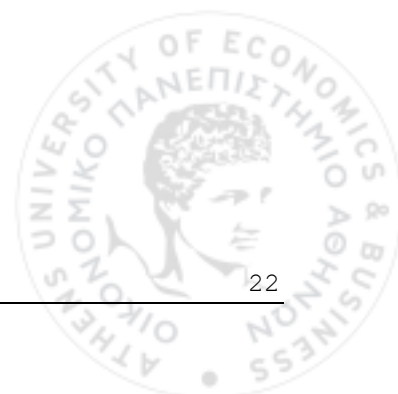


βάση την προηγούμενη οικονομική του συμπεριφορά και τα δημογραφικά του δεδομένα. Οι παραδοσιακές μέθοδοι χρησιμοποιούν μεροληπτικά ή ελλιπή δεδομένα για να λάβουν αποφάσεις που διατηρούν τις υπάρχουσες κοινωνικές ανισότητες και επηρεάζουν περισσότερο ορισμένες ομάδες. Οι αυτοματοποιημένες διαδικασίες προκαλούν ηθικό προβληματισμό επειδή δημιουργούν διακρίσεις και στερούνται διαφάνειας στις λειτουργίες τους. Η τεχνητή νοημοσύνη και η μηχανική μάθηση έχουν οδηγήσει στη δυνατότητα αύξησης της αποτελεσματικότητας και της δικαιοσύνης των συστημάτων αξιολόγησης πιστοληπτικής ικανότητας. (K. et al., 2025)

Στην βιβλιογραφία για την αντιμετώπιση μεροληψίας ακολουθείται μια σειρά. Εντοπίζεται πρώτα η μεροληψία χρησιμοποιώντας μετρικές δικαιοσύνης και έπειτα εφαρμόζεται κάποια τεχνική αντιμετώπισης της μεροληψίας για τη μείωση της αδικίας. Το τελικό βήμα της διαδικασίας περιλαμβάνει τη σύγκριση των αποτελεσμάτων πριν και μετά τη διαδικασία για να εκτιμηθεί η μεταβολή στη δικαιοσύνη και η επίδραση στην απόδοση του μοντέλου. (Moldovan, 2023)

Οι μέθοδοι αντιμετώπισης της μεροληψίας χωρίζονται σε τρεις κατηγορίες, ανάλογα με το σημείο της διαδικασίας εκπαίδευσης και αξιοποίησης του μοντέλου στο οποίο εφαρμόζονται. Οι τεχνικές προεπεξεργασίας (pre-processing) είναι η πρώτη κατηγορία οι οποίες εφαρμόζονται πριν από την εκπαίδευση και επικεντρώνονται σε παρεμβάσεις στα δεδομένα για την ελαχιστοποίηση της μεροληψίας πριν αυτά μεταφερθούν στο μοντέλο. Η δεύτερη κατηγορία είναι οι τεχνικές κατά την επεξεργασία (in-processing) όπου η δικαιοσύνη εισάγεται ενεργά στη διαδικασία εκπαίδευσης με περιορισμούς δικαιοσύνης (fairness constraints) ή συγκεκριμένες τροποποιήσεις του αλγορίθμου. Τέλος οι τεχνικές μετά την επεξεργασία (post-processing) που εφαρμόζονται μετά την εκπαίδευση, τροποποιώντας τις τελικές προβλέψεις του μοντέλου για την αντιμετώπιση της μεροληψίας. (Moldovan, 2023)

Οι τεχνικές pre-processing και post-processing μπορούν να χρησιμοποιηθούν ανεξάρτητα από τον αλγόριθμο ταξινόμησης καθώς λειτουργούν σε επίπεδο δεδομένων ή σε επίπεδο αποτελέσματος. Ενώ οι in-processing μέθοδοι είναι συνήθως πιο συσχετισμένες με συγκεκριμένους αλγορίθμους καθώς η παρέμβαση γίνεται κατά τη διάρκεια της εκπαίδευσης. Παρά την αποτελεσματικότητα των τεχνικών in-processing αυτές απαιτούν πιο περίπλοκη υλοποίηση και περιορίζουν την ελευθερία επιλογής μοντέλων και αυτό καθίσταται σημαντικό ζήτημα σε τομείς όπως η αξιολόγηση πιστοληπτικής ικανότητας. (Moldovan, 2023)



3.5.1 *Pre-processing τεχνικές*

3.5.1.1 *Reweighting*

Μία από τις παραδοσιακές μεθόδους pre-processing για την αντιμετώπιση της μεροληψίας είναι η μέθοδος της reweighting. Αναθέτει διαφορετικά βάρη στις μεταβλητές του dataset χωρίς να αλλάζει τις ετικέτες ή τις τιμές των χαρακτηριστικών για να μειώσει τις διακρίσεις μετατοπίζοντας τα βάρη μεταξύ των προστατευόμενων και μη προστατευόμενων ομάδων. (Moldovan, 2023). Η μέθοδος έχει το σημαντικό πλεονέκτημα ότι χρησιμοποιείται πριν από την εκπαίδευση έτσι ώστε να μπορεί να γίνει ανεξάρτητα από τον αλγόριθμο ταξινόμησης και σε πολλές περιπτώσεις μπορεί να χρησιμοποιηθεί για τη μείωση της μεροληψίας χωρίς σημαντική μείωση στην απόδοση του μοντέλου. Ωστόσο πειραματικά στοιχεία δείχνουν ότι η μέθοδος μπορεί να μην είναι πάντα σταθερή καθώς μπορεί να βελτιώσει ορισμένους δείκτες δικαιοσύνης ενώ άλλοι να παραμένουν οι ίδιοι, πράγμα που εξαρτάται κάθε φορά από το σύνολο δεδομένων και τον εφαρμοζόμενο ταξινομητή (Mariscal et al., 2024).

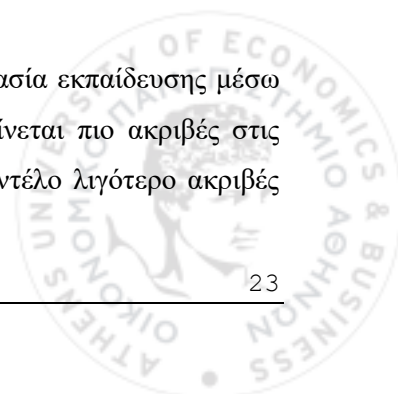
3.5.1.2 *Disparate Impact Remover (DIR)*

Το disparate impact remover επιδιώκει να μειώσει τις διακρίσεις πριν από την εκπαίδευση ενός μοντέλου. Η μέθοδος εντοπίζει το disparate impact και επιχειρεί να διορθώσει το σύνολο δεδομένων καθώς ο αλγόριθμος ανιχνεύει το disparate impact και προσπαθεί να διορθώσει τα δεδομένα ώστε να επιτευχθεί δικαιοσύνη. Η διόρθωση των δεδομένων πραγματοποιείται με τρόπο ώστε να διατηρείται όσο το δυνατόν περισσότερο η προβλεπτική χρησιμότητα και η κατάταξη των χαρακτηριστικών. (Moldovan, 2023). Σύμφωνα με τη βιβλιογραφία, η τεχνική μπορεί να εφαρμοστεί μόνο μέσω προσαρμογής αριθμητικών μεταβλητών με το προστατευόμενο χαρακτηριστικό και τη μεταβλητή-στόχο να παραμένουν σταθερά ενώ επιδιώκει την ενίσχυση της ομαδικής δικαιοσύνης και τη διατήρηση του rank ordering. Η αποτελεσματικότητά του δεν είναι η ίδια σε όλα τα μοντέλα και σε όλα τα datasets επειδή σε ορισμένες περιπτώσεις μπορεί να μειώσει τη μεροληψία ενώ σε άλλες μπορεί να επιδεινώσει συγκεκριμένες μετρικές δικαιοσύνης. (Mariscal et al., 2024). Για αυτό το DIR μπορεί να εφαρμοστεί και ως partial repair για να υπάρχει trade-off μεταξύ fairness και accuracy (Moldovan, 2023).

3.5.2 *In-processing τεχνικές*

3.5.2.1 *Adversarial Debiasing*

Η τεχνική adversarial debiasing περιλαμβάνει δίκαιη μεταχείριση στη διαδικασία εκπαίδευσης μέσω ενός μηχανισμού adversarial learning. Το μοντέλο εκπαιδεύεται ώστε να γίνεται πιο ακριβές στις προβλέψεις του και ταυτόχρονα προσπαθεί να κάνει ένα αντιπαραθετικό μοντέλο λιγότερο ακριβές



στο προβλεπόμενο προστατευόμενο χαρακτηριστικό ελαχιστοποιώντας την πιθανότητα του αντιπαραθετικού μοντέλου να προβλέψει το προστατευόμενο χαρακτηριστικό. (Moldovan, 2023). Πρακτικά η μέθοδος βασίζεται σε δύο δίκτυα, όπου το predictor εκπαιδεύεται σε μη προστατευόμενα χαρακτηριστικά για να προβλέψει το y και ένα δεύτερο δίκτυο (adversary) προσπαθεί να προβλέψει το προστατευόμενο χαρακτηριστικό με βάση το αποτέλεσμα του predictor. Η μέθοδος φαίνεται να μειώνει τη μεροληψία, ειδικά την μετρική Demographic Parity Difference, αλλά μπορεί να οδηγήσει σε μείωση του AUC. (Mariscal et al., 2024).

3.5.2.2 *Meta Fair Classifier*

Ο meta fair classifier είναι μια προσέγγιση που επιδιώκει να επιλύσει πρόσθετα προβλήματα ταξινόμησης με περιορισμούς δικαιοσύνης. Η μέθοδος βασίζεται στην ιδέα της ανάπτυξης ενός αλγορίθμου που εφαρμόζεται σε ένα ευρύ φάσμα προβλημάτων ταξινόμησης και ενσωματώνει fairness constraints στη διαδικασία της εκπαίδευσης. Ο χρήστης μπορεί να προσαρμόσει την παράμετρο περιορισμού για να ελέγξει το επίπεδο αυστηρότητας του fairness constraint και να επιτύχει έναν πρακτικό συμβιβασμό μεταξύ fairness και accuracy. Ο αλγόριθμος είναι πιο ευέλικτος επειδή ο χρήστης μπορεί να καθορίσει πόσο σημαντική θα πρέπει να είναι η μετρική δικαιοσύνης. Αν και ο meta fair classifier έχει καλά αποτελέσματα δεν είναι πάντα ένας σταθερός αλγόριθμος. Ο αλγόριθμος μπορεί να εμφανίζει σημαντικές διαφοροποιήσεις στα αποτελέσματα όταν εκτελείται αρκετές φορές στο ίδιο σύνολο δεδομένων και σε ορισμένες περιπτώσεις δεν πληροί τους περιορισμούς δικαιοσύνης. (Moldovan, 2023).

3.5.3 *Post-processing τεχνικές*

3.5.3.1 *Reject Option Classification*

Η μέθοδος reject option classification είναι από τις πρώτες post-processing τεχνικές που προτάθηκαν για τη μείωση διακρίσεων. Εφαρμόζεται μετά την εκπαίδευση ενός μοντέλου και χρησιμοποιεί τις εκ των υστέρων πιθανότητες (posterior probabilities) για να τροποποιήσει τις ετικέτες εξόδου με σκοπό την ελαχιστοποίηση των διακρίσεων. Έμφαση δίνεται σε καταστάσεις που εμπίπτουν σε μια κρίσιμη περιοχή δηλαδή σε περιπτώσεις όπου το μοντέλο είναι πιο αβέβαιο και επομένως πιο πιθανό να επηρεαστεί από μεροληψία. Σε αυτήν την περίπτωση οι παρατηρήσεις που βρίσκονται στη κρίσιμη περιοχή μπορούν να επαναχαρακτηριστούν (relabeling) με βάση το αν ανήκουν σε προστατευόμενη ή μη προστατευόμενη ομάδα. Η διαδικασία αναπτύσσει δύο διαφορετικούς πίνακες ευαισθητούς στο κόστος, έναν για τις deprived ομάδες και έναν για τις favored ομάδες, και η τελική βελτίωση πραγματοποιείται βελτιστοποιώντας τις συναρτήσεις απώλειας. Ακόμη η μέθοδος παρέχει ένα μέσο συμβιβασμού μεταξύ δικαιοσύνης και ακρίβειας με έναν συντελεστή trade-off (θ) ο οποίος αποδίδει

το βάρος της μείωσης των διακρίσεων και της διατήρησης της απόδοσης του μοντέλου. (Moldovan, 2023)

3.5.3.2 *Calibrated Equalized Odds Post-processing*

Η τεχνική calibrated equalized odds post-processing τροποποιεί τις τελικές ετικέτες μετά την ταξινόμηση με στόχο τη διατήρηση της δικαιοσύνης και βασίζεται στην έννοια του equalized odds. Με αυτή την μέθοδο επιδιώκεται η αντιμετώπιση της μεροληψίας τροποποιώντας το τελικό αποτέλεσμα χωρίς να χρειάζεται να τροποποιηθούν τα δεδομένα ή να επανεκπαιδευτεί το μοντέλο. Ένα στοιχείο που έχει ενσωματωθεί είναι η calibration η οποία παρέχει τη δυνατότητα επίτευξης δικαιοσύνης μεταξύ των προστατευόμενων και μη προστατευόμενων ομάδων χωρίς να ενθαρρύνεται η χρήση του προστατευόμενου χαρακτηριστικού με τρόπο που θα οδηγούσε σε διακρίσεις. Η προσέγγιση παρέχει επίσης στον χρήστη την ευελιξία να επιλέξει το επίπεδο του fairness constraint, κάτι που είναι χρήσιμο επειδή η εφαρμογή calibration σε ορισμένες περιπτώσεις μειώνει την ακρίβεια και ο χρήστης πρέπει να επιλέξει μεταξύ δικαιοσύνης και απόδοσης (Moldovan, 2023).

3.6 *Proxy μεταβλητές*

Proxy μεταβλητή είναι μια μεταβλητή που περιέχει πληροφορίες που μπορούν να χρησιμοποιηθούν για την έμμεση εξαγωγή ενός προστατευόμενου χαρακτηριστικού πχ φύλο, ηλικία, εθνικότητα, χωρίς να αποτελεί η ίδια προστατευόμενο χαρακτηριστικό. Ακόμη και όταν το προστατευόμενο χαρακτηριστικό αφαιρεθεί από τα δεδομένα οι πληροφορίες θα συνεχίσουν να βρίσκονται σε άλλα χαρακτηριστικά και το μοντέλο θα είναι σε θέση να επηρεάσει τις προβλέψεις του ανάλογα με την ομάδα του ατόμου με έμμεσο τρόπο. Σύμφωνα με αυτή την ιδιότητα η απλή αφαίρεση του προστατευόμενου χαρακτηριστικού (fairness through unawareness) δεν μπορεί να εξαλείψει τις διακρίσεις. Η ύπαρξη μιας proxy μεταβλητής σημαίνει ότι υπάρχουν πλεονάζουσες κωδικοποιήσεις (redundant encodings) που επιτρέπουν την επαναφορά της προστατευόμενης μεταβλητής χρησιμοποιώντας τα υπόλοιπα χαρακτηριστικά. Επομένως ένα μοντέλο μπορεί να συνεχίζει να εμφανίζει μεροληψία ακόμη και σε περιπτώσεις όπου η προστατευόμενη μεταβλητή δεν αποτελεί άμεση είσοδο στο σύστημα. (Hardt et al., 2016)

Σε μια τυπική διαδικασία επιβλεπόμενης μάθησης τα μοντέλα προβλέπουν το αποτέλεσμα χρησιμοποιώντας τα δεδομένα ώστε να προβλέψουν το πραγματικό αποτέλεσμα ενώ υπάρχει ένα προστατευόμενο χαρακτηριστικό. Ένα θεωρητικό κριτήριο για να αποφευχθούν διακρίσεις είναι να διασφαλιστεί ότι η πρόβλεψη δεν εξαρτάται από το προστατευόμενο χαρακτηριστικό πέρα από ό,τι δικαιολογείται από το πραγματικό αποτέλεσμα. Όταν το προστατευόμενο χαρακτηριστικό ή proxies του χρησιμοποιούνται για τη βελτίωση της ακρίβειας της πρόβλεψης μπορεί να οδηγήσει σε

διαφορετικές αποφάσεις μεταξύ ομάδων που δεν σχετίζονται με το πραγματικό αποτέλεσμα που επιδιώκουμε να προβλέψουμε. (Hardt et al., 2016)

Στην πραγματικότητα το πρόβλημα με τις proxy μεταβλητές φαίνεται πολύ έντονα σε εφαρμογές όπως η αξιολόγηση πιστοληπτικής ικανότητας. Παρόλο που μια μεταβλητή δεν είναι προστατευόμενο χαρακτηριστικό μπορεί να συσχετιστεί με αυτήν με τέτοιο τρόπο ώστε το μοντέλο να τη χρησιμοποιεί για να λαμβάνει τις αποφάσεις του στις δύο ομάδες διαφορετικά έστω και έμμεσα. Για παράδειγμα μεταβλητή όπως η τοποθεσία ή η περιοχή μπορεί να χρησιμεύσει ως υποκατάστατο της εθνικότητας. Επομένως η διάκριση δεν συμβαίνει επειδή το μοντέλο βλέπει το προστατευόμενο χαρακτηριστικό αλλά επειδή μπορεί να το ανακατασκευάσει χρησιμοποιώντας άλλα δεδομένα. Επομένως με την απαγόρευση χρήσης των προστατευόμενων χαρακτηριστικών (input scrutiny) δεν λύνεται το πρόβλημα. Παρόλο που το προστατευόμενο χαρακτηριστικό μπορεί να αφαιρεθεί από τα δεδομένα άλλες μεταβλητές μπορεί να εξακολουθούν να περιέχουν πληροφορίες που δημιουργούν μεροληψία. Μια προσέγγιση σχετικά με τις proxy μεταβλητές είναι να θεωρούμε ότι μπορεί να χρειάζεται μερική αφαίρεση του proxy signal δηλαδή να μην εξαλειφθεί ολόκληρη η proxy μεταβλητή αλλά να εξαλειφθεί το μέρος της πληροφορίας που προκαλεί άδικες διαφορές μεταξύ προστατευόμενων ομάδων. (Johnson et al., 2016)

3.7 Explainable AI (XAI)

Σε αντίθεση με τα κλασικά στατιστικά μοντέλα, τα μοντέλα μηχανικής μάθησης συχνά χαρακτηρίζονται από περιορισμένη διαφάνεια λόγω της πολυπλοκότητας και της ιδιαίτερης αρχιτεκτονικής τους που καθιστούν αδύνατη την απλή κατανόηση του εσωτερικού τους μηχανισμού. Τα ML μοντέλα προσπαθούν να εντοπίσουν σύνθετα, μη-γραμμικά πρότυπα, βελτιώνοντας δραματικά την ακρίβεια πρόβλεψης, ιδίως σε τομείς όπως είναι η ανάλυση πιστωτικού κινδύνου. Όμως η βελτίωση αυτή στην πρόβλεψη των μοντέλων έχει αντίκτυπο στην ερμηνευσιμότητα των αποτελεσμάτων. (André Aoun Montevechi et al.)

Διαφορετικές ομάδες ενδιαφερομένων προσπαθούν να κατανοήσουν πώς λειτουργεί και πώς λαμβάνει αποφάσεις ένα μοντέλο τεχνητής νοημοσύνης, και η εξηγησιμότητα συνδέεται ακριβώς με τα ερωτήματα αυτά. Για να θεωρηθεί ένα μοντέλο εξηγήσιμο, θα πρέπει να έχει καθαρές, τεκμηριωμένες εξηγήσεις που ξεδιπλώνουν τους εσωτερικούς μηχανισμούς του και τα αποτελέσματα που παράγει, ικανοποιώντας τις ερωτήσεις προγραμματιστών, διοικητικών στελεχών, ελεγκτών και ρυθμιστικών αρχών (Chen et al.). Ένα μοντέλο τεχνητής νοημοσύνης (AI) μπορεί να θεωρηθεί εξηγήσιμο εφόσον καθιστά σαφή τη λογική που οδηγεί στην αξιολόγηση της πιστοληπτικής ικανότητας και ειδικότερα όταν παρέχει ουσιώδεις πληροφορίες για τους παράγοντες που τις διαμορφώνουν (Chen et al.).

Παρατηρείται μια ανησυχία από τις ρυθμιστικές αρχές σχετικά με το κατά πόσο τα σύγχρονα μοντέλα ML μπορούν να προσφέρουν επαρκή ερμηνεία των αποφάσεων στη χορήγηση δανείων. Σε αρκετές χώρες θεσπίζονται νομικές διατάξεις που διασφαλίζουν τη δίκαιη πρόσβαση σε πίστωση, καθώς και την παρουσία ανθρώπινης εποπτείας στη λειτουργία των συστημάτων πιστοληπτικής αξιολόγησης (André Aoun Montevechi et al.). Στην Ευρώπη όπως αναφέραμε αναλυτικά στην προηγούμενη ενότητα τα κανονιστικά νομικά πλαίσια είναι το GDPR και AI Act. Οι απαιτήσεις αυτές καθιστούν αναγκαία την ανάπτυξη διαφανών και επεξηγήσιμων μοντέλων, ώστε οι αποφάσεις να είναι κατανοητές και δικαιολογημένες (André Aoun Montevechi et al.). Επομένως, η διαφάνεια των μοντέλων εμφανίζεται ως σημαντική προτεραιότητα και τα πιστωτικά ιδρύματα τείνουν ακόμα να προτιμούν πιο απλά, κυρίως γραμμικά, στατιστικά μοντέλα, ώστε η επίδραση κάθε μεταβλητής στην πιθανότητα αθέτησης πληρωμών να είναι άμεσα κατανοητή (André Aoun Montevechi et al.).

Γενικά στις διαδικασίες αξιολόγησης της πιστοληπτικής ικανότητας, τα χρηματοπιστωτικά συστήματα παραδοσιακά χρησιμοποιούν τα μοντέλα λογιστικής παλινδρόμησης (logistic regression) (Dastile, Celik, & Potsane, 2020). Η σχετική απλότητά τους επιτρέπει την εύκολη παρακολούθηση της επίδρασης κάθε μεταβλητής. Το μειονέκτημα είναι όμως ότι αυτή η απλότητα τείνει να περιορίζει την προβλεπτική τους ισχύ αφήνοντας την ακρίβεια σε πιο μέτρια επίπεδα. Σε αντίθεση, οι σύγχρονες τεχνικές ML, όπως τα Random Forests, σπρώχνουν την ακρίβεια σε υψηλότερα όρια, αλλά η πολυπλοκότητά τους καθιστά την ερμηνεία τους πιο δύσκολη. Για να ξεπεραστεί αυτή η δυσκολία, προτείνεται η χρήση εργαλείων μεταγενέστερης ανάλυσης, όπως το SHAP και το LIME που προσδίδουν στα ML μοντέλα μεγαλύτερη διαφάνεια και ερμηνευσιμότητα, χωρίς να παραμερίζεται η ακρίβεια των προβλέψεών τους (Chen et al.).

Ο τρόπος με τον οποίο μπορούν να εξηγηθούν τα μοντέλα μηχανικής μάθησης δηλαδή γιατί κατέληξαν σε ένα συγκεκριμένο αποτέλεσμα μπορεί να γίνει σε διαφορετικά επίπεδα κατανόησης. Η εξηγησιμότητα των μοντέλων μπορεί να αξιολογηθεί μέσω δύο συμπληρωματικών αναλυτικών επιπέδων, τα οποία περιλαμβάνουν το παγκόσμιο (global) επίπεδο όπου εστιάζουμε στη συνολική συμπεριφορά του μοντέλου και το τοπικό (local) όπου εστιάζουμε στην εξήγηση μιας συγκεκριμένης πρόβλεψης για μια μεμονωμένη παρατήρηση (Moscatto et al.). Αυτό αποτελεί την βάση για την κατηγοριοποίηση των μεθόδων Explainable AI παρακάτω.

3.7.1 *Partial Dependence Plots (PDP)*

Τα Partial Dependence Plots (PDP) είναι μια μέθοδος ερμηνείας ML μοντέλων και χρησιμοποιούνται για την ανάλυση της επίδρασης κάθε χαρακτηριστικού στην πρόβλεψη που παράγει το μοντέλο. Τα PDPs βασίζονται στην ιδέα ότι για να αναλυθεί η επίδραση ενός συγκεκριμένου χαρακτηριστικού

στην πρόβλεψη του μοντέλου η τιμή του χαρακτηριστικού διατηρείται σε διαδοχικές τιμές, ενώ οι υπόλοιπες μεταβλητές διατηρούνται όπως στα πραγματικά δεδομένα. Σε κάθε μία από αυτές τις τιμές, το μοντέλο προβλέπει όλους τους πιθανούς συνδυασμούς των άλλων χαρακτηριστικών και οι προβλέψεις υπολογίζονται ως μέσοι όροι. Αυτός ο μέσος όρος σε αυτήν τη συγκεκριμένη τιμή του χαρακτηριστικού γίνεται σημείο της καμπύλης PDP. Τα PDPs δίνουν μια συνολική εικόνα της συμπεριφοράς του μοντέλου (global) ανά χαρακτηριστικό και μπορούν να χρησιμοποιηθούν με διάφορα μοντέλα. Ωστόσο επειδή χρησιμοποιεί μέσους όρους, μπορεί να αποκρύψει μεγάλη ετερογένεια στις επιμέρους προβλέψεις και δεν αποτυπώνει πώς λειτουργεί το μοντέλο με ένα συγκεκριμένο άτομο. Ένας περιορισμός τους είναι ότι τα αποτελέσματα που θα δώσουν μπορεί να είναι παραπλανητικά όταν τα χαρακτηριστικά είναι ισχυρά συσχετισμένα. Σε αυτή την περίπτωση τα PDPs χρησιμοποιούν συνδυασμούς που μπορούν να μην εμφανίζονται ποτέ σε πραγματικά δεδομένα οδηγώντας έτσι σε ερμηνείες που δεν ανταποκρίνονται σε ρεαλιστικές καταστάσεις. (Gero Szepannaek & Karsten Lübke, 2023)

3.7.2 *Individual Conditional Expectation (ICE)*

Τα διαγράμματα Individual Conditional Expectation (ICE) σε συνδυασμό με τα διαγράμματα PDP απεικονίζουν τον τρόπο με τον οποίο αλλάζει η πρόβλεψη του μοντέλου μιας μεμονωμένης παρατήρησης όταν τροποποιείται μια συγκεκριμένη μεταβλητή. Τα PDP βασίζονται σε μέσους όρους ενώ τα διαγράμματα ICE δημιουργούν μια ξεχωριστή καμπύλη ανά παρατήρηση αλλάζοντας ένα χαρακτηριστικό και διατηρώντας τα άλλα σταθερά στις αρχικές τους τιμές. Με τον τρόπο αυτό τα ICE plots μας επιτρέπουν να έχουμε οπτική ανάλυση της κρυφής μεταβλητότητας που δεν είναι ορατή στην καμπύλη PDP και βοηθούν να εντοπίσουμε αν η επίδραση μιας μεταβλητής είναι ομοιόμορφη ή διαφέρει μεταξύ διαφορετικών παρατηρήσεων. Ωστόσο η ανάλυση των ICE βασίζεται μόνο στην οπτική παρατήρηση χωρίς να υπάρχει αντικειμενικό μέτρο ποσοτικοποίησης της διαφοροποίησης ενώ επιπλέον οι τιμές του χαρακτηριστικού μεταβάλλονται ανεξάρτητα από το αν οι αντίστοιχοι συνδυασμοί είναι ρεαλιστικοί για τη συγκεκριμένη παρατήρηση πράγμα που μπορεί να οδηγήσει σε παραπλανητικές ερμηνείες. (Gero Szepannaek & Karsten Lübke, 2023)

3.7.3 *Shapley Additive exPlanations (SHAP)*

Η μέθοδος ερμηνείας ML μοντέλων *Shapley Additive exPlanations (SHAP)* προσφέρει ένα πολύτιμο επίπεδο διαφάνειας στα μοντέλα επειδή καθιστά εφικτή την ερμηνεία τους τόσο σε παγκόσμιο όσο και σε τοπικό επίπεδο. Σε παγκόσμιο επίπεδο αποτυπώνει πώς συνεισφέρει κάθε μεταβλητή, θετικά ή αρνητικά, στο τελικό αποτέλεσμα δείχνοντας την ακριβή βαρύτητα της. Σε τοπικό επίπεδο δείχνει γιατί μια συγκεκριμένη παρατήρηση ταξινομήθηκε σε μια συγκεκριμένη κατηγορία και ποια ήταν η συμβολή των μεταβλητών στην εν λόγω απόφαση. Επιπλέον η μέθοδος SHAP βασίζεται στις Shapley values από την θεωρία παιγνίων και δίνει τη συμβολή χαρακτηριστικών που είναι θεωρητικά

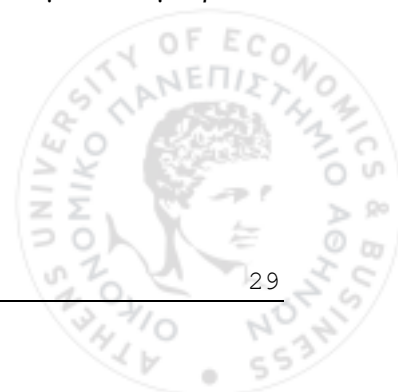
υποστηριζόμενη και αξιόπιστη έχοντας ιδιότητες τοπικής ακρίβειας και συνέπειας. Ακόμη η ανάλυση SHAP επιτρέπει τον εντοπισμό μη γραμμικών σχέσεων, διασποράς και δομικών αλλαγών μεταξύ των χαρακτηριστικών και της μεταβλητής-στόχου, πράγμα που καθιστά τη μέθοδο ιδιαίτερα χρήσιμη σε εφαρμογές πιστωτικού κινδύνου σε ρυθμιζόμενα περιβάλλοντα. (Ariza-Garzon et al.).

3.7.4 Local Interpretable Model-agnostic Explanations (LIME)

Από την άλλη μεριά η μεθοδολογία Local Interpretable Model-agnostic Explanations (LIME) εφαρμόζει υποκατάστατα μοντέλα στην τοπική περιοχή για να εξηγήσει με απλούς όρους τον τρόπο με τον οποίο το αρχικό μοντέλο έκανε μια συγκεκριμένη πρόβλεψη. Στόχος της είναι να εντοπίσει ποιοι ακριβείς παράγοντες οδήγησαν το μοντέλο σε μια συγκεκριμένη πρόβλεψη βασισόμενη στην υπόθεση ότι σε μικρότερη κλίμακα οι σχέσεις εμφανίζουν γραμμικότητα. Η μέθοδος αυτή στηρίζεται στη δημιουργία παραλλαγών των δειγμάτων δεδομένων και στην εφαρμογή τοπικών γραμμικών προσεγγίσεων, ώστε να αποκαλύψει πώς οι αλλαγές στα δεδομένα επηρεάζουν τα αποτελέσματα. Ακόμη το LIME αποτελεί μια model-agnostic τεχνική ερμηνείας που βοηθά να εξηγηθεί τι κάνει το μοντέλο και είναι ανεξάρτητο από την εσωτερική δομή του. Παρ' όλα αυτά, η περιγραφή που προσφέρει είναι μόνο σε υψηλό επίπεδο και δεν περιγράφει ολόκληρη τη συμπεριφορά του μοντέλου. (Ariza-Garzon et al.).

3.7.5 Surrogate models

Τα surrogate models (ή meta-models) είναι απλοποιημένα μοντέλα που χρησιμοποιούνται για την αναπαραγωγή της συμπεριφοράς ενός σύνθετου black box μοντέλου. Η λειτουργία τους είναι να κάνουν το σύνθετο μοντέλο διαφανές, ώστε να μπορούμε να κατανοήσουμε τη διαδικασία λήψης αποφάσεων. Αρχικά, στη λογική του global surrogate, παρατηρούμε το black box χρησιμοποιώντας διάφορα εργαλεία και μετά με βάση αυτή τη γνώση, κατασκευάζουμε ένα απλό κατανοητό μοντέλο που είναι ικανό να προσεγγίζει την αρχική συμπεριφορά. Σε τοπικό επίπεδο ένα παράδειγμα surrogate model είναι και το LIME. Τα global surrogate models είναι σχεδιασμένα για την προσέγγιση της συνολικής συμπεριφοράς του μοντέλου, ενώ τα local surrogate models επικεντρώνονται στην εξήγηση συγκεκριμένες προβλέψεις σε μια συγκεκριμένη περιοχή των δεδομένων. Από την μια μεριά surrogate model μπορεί να χρησιμοποιηθεί αντί του black box, έχει περίπου ίδιο επίπεδο ακρίβειας και είναι πολύ πιο διαφανές, κάτι που είναι κρίσιμο σε κλάδους με υψηλή ρύθμιση, όπως τα χρηματοπιστωτικά συστήματα. Από την άλλη μεριά η χρησιμότητα του εξαρτάται από το πόσο πιστά προσεγγίζει τη συμπεριφορά του αρχικού μοντέλου και ένα σε μεγάλο βαθμό απλοποιημένο μοντέλο μπορεί να οδηγήσει σε παραπλανητικές ερμηνείες.



3.7.6 *Counterfactual explanations*

Οι *counterfactual explanations* είναι επίσης μια τεχνική Explainable AI που στοχεύει να εξηγήσει πώς μικρές τροποποιήσεις στα δεδομένα εισόδου μπορούν να οδηγήσουν σε εναλλακτικά αποτελέσματα. Επομένως δεν δείχνουν μόνο πώς ένα μοντέλο κατέληξε σε μια συγκεκριμένη πρόβλεψη αλλά απαντούν στο ερώτημα τι θα μπορούσε να αλλάξει για να επιτευχθεί μια διαφορετική απόφαση. Σε έναν αιτούντα δανείου αυτή η τεχνική μπορεί να υποδεικνύει τις αλλαγές που απαιτούνται στα χαρακτηριστικά του αιτούντος για να μετατραπεί μια απόρριψη σε έγκριση, παρέχοντας με αυτό τον τρόπο εφαρμόσιμη ανατροφοδότηση (algorithmic recourse). Ένα αποτελεσματικό counterfactual πρέπει να διαθέτει ελάχιστες τροποποιήσεις στα χαρακτηριστικά έτσι ώστε η προτεινόμενη τροποποίηση να είναι ρεαλιστική και πρακτικά εφαρμόσιμη. Επιπλέον οι counterfactual explanations μπορούν να εφαρμοστούν και σε black-box μοντέλα στα οποία οι εσωτερικές λειτουργίες του μοντέλου δεν είναι γνωστές, και είναι νομικά αποδεκτές ώστε να είναι ελκυστικές για χρήση σε ρυθμιζόμενους τομείς όπως τα χρηματοπιστωτικά συστήματα. (Verma et al., 2020)

Από μια πρακτική προοπτική, οι Ariza-Garzon et al. αναφέρουν ορισμένες έννοιες που αφορούν την ερμηνευσιμότητα και την εξηγησιμότητα, και έχουν προτείνει διάφορες τεχνικές που μπορούν να χρησιμοποιηθούν στην εφαρμογή τους. Ταυτόχρονα, στο πιο εξειδικευμένο πλαίσιο της προσέγγισής αυτής, έχουν προταθεί τρία βασικά κριτήρια που πρέπει να ακολουθούνται κατά τη δημιουργία ερμηνεύσιμων μοντέλων: η γραμμικότητα, η μονοτονικότητα και η αλληλεπίδραση. Η γραμμικότητα όπως τη συναντάμε δείχνει μια σαφώς καθορισμένη ευθεία συσχέτιση μεταξύ μιας ανεξάρτητης μεταβλητής και της εξαρτημένης μεταβλητής. Η έννοια της μονοτονικότητας αποκαλύπτει ότι η αλληλεπίδραση μεταξύ μιας δεδομένης εισόδου και της επιδιωκόμενης εξαρτημένης μεταβλητής παραμένει αμετάβλητα ευθυγραμμισμένη σε όλο το φάσμα τιμών του σχετικού χαρακτηριστικού. Και τέλος, η έννοια της αλληλεπίδρασης αποτυπώνει το πώς το μοντέλο ενσωματώνει αυτόματα τις σχέσεις μεταξύ των χαρακτηριστικών προκειμένου να βελτιώσει τις προβλέψεις για τη εξαρτημένη μεταβλητή.

4 *Μεθοδολογία*

4.1 *Δεδομένα*

Τα δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση και την αξιολόγηση των μοντέλων ήταν το German Credit Dataset το οποίο προέρχεται από τη βάση δεδομένων UCI Machine Learning Repository και περιλαμβάνει πληροφορίες για αιτούντες δανείων. Περιέχει 1.000 παρατηρήσεις και 20 μεταβλητές που περιγράφουν δημογραφικά, οικονομικά και πιστωτικά χαρακτηριστικά. Οι μεταβλητές είναι αριθμητικές και κατηγορικές. Η εξαρτημένη μεταβλητή λαμβάνει δυαδική τιμή και δηλώνει εάν ο δανειολήπτης αξιολογείται ως «καλός» ή «κακός» πιστωτικός κίνδυνος.

4.2 *Προεπεξεργασία Δεδομένων*

Η προεπεξεργασία και η ανάλυση των δεδομένων έγιναν σε περιβάλλον Python. Ως πρώτο βήμα, για την διευκόλυνση της επεξερισμότητας, στο κωδικοποιημένο dataset προστέθηκαν τα ονόματα των μεταβλητών και των τιμών. Έπειτα ελέγχθηκε ποια είναι η φύση των μεταβλητών, ποιες είναι αριθμητικές και ποιες κατηγορικές. Στη συνέχεια έγινε ένας έλεγχος στο σύνολο των δεδομένων και διαπιστώθηκε ότι δεν υπάρχουν ελλιπείς τιμές. Για την καλύτερη κατανόηση των μεταβλητών έγιναν οπτικοποιήσεις των δεδομένων.

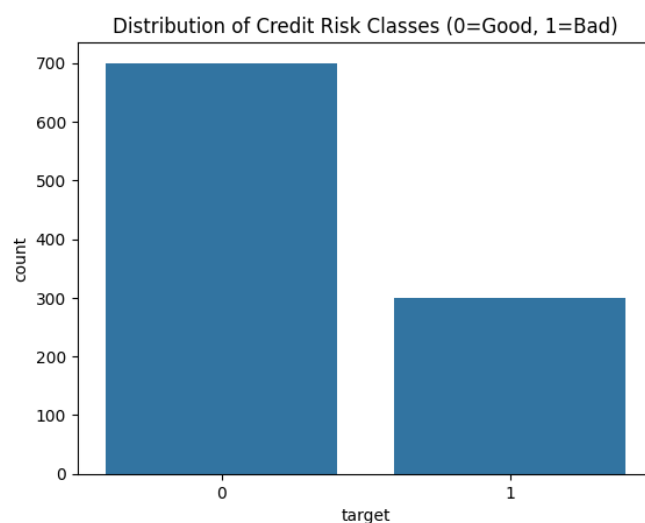
Τα δεδομένα που χρησιμοποιήθηκαν για την ανάλυση είναι τα ακόλουθα

A/A	Μεταβλητή	Περιγραφή
1	account_status	Κατάσταση του υπάρχοντος λογαριασμού όψεως
2	duration_months	Διάρκεια του δανείου σε μήνες.
3	credit_history	Ιστορικό προηγούμενων πιστώσεων του αιτούντα

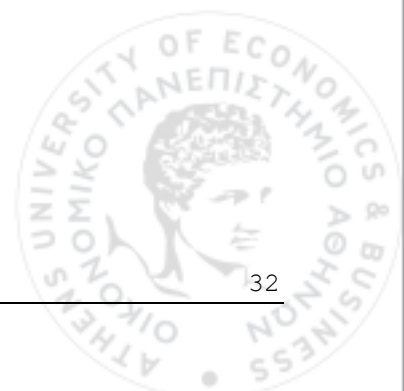
4	purpose	Σκοπός του αιτούμενου δανείου
5	credit_amount	Συνολικό ποσό του αιτούμενου δανείου.
6	savings_account	Κατάσταση αποταμιεύσεων ή ομολόγων
7	employment_since	Διάρκεια της τρέχουσας απασχόλησης του αιτούντα.
8	installment_rate_income	Ποσοστό της μηνιαίας δόσης σε σχέση με το διαθέσιμο εισόδημα του αιτούντα.
9	sex_only	Φύλο του αιτούντα
10	residence_since	Διάρκεια παραμονής του αιτούντα στην τρέχουσα κατοικία.
11	property	Τύπος και αξία περιουσιακών στοιχείων που κατέχει ο αιτών.
12	age_years	Ηλικία του αιτούντα σε έτη
13	other_installment_plans	Ύπαρξη άλλων ενεργών προγραμμάτων αποπληρωμής.
14	housing	Καθεστώς στέγασης
15	num_existing_credits	Αριθμός υφιστάμενων πιστώσεων του αιτούντα στο ίδιο πιστωτικό ίδρυμα.
16	job	Επαγγελματική κατάσταση και επίπεδο ειδίκευσης του αιτούντα.
17	num_dependents	Αριθμός ατόμων που εξαρτώνται οικονομικά από τον αιτούντα.
18	target	Εξαρτημένη μεταβλητή που εκφράζει την πιστοληπτική αξιολόγηση του αιτούντα («good» ή «bad» πιστωτικός κίνδυνος).

Εικόνα 2. Πίνακας με τις μεταβλητές του dataset

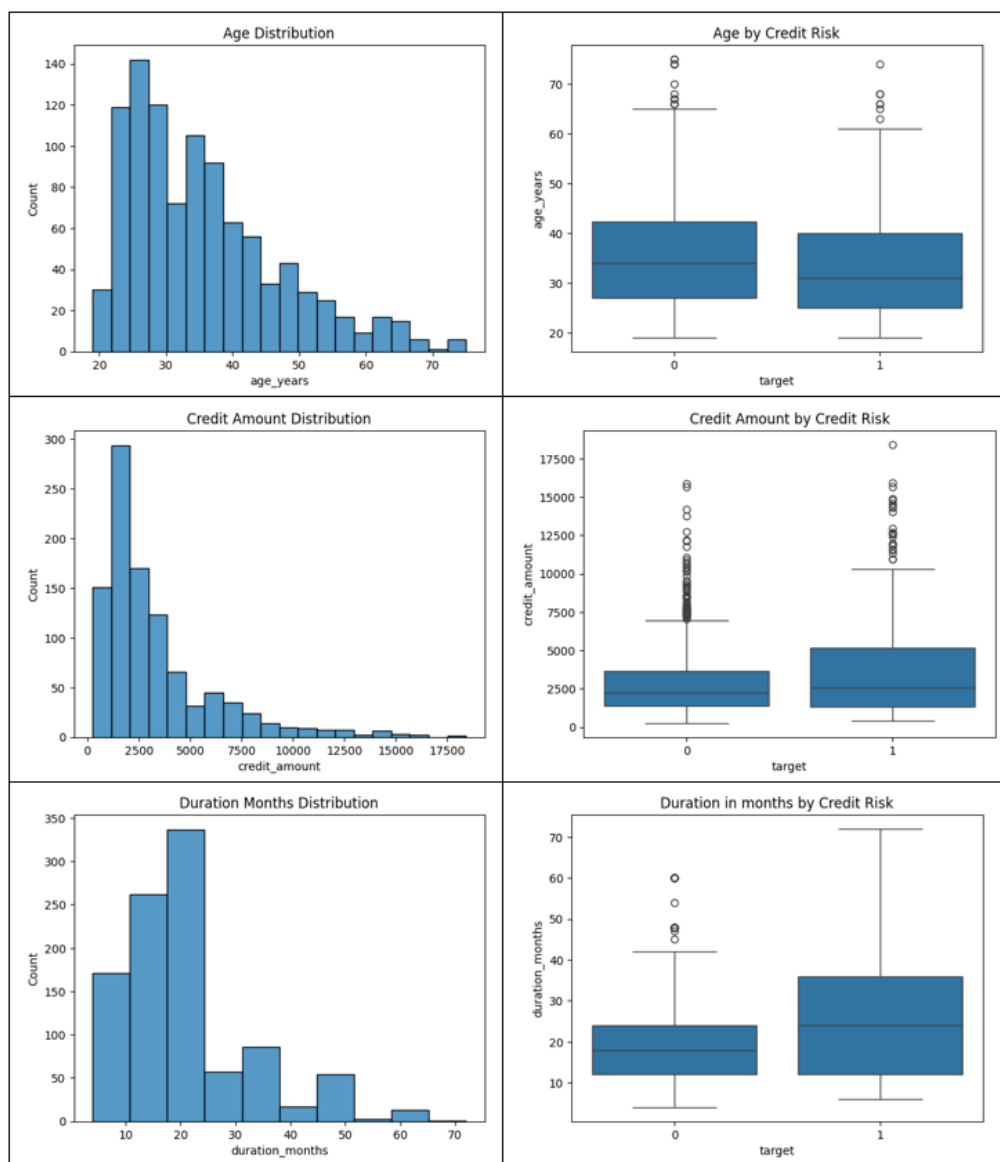
Το dataset παρουσιάζει ανισορροπία κλάσεων το οποίο φαίνεται στο παρακάτω διάγραμμα.



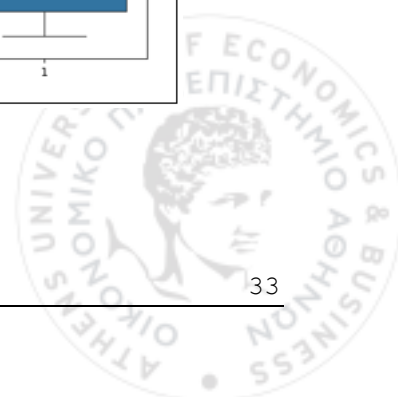
Εικόνα 4. Κατανομή Κλάσεων Πιστωτικού Κινδύνου



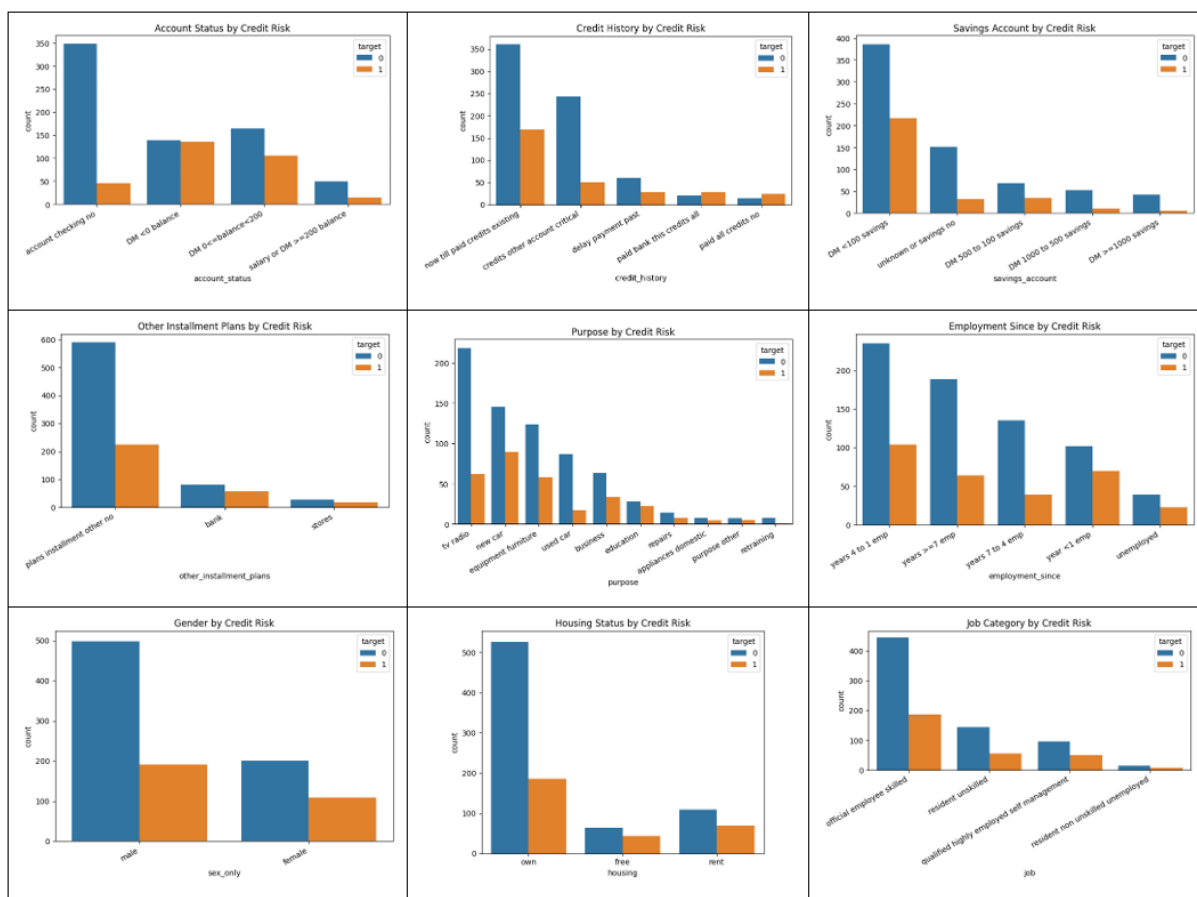
Οι αριθμητικές μεταβλητές είναι οι `duration_months`, `credit_amount`, `installment_rate_income`, `residence_since`, `age_years`, `num_existing_credits` και `num_dependents`. Σε αυτές τις μεταβλητές εφαρμόστηκε `standard scaling` στα μοντέλα `Logistic Regression`, `Support Vector Machines` και `Artificial Neural Networks` ενώ για τα μοντέλα `Decision Tree`, `Random Forest` και `XGBoost` δεν ήταν απαραίτητο διότι τα μοντέλα αυτά δεν επηρεάζονται από την κλίμακα των χαρακτηριστικών. Επιπλέον στις αριθμητικές μεταβλητές δεν έγινε `encoding`. Η μεταβλητή `credit_amount` παρουσίασε έντονη δεξιά ασυμμετρία. Για τον λόγο αυτό, εφαρμόστηκε λογαριθμικός μετασχηματισμός στη συγκεκριμένη μεταβλητή με στόχο τη μείωση της στρέβλωσης και τη βελτίωση της γραμμικότητας. Η συγκεκριμένη διαδικασία έγινε πάλι μόνο για τα μοντέλα `Logistic Regression`, `Support Vector Machines` και `Artificial Neural Networks`, τα οποία είναι ευαίσθητα στη διαφορά κλίμακας των χαρακτηριστικών.



Εικόνα 5. EDA στις αριθμητικές μεταβλητές



Οι κατηγορικές μεταβλητές, όπως οι `account_status`, `credit_history`, `purpose`, `savings_account`, `employment_since`, `sex_only`, `property`, `other_installment_plans`, `housing` και `job` κωδικοποιήθηκαν με τη μέθοδο One-Hot Encoding. Η επιλογή της συγκεκριμένης μεθόδου επιτρέπει τη μετατροπή των κατηγορικών δεδομένων σε δυαδικές μεταβλητές, χωρίς να εισάγεται τεχνητή διάταξη ή ιεράρχηση μεταξύ των κατηγοριών, γεγονός που είναι ιδιαίτερα σημαντικό για γραμμικά και μη γραμμικά μοντέλα. Ενώ στις κατηγορικές μεταβλητές κλίμακας τάξης όπως `savings_account` και `employment_since` εφαρμόστηκε η μέθοδος Ordinal Encoding και ορίστηκε σαφής σειρά των τιμών.



Εικόνα 6. EDA στις κατηγορικές μεταβλητές

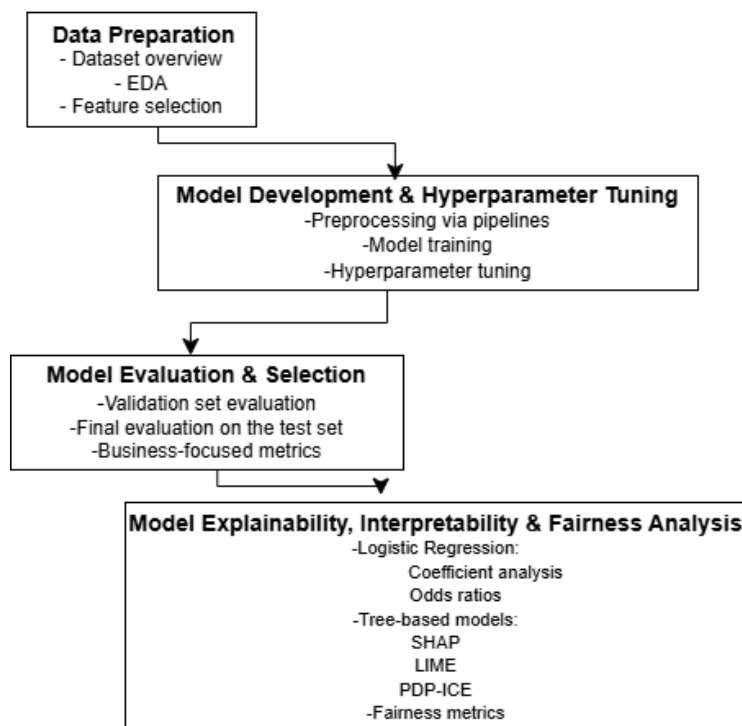
Οι μεταβλητές `sex_only` και `age_years` θεωρούνται ευαίσθητα χαρακτηριστικά και δεν αφαιρέθηκαν από το σύνολο των δεδομένων στην προεργασία. Χρησιμοποιούνται στο πλαίσιο της παρούσας μελέτης για την ανάλυση ερμηνευσιμότητας και τη διερεύνηση πιθανής μεροληψίας και ζητημάτων δικαιοσύνης στις αποφάσεις των μοντέλων μηχανικής μάθησης.

Στη συνέχεια, το σύνολο δεδομένων διαχωρίστηκε σε σύνολο εκπαίδευσης (training set) για την εκπαίδευση και τον συντονισμό των υπερπαραμέτρων μέσω διασταυρωμένης επικύρωσης (cross-validation), σύνολο επικύρωσης (validation set) για την συγκριτική αξιολόγηση απόδοσης και το σύνολο ελέγχου (test set) για την τελική αξιολόγηση απόδοσης των μοντέλων με αναλογία 60/20/20. Ο διαχωρισμός πραγματοποιήθηκε με τυχαία δειγματοληψία διατηρώντας την αναλογία των κλάσεων

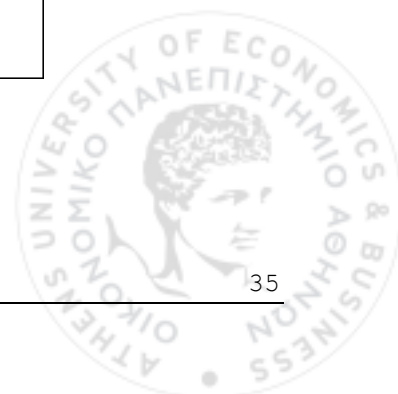
της εξαρτημένης μεταβλητής ώστε να αποφευχθεί πιθανή στρέβλωση στα αποτελέσματα σε όλα τα μοντέλα εκτός από το Logistic Regression στο οποίο εφαρμόστηκε oversampling (SMOTE).

Η ρύθμιση των υπερπαραμέτρων (hyperparameter tuning) πραγματοποιήθηκε μόνο στο training set και χρησιμοποιήθηκαν 5-folds Stratified Cross-Validation για όλα τα μοντέλα. Με τον τρόπο αυτό διασφαλίστηκε η δίκαιη σύγκριση τους και αποφεύχθηκε η διαρροή πληροφοριών (data leakage). Για το SVM εφαρμόστηκε η μέθοδος Grid Search για να ελεγχθούν μεθοδικά συνδυασμοί των βασικών υπερπαραμέτρων του μοντέλου. Στα υπόλοιπα μοντέλα εφαρμόστηκε η μέθοδος Random Search που επιτρέπει αποτελεσματικότερη εξερεύνηση του μεγάλου χώρου υπερπαραμέτρων. Όσον αφορά τη Logistic Regression, σε αυτό το μοντέλο δεν έγινε ρύθμιση υπερπαραμέτρων, καθώς χρησιμοποιήθηκε ως έχει. Οι καλύτερες υπερπαραμέτροι που ανακαλύφθηκαν εφαρμόστηκαν στη συνέχεια για να δοκιμαστούν τα μοντέλα στα validation και test set.

Όλα τα μοντέλα έτρεξαν σε ενιαίο υπολογιστικό πλαίσιο (pipeline) για να υπάρξει συνεπής εφαρμογή των ίδιων μετασχηματισμών και να διασφαλιστεί η συγκρισιμότητα των αποτελεσμάτων. Η τελική επιλογή και σύγκριση των μοντέλων βασίστηκε στα αποτελέσματα του validation set. Μετά την εκπαίδευση και ρύθμιση τα επιλεγμένα μοντέλα αξιολογήθηκαν στο test set το οποίο χρησιμοποιήθηκε αποκλειστικά για την τελική αποτίμηση της απόδοσης και της γενικευσιμότητας τους.



Εικόνα 7. Steps of model development and assessment



5 *Αξιολόγηση*

Σε αυτό το κεφάλαιο παρουσιάζονται τα αποτελέσματα αξιολόγησης των μοντέλων μηχανικής μάθησης που χρησιμοποιούνται στην ανάλυση πιστωτικού κινδύνου. Οι στόχοι της αξιολόγησης είναι η σύγκριση των προβλεπτικών ικανοτήτων των μοντέλων και η ανάλυση της διαφάνειας και της δικαιοσύνης τους στη λήψη αποφάσεων. Αρχικά, περιγράφονται τα κριτήρια αξιολόγησης και η διαδικασία αξιολόγησης, με τα πειραματικά αποτελέσματα να παρουσιάζονται και να συζητούνται στη συνέχεια.

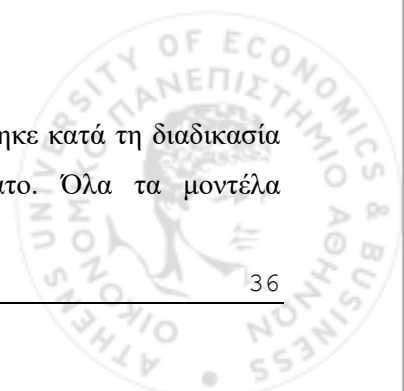
5.1 Παράμετροι αξιολόγησης

Τα μοντέλα αξιολογήθηκαν με βάση τις μετρικές που παρουσιάστηκαν αναλυτικά στο Κεφάλαιο 3 και είναι κατάλληλες για την ανάλυση πιστωτικού κινδύνου. Επειδή τα δεδομένα παρουσίασαν ανισορροπία κλάσεων και τέτοια προβλήματα έχουν διαφορετικό κόστος σφαλμάτων ταξινόμησης η προτεραιότητα δόθηκε στη μετρική ROC–AUC. Αυτό καταδεικνύει τη διακριτική δύναμη των μοντέλων και δεν εξαρτάται από μια συγκεκριμένη επιλογή ορίου απόφασης.

Ύστερα ακολούθησαν οι μετρικές recall και precision της κλάσης «bad» διότι είναι σημαντικό να εντοπίσουμε τους κακούς δανειολήπτες και να περιορίσουμε λανθασμένες κακές προβλέψεις. Το F1-score χρησιμοποιήθηκε ως συνδυασμός των δύο παραπάνω. Το accuracy εμφανίζεται συμπληρωματικά γιατί από μόνο του σε τέτοια προβλήματα με ανισορροπία κλάσεων δεν είναι επαρκές.

5.2 Σύστημα αξιολόγησης

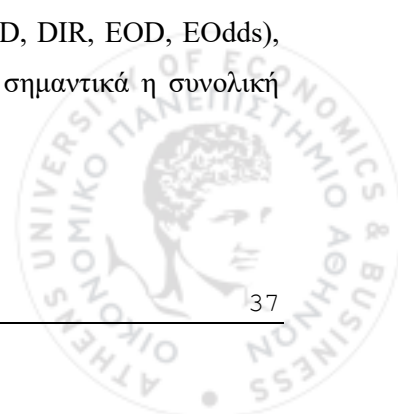
Η τελική αξιολόγηση των μοντέλων έγινε στο test set που δεν χρησιμοποιήθηκε κατά τη διαδικασία εκπαίδευσης ή τον πειραματισμό με υπερπαραμέτρους και ήταν αόρατο. Όλα τα μοντέλα



εκπαιδευτήκαν και αξιολογήθηκαν στο ίδιο σύνολο δεδομένων, ενώ χρησιμοποιήθηκε η ίδια διάσπαση σε σύνολα εκπαίδευσης, επικύρωσης και ελέγχου. Με αυτή τη προσέγγιση διασφαλίζεται ότι τα αποτελέσματα της αξιολόγησης δείχνουν την ικανότητα των μοντέλων να γενικεύονται σε περισσότερα νέα δεδομένα που δεν έχουν παρατηρηθεί.

Η αξιολόγηση βασίστηκε σε τρεις πρόσθετες πτυχές, την ακρίβεια πρόβλεψης, την επιχειρησιακή αποτελεσματικότητα και τη διαφάνεια των αποφάσεων. Η ακρίβεια πρόβλεψης αξιολογήθηκε με βάση τις μετρικές που παρουσιάστηκαν στο 5.1. Για την επιχειρησιακή αποτελεσματικότητα υπολογίστηκε το cumulative lift για τρία μοντέλα. Αυτά είναι τα Random Forest που είχε την καλύτερη συνολική απόδοση, το XGBoost που χρησιμοποιείται γενικά σε credit scoring και είναι αρκετά ερμηνεύσιμο με SHAP και Logistic Regression που είναι το baseline μοντέλο. Τέλος η διαφάνεια αποφάσεων έγινε με χρήση τεχνικών Explainable AI για να εξηγήσει τι προβλέπουν τα μοντέλα. Οι μέθοδοι SHAP και LIME χρησιμοποιήθηκαν στα μοντέλα Random Forest και XGBoost, τα οποία θεωρούνται black-box. Επιπλέον δημιουργήθηκαν PDP και ICE plots για το μοντέλο με την καλύτερη απόδοση για συμπληρωματική ανάλυση της ερμηνευσιμότητας καθώς μας παρέχουν την εικόνα της μέσης επίδρασης κάθε μεταβλητής και αποκαλύπτουν την επίδραση σε επίπεδο μεμονωμένων παρατηρήσεων. Το μοντέλο Logistic Regression, που είναι από την φύση του ερμηνεύσιμο, αναλύθηκε χρησιμοποιώντας ανάλυση των συντελεστών του μοντέλου και των αντίστοιχων odds ratios. Μέσω αυτών των τεχνικών μπορεί κανείς να κατανοήσει τους παράγοντες που επηρεάζουν τις αποφάσεις του μοντέλου καθώς και την αξιολόγηση της διαφάνειας της πρόβλεψης.

Επιπλέον, πραγματοποιήθηκε ανάλυση δικαιοσύνης πριν και μετά την εκπαίδευση των μοντέλων, με στόχο την ανάλυση πιθανής μεροληψίας σε ευαίσθητες μεταβλητές όπως το φύλο και η ηλικία. Στόχος της ανάλυσης ήταν η ανίχνευση της ήδη υπάρχουσας μεροληψίας στα δεδομένα, ελέγχοντας ευαίσθητα χαρακτηριστικά, όπως το φύλο και η ηλικιακή ομάδα. Εάν υπάρχει μεροληψία από την αρχή στα δεδομένα μπορεί να μεταφερθεί και να ενισχυθεί από τα μοντέλα οπότε είναι απαραίτητο να κατανοήσουμε τα δεδομένα. Για να επιτευχθεί αυτό, χρησιμοποιήθηκαν μετρικές όπως η Statistical Parity Difference (SPD), ο Disparate Impact Ratio (DIR), η Equal Opportunity Difference (EOD) και η Equalized Odds Difference (EOdds), προκειμένου να αξιολογηθεί η ίση μεταχείριση διαφορετικών ομάδων στις προβλέψεις των μοντέλων. Αφού ολοκληρώθηκε η αρχική αξιολόγηση εφαρμόστηκαν τεχνικές αντιμετώπισης μεροληψίας (bias mitigation) με σκοπό να περιοριστούν οι διαφορές που παρατηρούνται μεταξύ ομάδων και να βελτιωθεί η δικαιοσύνη των προβλέψεων. Στη συνέχεια η επίδραση αυτών των παρεμβάσεων ελέγχθηκε ξανά με τις ίδιες μετρικές (SPD, DIR, EOD, EOdds), ώστε να φανεί αν η μεροληψία μειώθηκε ουσιαστικά χωρίς να επηρεαστεί σημαντικά η συνολική απόδοση των μοντέλων.



5.3 Αποτελέσματα

5.3.1 Fairness analysis στα δεδομένα

Pre-model Fairness Summary (Default = 1)

Sensitive Variable	Disadvantaged Group	Disadvantaged Default Rate	Reference Group	Reference Default Rate	SPD	DIR	p-value	Interpretation
Gender	Female	0.352	Male	0.277	0.075	1.27	0.0207	Statistically significant bias
Age Group	18-25	0.421	36-50	0.238	0.184	1.77	0.0002	Strong age-based bias

Εικόνα 8. Ανάλυση δικαιοσύνης στα δεδομένα για φύλο και ηλικιακή ομάδα

Τα αποτελέσματα στον πίνακα δείχνουν ότι υπάρχουν στατιστικά σημαντικές ανισότητες ως προς την ηλικία και το φύλο. Οι γυναίκες έχουν ποσοστό αθέτησης 0,352 σε σύγκριση με τους άνδρες που έχουν ποσοστό αθέτησης 0,277. Η τιμή DIR είναι 1,27 (λόγος ποσοστών αθέτησης disadvantaged/reference) και το SPD είναι 0,075, υποδεικνύοντας ότι οι γυναίκες (η προστατευόμενη ομάδα) βιώνουν πιο αρνητικά αποτελέσματα. Αν εξετάσουμε τις ηλικιακές ομάδες, το ποσοστό αθέτησης μεταξύ 18 και 25 ετών είναι περίπου 0,421 έναντι του ποσοστού αθέτησης 0,238 για άτομα ηλικίας μεταξύ 36 και 50 ετών. Το SPD είναι 0,184 και το DIR είναι 1,77. Επιπλέον οι τιμές p-values δείχνουν ότι αυτές οι διαφορές είναι στατιστικά σημαντικές και όχι τυχαίες. Τα αποτελέσματα δείχνουν ότι το σύνολο δεδομένων περιέχει εγγενείς ανισότητες.

5.3.2 Συγκριτική αξιολόγηση απόδοσης μοντέλων

Performance on Test Set

Model	ROC-AUC	Recall (Bad)	Precision (Bad)	F1-score	Accuracy
Random Forest	0.805	0.617	0.638	0.627	0.78
SVM (RBF)	0.803	0.783	0.49	0.603	0.69
Logistic Regression	0.798	0.75	0.542	0.629	0.735
XGBoost	0.791	0.7	0.519	0.596	0.715
ANN (MLP)	0.76	0.717	0.506	0.593	0.705
Decision Tree	0.722	0.433	0.553	0.486	0.725

Εικόνα 9. Τελική αξιολόγηση μοντέλων στο test set

Το Random Forest είχε την υψηλότερη τιμή ROC-AUC (0.805) μεταξύ όλων των μοντέλων. Είναι το πιο αποτελεσματικό στη διάκριση καλών και κακών δανειοληπτών στο test set. Επιπλέον δείχνει σχετικά ισορροπημένη συμπεριφορά ως προς τα Recall (Bad = 0.617) και Precision (Bad = 0.638).

Αυτό δείχνει ότι αναγνωρίζει ένα μεγάλο ποσοστό της αθέτησης δανείων και δεν θεωρεί πολλούς δανειολήπτες υψηλού κινδύνου. Το ίδιο δείχνει και το F1-score που είναι 0.627. Το Accuracy (0.78) είναι το υψηλότερο απ' όλα τα μοντέλα.

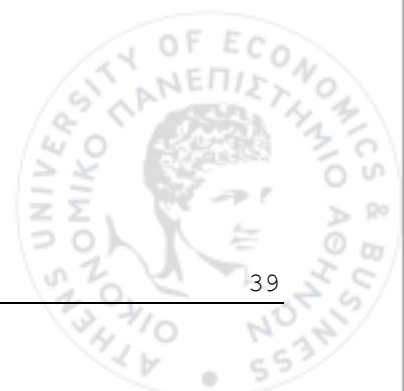
Δεύτερο στην σειρά είναι το SVM το οποίο επιδεικνύει υψηλό Recall (0.783) δείχνοντας ότι είναι καλό στην αναγνώριση κακών δανειοληπτών. Το χαμηλό όμως Precision (0,49) παρουσιάζει ότι το μοντέλο αναγνωρίζει λανθασμένα πολλούς καλούς δανειολήπτες ως κακούς. Αυτό προκαλεί χαμηλό Accuracy (0,69), υποδεικνύοντας ότι το μοντέλο αναγνωρίζει λανθασμένα τους καλούς δανειολήπτες ως κακούς για να εντοπίσει περισσότερους κακούς.

Η Logistic Regression εμφανίζει αρκετά ισορροπημένη απόδοση. Το Recall (Bad = 0.75) είναι υψηλό και το Precision (Bad = 0.542) είναι λίγο πιο χαμηλό, από όσους το μοντέλο χαρακτηρίζει ως Bad, μόνο το 54.2% είναι πράγματι. Το F1-score (0.629) είναι το υψηλότερο σε σχέση με τα άλλα μοντέλα και δείχνει ότι το μοντέλο επιτυγχάνει την καλύτερη δυνατή ισορροπία μεταξύ Recall και Precision. Το μοντέλο είναι ικανοποιητικό με Accuracy 0.735.

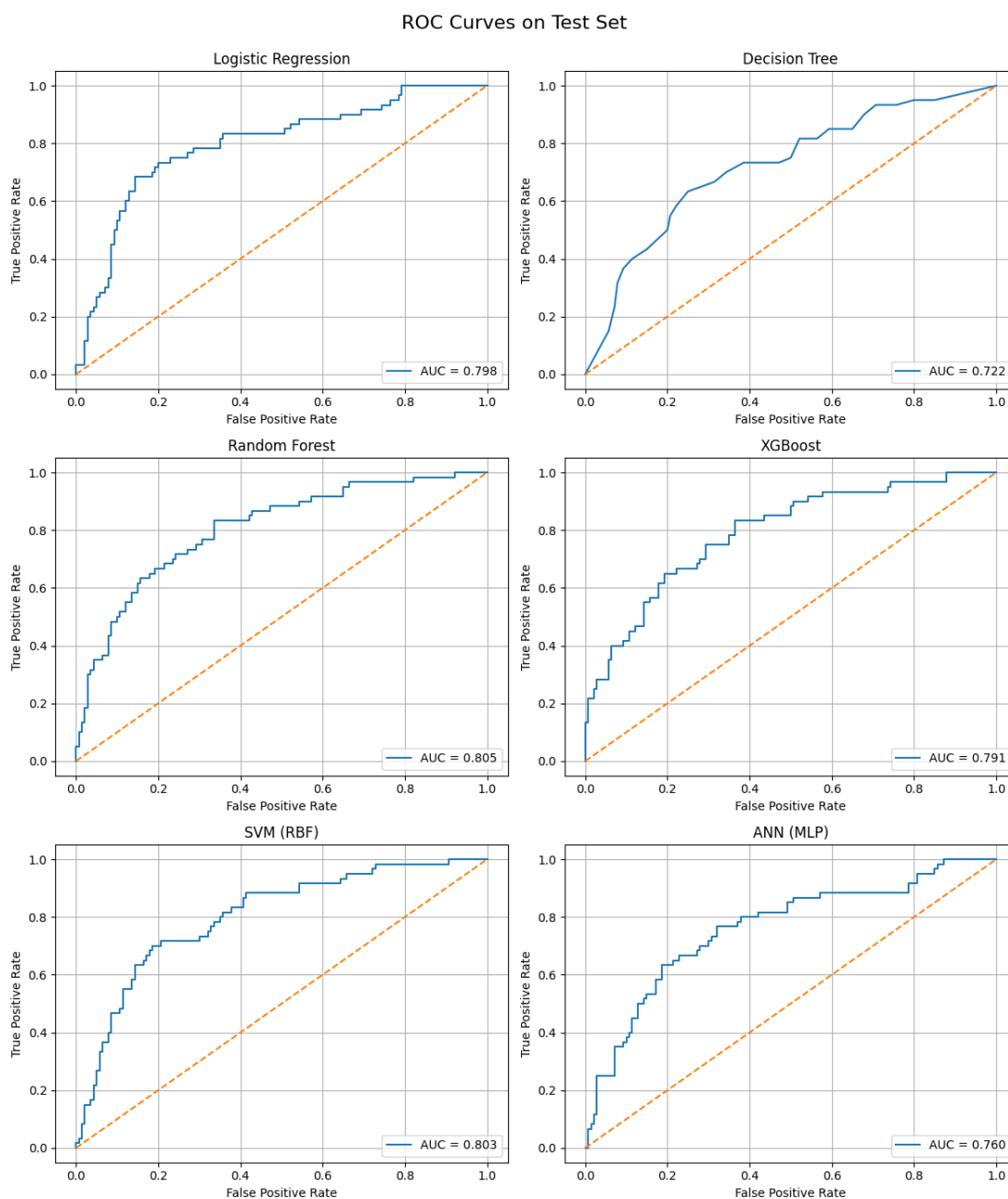
Το XGBoost έχει και αυτό μια γενική καλή απόδοση. Το ROC–AUC είναι 0.791, το Recall 0.7 και το Precision 0.51. Σε συνδυασμό με το F1-score που είναι 0.596 δείχνει ότι το μοντέλο διατηρεί μια μέση ισορροπία αναγνωρίζοντας ένα μεγάλο ποσοστό της αθέτησης δανείων και δεν θεωρεί πολλούς δανειολήπτες υψηλού κινδύνου. Έχει Accuracy 0,715 που σημαίνει ότι είναι γενικά σταθερό.

Το ANN εμφανίζει μια μέτρια απόδοση σε όλες τις μετρήσεις. Εντοπίζει επαρκή αριθμό κακών δανειοληπτών όπως αποδεικνύεται από το Recall (Bad = 0,717), αλλά το Precision (0,506) και το F1 (0,593) δείχνουν ότι δεν είναι τόσο ισχυρό όσο τα άλλα μοντέλα. Αυτή η εικόνα επιβεβαιώνεται από το Accuracy (0,705).

Το Decision Tree παρουσιάζει τη χαμηλότερη απόδοση. Το χαμηλό Recall (Bad = 0.433) δείχνει ότι το μοντέλο δεν εντοπίζει τους περισσότερους κακούς δανειολήπτες και το F1 (0,486) είναι το χειρότερο από όλα τα μοντέλα. Αν και το Accuracy (0.725) δεν είναι ιδιαίτερα χαμηλό, το μοντέλο δεν επιτυγχάνει στη διαχείριση του πιστωτικού κινδύνου επειδή οι περισσότερες από τις σωστές προβλέψεις του οφείλονται στο γεγονός ότι η πλειοψηφική κατηγορία έχει ταξινομηθεί σωστά.



5.3.3 Καμπύλες ROC

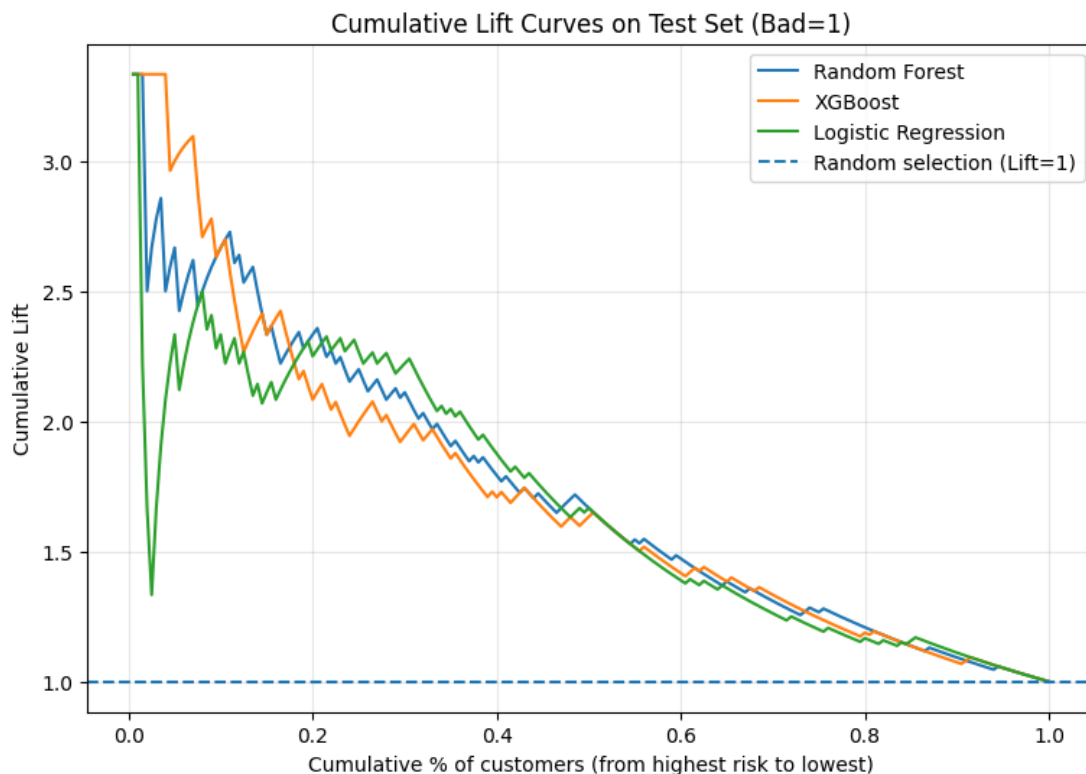


Εικόνα 10. Καμπύλες ROC στο test set

Τα διαγράμματα ROC απεικονίζουν τη σχέση μεταξύ των πραγματικών θετικών και ψευδώς θετικών σε διαφορετικά κατώφλια απόφασης. Τα αποτελέσματα δείχνουν ότι όλες οι καμπύλες βρίσκονται πολύ πάνω από τη διαγώνιο τυχαίας πρόβλεψης και επομένως όλα τα μοντέλα έχουν ισχυρή ικανότητα πρόβλεψης. Μια καμπύλη πιο κοντά στην επάνω αριστερή γωνία του διαγράμματος παρουσιάζεται από το Random Forest και το SVM (RBF), γεγονός που δείχνει ότι είναι σε θέση να επιτύχουν υψηλά ποσοστά ανίχνευσης κακών δανειοληπτών με χαμηλά ποσοστά ψευδώς θετικών αποτελεσμάτων. Το Decision Tree έχει μια καμπύλη που είναι πιο κοντά στη διαγώνιο και αυτό

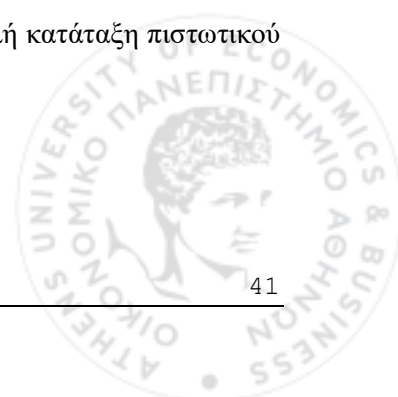
σημαίνει ότι έχει χαμηλότερη διακριτική ικανότητα. Τα υπόλοιπα μοντέλα παρουσιάζουν μια σταθερή συμπεριφορά.

5.3.4 Cumulative Lift Analysis



Εικόνα 11. Cumulative lift καμπύλες στο test set

Το διάγραμμα cumulative lift απεικονίζει πόσο καλά τα μοντέλα βρίσκουν τους κακούς δανειολήπτες όταν οι πελάτες κατατάσσονται από τον υψηλότερο στον χαμηλότερο προβλεπόμενο κίνδυνο. Τα τρία μοντέλα είναι καλύτερα από την τυχαία επιλογή στα πρώτα ποσοστά του πληθυσμού δείχνοντας ότι βρίσκουν πελάτες με υψηλό κίνδυνο νωρίς. Το Random Forest έχει ένα υψηλό και γενικά σταθερό cumulative lift, δείχνοντας ότι κατατάσσει καλά τους δανειολήπτες με τον υψηλότερο κίνδυνο στις πρώτες θέσεις. Το XGBoost έχει και αυτό ένα υψηλό cumulative lift αλλά το σχήμα της καμπύλης του μεταβάλλεται περισσότερο. Η Logistic Regression από την άλλη πλευρά παρουσιάζει αυξημένη αστάθεια στο αρχικό μέρος του πληθυσμού το οποίο δείχνει ανακριβή κατάταξη στο πολύ αρχικό τμήμα υψηλού κινδύνου. Στη συνέχεια, το μοντέλο υπερέχει έναντι των άλλων, στο ποσοστό μεταξύ 20% και 40% του πληθυσμού, έχοντας υψηλότερο cumulative lift. Η τάση αυτή δείχνει ότι το μοντέλο είναι αποδοτικότερο σε μεσαία επίπεδα κινδύνου και παρέχει μια καλή κατάταξη πιστωτικού κινδύνου για ένα σημαντικό ποσοστό του πληθυσμού.



Lift Table (Random Forest – Test Set)

decile	customers	bads	bad_rate	cumulative_bads	cumulative_bad_rate	lift
D1	20	16	0.8	16	0.267	2.667
D2	20	12	0.6	28	0.467	2.0
D3	20	10	0.5	38	0.633	1.667
D4	20	5	0.25	43	0.717	0.833
D5	20	7	0.35	50	0.833	1.167
D6	20	3	0.15	53	0.883	0.5
D7	20	2	0.1	55	0.917	0.333
D8	20	3	0.15	58	0.967	0.5
D9	20	1	0.05	59	0.983	0.167
D10	20	1	0.05	60	1.0	0.167

Εικόνα 12. Πίνακας lift για το μοντέλο Random Forest

5.3.5 Explainability

5.3.5.1 Baseline model: Logistic Regression

Logistic Regression Explainability (Top 10 features by $|\beta|$)

Feature	β (log-odds)	Odds Ratio (e^{β})	Effect
nom_account_status_no_checking_account	-1.1152	0.328	↓ decreases risk (Good=0)
nom_purpose_car_used	-0.9815	0.375	↓ decreases risk (Good=0)
nom_purpose_education	0.9636	2.621	↑ increases risk (Bad=1)
nom_account_status_balance_<0_DM	0.8612	2.366	↑ increases risk (Bad=1)
nom_credit_history_critical_account_other_credits	-0.7982	0.450	↓ decreases risk (Good=0)
nom_credit_history_no_credits_all_paid	0.7701	2.160	↑ increases risk (Bad=1)
nom_property_no_property_or_unknown	0.6324	1.882	↑ increases risk (Bad=1)
nom_purpose_repairs	0.5040	1.655	↑ increases risk (Bad=1)
nom_housing_rent	0.4836	1.622	↑ increases risk (Bad=1)
num_scale_rest_duration_months	0.4508	1.570	↑ increases risk (Bad=1)

Εικόνα 13. Ερμηνευσιμότητα Λογιστικής Παλινδρόμησης – Κορυφαία 10 Χαρακτηριστικά βάσει $|\beta|$

Το μοντέλο Logistic Regression είναι ερμηνεύσιμο και βασίζεται στους συντελεστές του και στα αντίστοιχα odds ratios. Αυτά επιτρέπουν να προσδιοριστεί η κατεύθυνση και η ένταση της επίδρασης κάθε χαρακτηριστικού στον πιστωτικό κίνδυνο. Όπως φαίνεται από τον πίνακα η πιθανότητα αθέτησης επηρεάζεται σε μεγάλο βαθμό από το σκοπό του δανείου, την κατάσταση λογαριασμού, το πιστωτικό ιστορικό, την ιδιοκτησία ακινήτου και άλλους δείκτες χρηματοοικονομικής σταθερότητας. Για παράδειγμα βλέπουμε ότι η μεταβλητή `nom_account_status_no_checking_account` έχει $\text{coef} = -1.115$ και $\text{odds ratio} = 0.328$. Αυτό σημαίνει ότι οι δανειολήπτες χωρίς checking account έχουν μικρότερη πιθανότητα να χαρακτηριστούν ως υψηλού κινδύνου. Ο συντελεστής β είναι αρνητικός οπότε μειώνει τα odds κατάταξης ενός δανειολήπτη σε «bad» και το odds ratio μικρότερο του 0.5 υποδηλώνει ισχυρή μείωση του σχετικού κινδύνου, με τα odds αθέτησης να είναι περίπου 67% χαμηλότερα σε σύγκριση με την κατηγορία αναφοράς.

5.3.5.2 Black box models

Οι τεχνικές Explainable AI εφαρμόστηκαν στα μοντέλα που είναι πιο σύνθετα στην ανάλυση, στο Random Forest, το μοντέλο που είχε την υψηλότερη ακρίβεια πρόβλεψης, και στο XGBoost, που θεωρείται περισσότερο black box καθώς είναι πιο περίπλοκο να εξηγηθούν και να επιβεβαιωθούν οι προβλέψεις του. Με την επιλογή των συγκεκριμένων μοντέλων αποσκοπείτε η αξιολόγηση του κατά πόσο οι μέθοδοι ερμηνευσιμότητας μπορούν να παρέχουν συνεπείς και αξιόπιστες εξηγήσεις σε διαφορετικές αρχιτεκτονικές μηχανικής μάθησης. Μέσω αυτής της ανάλυσης, είναι δυνατό να αναγνωρίσουμε τους παράγοντες που επηρεάζουν τις προβλέψεις των μοντέλων, καθώς και να συγκρίνουμε τη σχετική σημασία των χαρακτηριστικών, συμβάλλοντας στη διαφάνεια, την ερμηνευσιμότητα και την εμπιστοσύνη στη διαδικασία λήψης αποφάσεων.

5.3.5.2.1 LIME

XGBoost

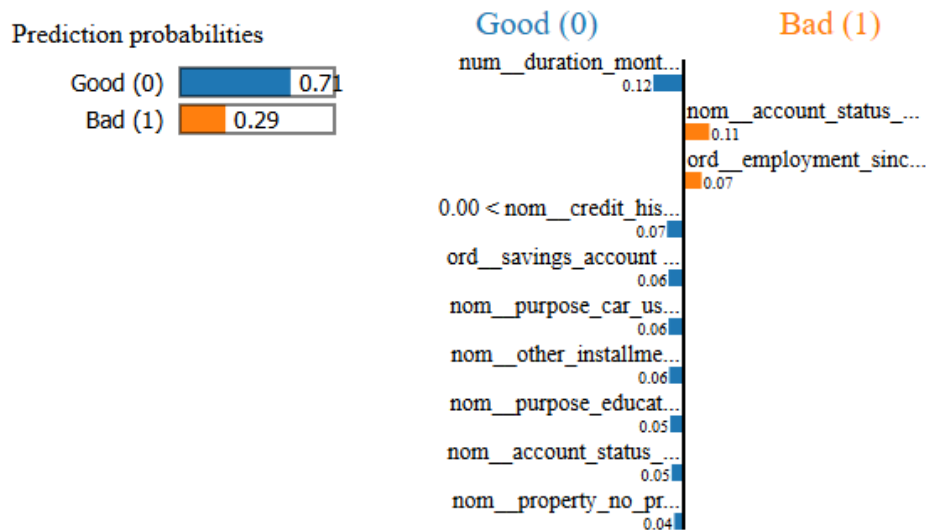
Feature	Value
num_duration_months	12.00
nom_account_status_no_checking_account	0.00
ord_employment_since	0.00
nom_credit_history_critical_account_other_credits	1.00
ord_savings_account	1.00
nom_purpose_car_used	1.00
nom_other_installment_plans_bank	0.00
nom_purpose_education	0.00
nom_account_status_balance_<0_DM	0.00
nom_property_no_property_or_unknown	0.00

LIME XGBoost

Εικόνα 14. Πίνακας LIME XGBoost model

Ο παραπάνω πίνακας παρουσιάζει τα χαρακτηριστικά και τις τιμές τους για μια συγκεκριμένη παρατήρηση τα οποία αναλύονται με τη μέθοδο LIME στο μοντέλο XGBoost. Κάθε γραμμή στον πίνακα είναι μια μεταβλητή εισόδου και δείχνει την τιμή της μεταβλητής για τον συγκεκριμένο δανειολήπτη, η οποία χρησιμοποιείται από το μοντέλο για να δημιουργήσει την πρόβλεψή του. Η τιμή της μεταβλητής num_duration_months είναι 12, πράγμα που σημαίνει ότι το δάνειο είναι μικρής διάρκειας και αυτό επηρεάζει σημαντικά την πρόβλεψη. Το χαρακτηριστικό employment_since είναι ordinal-encoded και η τιμή 0 σημαίνει ότι το άτομο βρίσκεται στην αρχική κατηγορία της ιεραρχίας που έχει καθιερωθεί και ήταν «unemployed». Όσες είναι κατηγορικές μεταβλητές και έγινε one-hot encoded και έχουν οριστεί σε 0 σημαίνει ότι ο δανειολήπτης δεν εμπίπτει σε αυτές τις κατηγορίες.

Για παράδειγμα, μια τιμή 0 στη μεταβλητή `nom_account_status_no_checking_account` υποδεικνύει ότι ο δανειολήπτης έχει τρεχούμενο λογαριασμό.



LIME Prediction Probability Breakdown XGBoost
 Εικόνα 15. LIME Ανάλυση πιθανοτήτων πρόβλεψης

Το παραπάνω διάγραμμα LIME δείχνει με ποιο τρόπο σε τοπικό επίπεδο ερμηνεύεται μια συγκεκριμένη πρόβλεψη του μοντέλου XGBoost δείχνοντας πως τα ξεχωριστά χαρακτηριστικά επηρεάζουν την τελική πιθανότητα ταξινόμησης. Σε αυτήν την περίπτωση, το μοντέλο δίνει στην κλάση Good (0) πιθανότητα 0,71 και στην κλάση Bad (1) πιθανότητα 0,29, επομένως ο δανειολήπτης ταξινομείται ως πιστοληπτικά αξιόπιστος.

Το διάγραμμα δείχνει πώς οι διάφορες συνεισφορές των χαρακτηριστικών επηρεάζουν την απόφαση προς κάθε κατεύθυνση. Ο πιστωτικός κίνδυνος μειώνεται από τα μπλε χαρακτηριστικά που μετακινούν την πρόβλεψη προς την κατηγορία Good (0). Τα πορτοκαλί χαρακτηριστικά προσθέτουν κίνδυνο και αυξάνουν την πιθανότητα να ταξινομηθεί κάποιος στη κλάση Bad (1). Η επιρροή ενός χαρακτηριστικού σε μια πρόβλεψη φαίνεται από το μήκος της ράβδου και την αριθμητική τιμή της και αυτά δείχνουν τη σχετική ένταση ενός χαρακτηριστικού στη πρόβλεψη.

Η πιο θετική επίδραση στην κλάση Good (0) επιτυγχάνεται από τη μεταβλητή `num_duration_months` με τιμή συνεισφοράς 0,12. Η μειωμένη διάρκεια του δανείου μειώνει την χρονική έκθεση και αποτελεί το κύριο προστατευτικό στοιχείο σε αυτήν την περίπτωση. Τα άλλα χαρακτηριστικά με μπλε χρώμα έχουν μικρότερες αλλά σταθερά θετικές επιπτώσεις που χρησιμεύουν ως βοηθητικές για την τελική απόφαση. Αντίθετα, οι πορτοκαλί ράβδοι απεικονίζουν τα χαρακτηριστικά που προσθέτουν στον εκτιμώμενο πιστωτικό κίνδυνο. Σε αυτήν την πρόβλεψη, ο μεγαλύτερος αρνητικός παράγοντας στην πρόβλεψη είναι η μεταβλητή `nom_account_status_no_checking_account` που έχει τιμή 0,11 και

ακολουθεί η μεταβλητή `ord_employment_since` με τιμή 0,07. Αυτά τα αρνητικά χαρακτηριστικά δεν έχουν ισχυρό αντίκτυπο, καθώς η τιμή τους είναι χαμηλή σε σχέση με τις θετικές συνεισφορές. Σύμφωνα με την τοπική εξήγηση που χρησιμοποιεί το LIME, η τελική απόφαση του XGBoost βασίζεται σε έναν συνδυασμό πολυάριθμων χαρακτηριστικών που έχουν διαφορετική κατεύθυνση και ένταση επίδρασης.

Random Forest



LIME Prediction Probability Breakdown Random Forest

Εικόνα 16. Πίνακας και Ανάλυση πιθανοτήτων πρόβλεψης LIME

Για το μοντέλο Random Forest χρησιμοποιήθηκε για το LIME η ίδια παρατήρηση με αυτήν που εξετάζεται στο XGBoost. Σε αυτήν την περίπτωση, το μοντέλο αποδίδει πιθανότητα 0,58 στην κλάση Good (0) και 0,42 στην κλάση Bad (1). Παρόλο που η τελική πρόβλεψη εξακολουθεί να υποστηρίζει την κλάση Good, η διαφορά μεταξύ των δύο πιθανοτήτων είναι μικρότερη, πράγμα που σημαίνει ότι το μοντέλο είναι λιγότερο βέβαιο από ό,τι στο XGBoost.

Η μεταβλητή `nom_account_status_no_checking_account` δημιουργεί τη μεγαλύτερη αρνητική επίδραση στην κατηγορία Bad (1) εμφανίζοντας το μεγαλύτερο μήκος πορτοκαλί ράβδου και τιμή

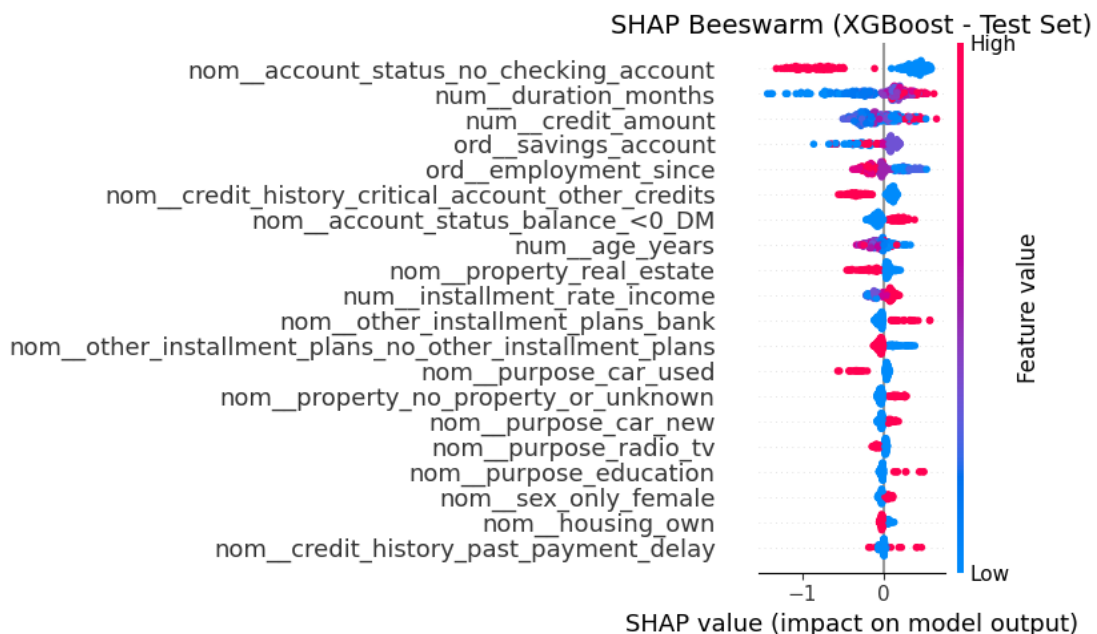
συνεισφοράς 0,15. Αυτό το αποτέλεσμα δείχνει ότι η κατάσταση του τραπεζικού λογαριασμού αποτελεί βασικό παράγοντα αύξησης του εκτιμώμενου πιστωτικού κινδύνου για αυτήν την παρατήρηση. Η μεταβλητή `ord_employment_since` επίσης αυξάνει τον κίνδυνο αλλά με μικρότερη ένταση από την κατάσταση του λογαριασμού με τιμή συνεισφοράς 0,05.

Η μεταβλητή `num_duration_months` είναι η μεταβλητή που επηρεάζει περισσότερο τη μείωση του πιστωτικού κινδύνου και δείχνει ότι η διάρκεια του δανείου είναι μια κρίσιμη μεταβλητή και στο μοντέλο Random Forest. Επιπλέον, μεταβλητές όπως `nom__account_status_balance_<0_DM` και `nom__credit_history_critical_account_other_credits` δίνουν μικρότερες αλλά σταθερές θετικές συνεισφορές στην κατηγορία Good (0) και βοηθούν στη λήψη της τελικής απόφασης.

5.3.5.2.2 SHAP

Το διάγραμμα SHAP beeswarm παρέχει τη συνολική εξήγηση του μοντέλου στο test set δείχνοντας τη σημασία των χαρακτηριστικών και τον τρόπο με τον οποίο επηρεάζουν το μοντέλο. Τα χαρακτηριστικά βρίσκονται στον κατακόρυφο άξονα, ταξινομημένα κατά φθίνουσα σειρά σπουδαιότητας και οι τιμές SHAP βρίσκονται στον οριζόντιο άξονα δείχνοντας την κατεύθυνση και το μέγεθος επίδρασης κάθε χαρακτηριστικού στην έξοδο του μοντέλου. Ο εκτιμώμενος πιστωτικός κίνδυνος αυξάνεται κατά τις θετικές τιμές SHAP και μειώνεται κατά τις αρνητικές τιμές.

XGBoost



Εικόνα 17. SHAP Beeswarm plot - XGBoost

Η πιο ισχυρή επίδραση στο μοντέλο XGBoost οφείλεται στο χαρακτηριστικό `nom__account_status_no_checking_account`. Σύμφωνα με την κατανομή των σημείων, οι υψηλές

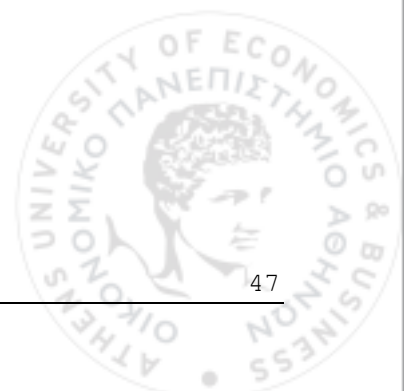
κόκκινες τιμές του χαρακτηριστικού σχετίζονται με αρνητικές τιμές SHAP, που σημαίνει ότι η απουσία τραπεζικού λογαριασμού μειώνει την εκτίμηση του πιστωτικού κινδύνου. Αντίθετα, τα μπλε σημεία στα δεξιά με θετικές τιμές SHAP δείχνουν την τάση αύξησης πιστωτικού κινδύνου.

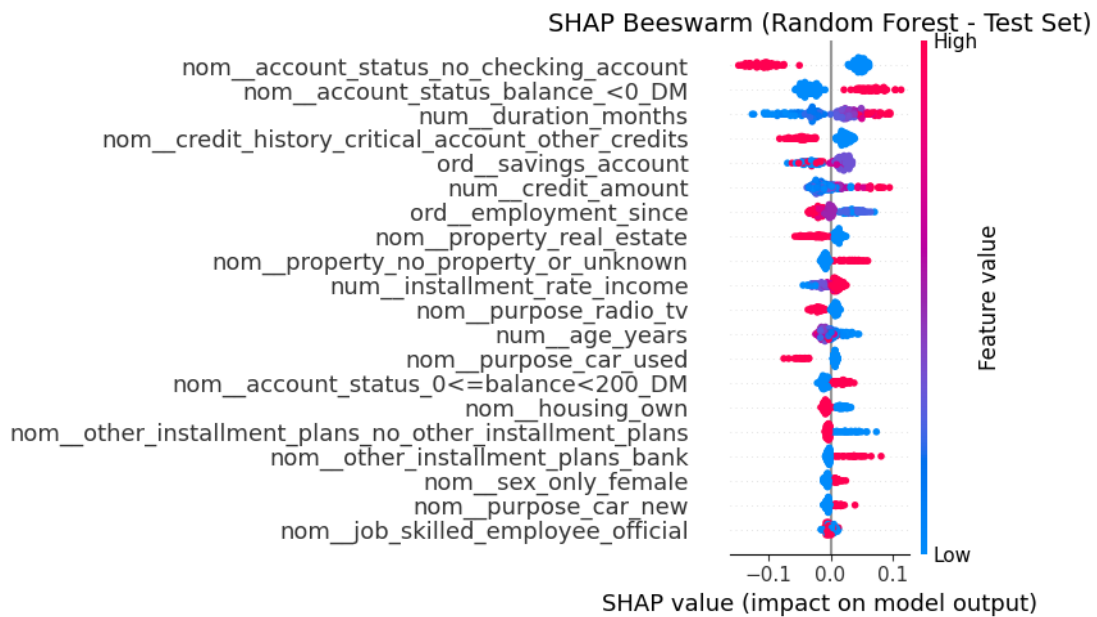
Η διάρκεια του δανείου (`num_duration_months`) είναι ο δεύτερος πιο σημαντικός παράγοντας. Το διάγραμμα δείχνει ότι οι μεγαλύτερες περιόδους δανείου (κόκκινα σημεία) αυξάνουν τον πιστωτικό κίνδυνο και οι μικρότερες περιόδους μειώνουν την τάση να ταξινομηθεί κάποιος ως κακός δανειολήπτης.

Ομοίως, το ποσό του δανείου (`num_credit_amount`) έχει ισχυρή επίδραση, δηλαδή οι υψηλές τιμές αυξάνουν τον πιστωτικό κίνδυνο. Επιπλέον η μεταβλητή `ord_savings_account` έχει χαμηλότερες μπλε τιμές στα αριστερά, αρνητικές τιμές SHAP, και υψηλότερες κόκκινες τιμές πάλι στα αριστερά αλλά πιο κοντά στη τιμή μηδέν. Αυτό δείχνει ότι η μεταβλητή λειτουργεί προστατευτικά ως προς την τάση ταξινόμησης στην κλάση Bad (1). Η διάρκεια απασχόλησης (`ord_employment_since`) δείχνει ότι όταν έχει χαμηλές τιμές, σύμφωνα με την σειρά που δηλώθηκε η χαμηλότερη είναι «unemployed», αυξάνει τον κίνδυνο και οι υψηλές τιμές, εργασία πάνω από 7 χρόνια, τον μειώνουν. Επιπλέον βλέπουμε ότι το κακό πιστωτικό ιστορικό (`nom_credit_history_critical_account_other_credits`) μειώνει την τάση ταξινόμησης στην κλάση Bad (1). Το αρνητικό υπόλοιπο λογαριασμού (`nom_account_status_balance_<0_DM`) αυξάνουν τον κίνδυνο αθέτησης.

Όσον αφορά τα δημογραφικά χαρακτηριστικά, η ηλικία (`num_age_years`) επηρεάζει ελαφρώς τις προβλέψεις του μοντέλου. Σύμφωνα με το διάγραμμα SHAP, οι τιμές νεαρής ηλικίας (μπλε σημεία) παράγουν θετικές τιμές SHAP που αυξάνουν την τάση ταξινόμησης στην κλάση Bad (1). Τα κόκκινα σημεία που αντιπροσωπεύουν υψηλότερες τιμές ηλικίας παράγουν αρνητικές τιμές SHAP που μειώνουν τον εκτιμώμενο πιστωτικό κίνδυνο. Το φύλο (`nom_sex_only_female`) παράγει ελάχιστη επίδραση επειδή οι τιμές SHAP παραμένουν γύρω στο μηδέν. Η κατανομή έντασης δείχνει ότι το XGBoost δεν χρησιμοποιεί το φύλο για να δημιουργήσει αλλαγές πρόβλεψης προς οποιαδήποτε κατεύθυνση.

Random Forest





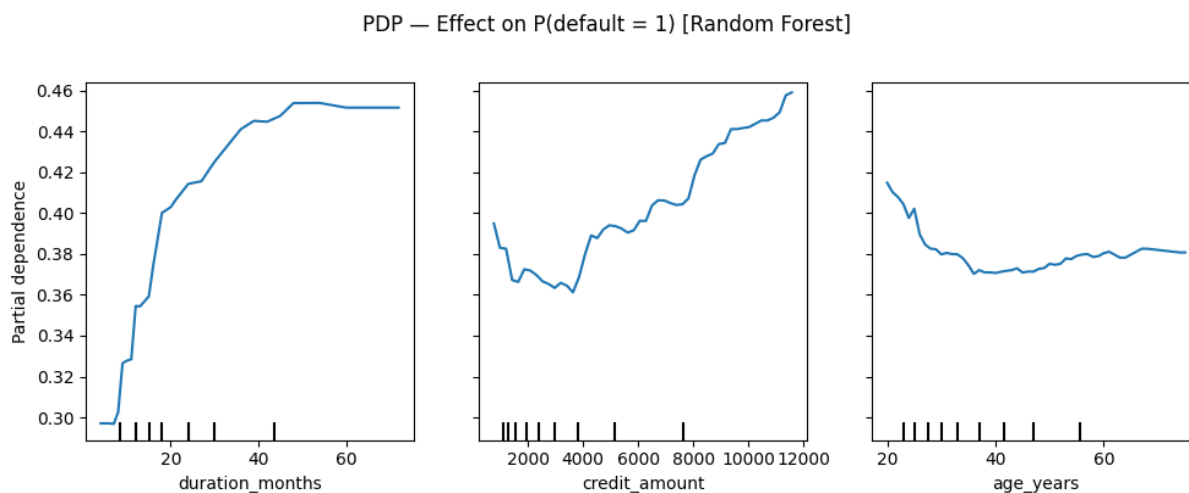
Εικόνα 18. SHAP Beeswarm plot – Random Forest

Και στο Random Forest το χαρακτηριστικό με τη μεγαλύτερη επιρροή είναι το `nom_account_status_no_checking_account`. Η κατανομή των σημείων δείχνει ότι οι υψηλές τιμές του χαρακτηριστικού προκαλούν αρνητικές τιμές SHAP. Η έλλειψη τρεχούμενου λογαριασμού οδηγεί σε μειωμένο προβλεπόμενο πιστωτικό κίνδυνο. Το επόμενο χαρακτηριστικό με μεγάλη σημασία είναι το `nom_account_status_balance_<0_DM`, το οποίο δείχνει τις υψηλότερες κόκκινες τιμές στα δεξιά, θετικό SHAP, που σημαίνει ότι τα αρνητικά υπόλοιπα λογαριασμών οδηγούν σε υψηλότερο κίνδυνο αθέτησης.

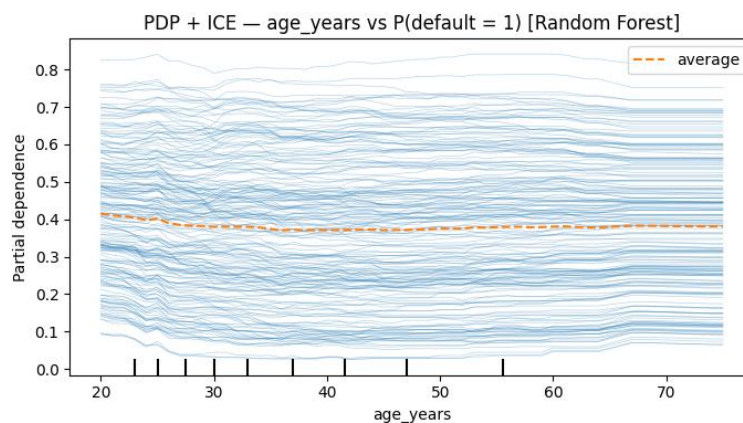
Η διάρκεια του δανείου (`num_duration_months`) είναι επίσης μια σημαντική μεταβλητή, καθώς οι περίοδοι με μεγαλύτερη διάρκεια έχουν θετικές τιμές SHAP, ενώ οι μικρότερες περιόδους μειώνουν τον πιστωτικό κίνδυνο. Η μεταβλητή `ord_savings_account` εμφανίζει περιορισμένη επίδραση στις προβλέψεις του μοντέλου γιατί τα σημεία της δεν ξεχωρίζουν σε υψηλές και χαμηλές αλλά είναι προς τα αριστερά, στις αρνητικές τιμές φτάνοντας περίπου $-0,1$. Επιπλέον, το `num_credit_amount` δείχνει ότι τα υψηλότερα ποσά δανείου αυξάνουν τον πιστωτικό κίνδυνο.

Όσον αφορά τα δημογραφικά χαρακτηριστικά `num_age_years` και `nom_sex_only_female` για αυτά ισχυριέται ότι και στο XGBoost. Η ηλικία επηρεάζει ελαφρώς τις προβλέψεις του μοντέλου και το φύλο παράγει ελάχιστη επίδραση επειδή οι τιμές SHAP παραμένουν γύρω στο μηδέν.

5.3.5.2.3 PDP και ICE plots



Εικόνα 19. PDPs για την πιθανότητα αθέτησης (P) - Random Forest



Εικόνα 20. PDP και ICE για τη μεταβλητή age_years στο Random Forest

Οι μεταβλητές duration_months και credit_amount που αναλύθηκαν χρησιμοποιώντας PDP επιλέχθηκαν με βάση τα αποτελέσματα της ανάλυσης SHAP και η μεταβλητή age_years για το ICE plot με βάση τη θεματική της διπλωματικής για δικαιοσύνη και μεροληψία προκειμένου να αναλυθεί η επίδρασή τους στις πιθανότητες αθέτησης.

Σύμφωνα με τα αποτελέσματα της καμπύλης PDP, η μεταβλητή duration_months, η οποία αναφέρεται στη διάρκεια του δανείου, είναι ένας σημαντικός παράγοντας που αυξάνει τις πιθανότητες αθέτησης του δανείου. Φαίνεται ότι όσο αυξάνεται η διάρκεια αυξάνεται και η πιθανότητα αθέτησης και παρουσιάζεται να σταθεροποιείται στους 50 και 60 μήνες. Στο credit_amount στα μικρά ποσά υπάρχει αρχικά μια μικρή μείωση του κινδύνου ενώ στη συνέχεια η πιθανότητα αθέτησης αυξάνεται ξανά σε υψηλές τιμές. Στη συνέχεια η ηλικία (age_years) επηρεάζει ελάχιστα και μη γραμμικά την πιθανότητα

αθέτησης σύμφωνα με την καμπύλη PDP, η οποία δείχνει αυξημένο κίνδυνο κατά τη νεαρή ηλικία και σταθεροποίηση μετά την ηλικία των 30 περίπου.

Όπως φαίνεται στο διαγράμματα ICE η επίδραση της ηλικίας στην πιθανότητα αθέτησης διαφοροποιείται σημαντικά μεταξύ μεμονωμένων δανειοληπτών. Η μέση καμπύλη παραμένει σταθερή με υψηλότερη πιθανότητα όταν η ηλικία είναι χαμηλή και σταθεροποιείται όταν η ηλικία υπερβαίνει μετά 30 έτη που σημαίνει ότι το μοντέλο δεν χρησιμοποιεί την ηλικία ως σημαντικό παράγοντα στην πρόβλεψη πιστωτικού κινδύνου. Οι μεμονωμένες καμπύλες δείχνουν επίσης υψηλό επίπεδο ετερογένειας στην επίδραση της ηλικίας, καθώς ορισμένοι δανειολήπτες επηρεάζονται από την ηλικία στην πρόβλεψη, ενώ άλλοι έχουν ελάχιστη επιρροή.

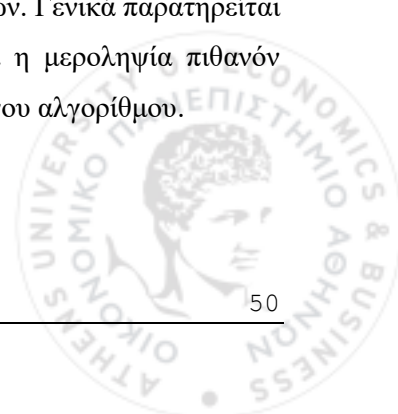
5.3.6 Fairness Analysis στο test set

Fairness Metrics (Pairwise) — Age (GOOD outcome)

Model	AUC	SPD	DIR	EOD	EOdds
Random Forest	0.805	-0.251	0.693	-0.228	0.128
Logistic Regression	0.798	-0.218	0.676	-0.204	0.109
XGBoost	0.791	-0.26	0.647	-0.268	0.137

Εικόνα 21. Fairness metrics ανά μοντέλο για την ηλικία

Τα αποτελέσματα της ανάλυσης fairness στο σύνολο ελέγχου για την ηλικιακή ομάδα δείχνουν μεροληψία εις βάρος της ομάδας 18–25 στα μοντέλα που εξετάστηκε. Οι τιμές Statistical Parity Difference (SPD) είναι αρνητικές (μεταξύ -0,22 και -0,26), που σημαίνει ότι οι νεότεροι δανειολήπτες λαμβάνουν λιγότερες θετικές προβλέψεις σε σύγκριση με την ομάδα αναφοράς 36-50 ετών. Οι τιμές Disparate Impact Ratio (DIR) κυμαίνονται μεταξύ 0,65 και 0,69, οι οποίες είναι χαμηλότερες από το όριο 0,8, το οποίο υποδηλώνει υψηλή μεροληψία έναντι της προστατευόμενης ομάδας. Επιπλέον οι αρνητικές τιμές της Equal Opportunity Difference (EOD) σημαίνουν ότι τα μοντέλα είναι λιγότερο ακριβή στην αναγνώριση πραγματικά αξιόπιστων πελατών στην ηλικιακή ομάδα 18-25 που δείχνει ανισότητα ευκαιριών. Η μετρική Equalized Odds Difference (EOdds) δείχνει μικρές αποκλίσεις στα σφάλματα του μοντέλου μεταξύ των ηλικιακών ομάδων χωρίς όμως σημαντικές διαφορές. Στη Logistic Regression παρατηρείται η πιο χαμηλή τιμή 0,109 που δείχνει ότι παρότι εξακολουθεί να υφίσταται μεροληψία το απλούστερο και γραμμικό μοντέλο παρουσιάζει την πιο ισορροπημένη συμπεριφορά ως προς τα σφάλματα ταξινόμησης μεταξύ των ηλικιακών ομάδων. Γενικά παρατηρείται ότι τα μοντέλα εμφανίζουν παρόμοια μοτίβα πράγμα που υποδηλώνει ότι η μεροληψία πιθανόν σχετίζεται με τη δομή των δεδομένων και όχι με την επιλογή ενός συγκεκριμένου αλγορίθμου.



Fairness Metrics (Pairwise) — Gender (GOOD outcome)

Model	AUC	SPD	DIR	EOD	EOdds
Random Forest	0.805	-0.102	0.863	-0.098	0.074
Logistic Regression	0.798	-0.111	0.821	-0.099	0.087
XGBoost	0.791	-0.101	0.838	-0.123	0.062

Εικόνα 22. Fairness metrics ανά μοντέλο για το φύλο

Τα αποτελέσματα της ανάλυσης fairness στο σύνολο ελέγχου για το φύλο δείχνουν ότι υπάρχει ήπια μεροληψία εις βάρος των γυναικών και στα τρία μοντέλα. Οι τιμές SPD είναι ελαφρώς μικρότερες από το μηδέν (μεταξύ -0,10 και -0,11) που σημαίνει ότι οι γυναίκες λαμβάνουν λίγο λιγότερα καλά αποτελέσματα από τους άνδρες. Οι τιμές DIR εμπίπτουν στο εύρος 0,82 έως 0,86, το οποίο είναι εντός του ενδεικτικού ορίου μεταξύ 0,8 και 1,25. Οι τιμές EOD είναι μικρές και αρνητικές που σημαίνει ότι τα μοντέλα δείχνουν χαμηλότερο True Positive Rate για τις γυναίκες. Η μετρική EOdds έχει χαμηλές τιμές, μεταξύ 0,06 και 0,09, που δείχνει ότι το μοντέλο κάνει παρόμοια λάθη κατά την ταξινόμηση ανδρών και γυναικών. Συνολικά, παρότι παρατηρούνται μικρές ενδείξεις μεροληψίας ως προς το φύλο η ένταση του φαινομένου είναι σαφώς πιο ήπια σε σύγκριση με τη μεροληψία που παρατηρήθηκε ως προς την ηλικία. Επομένως οι fairness μετρικές για το φύλο παραμένουν σε γενικά αποδεκτά επίπεδα και δεν εφαρμόστηκε bias mitigation για τη συγκεκριμένη προστατευόμενη μεταβλητή.

5.3.7 Post-processing Bias Mitigation για την ηλικία

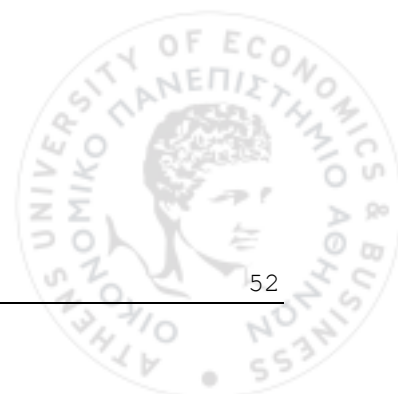
Post-processing Bias Mitigation — Age (cut=36), GOOD outcome

Model	SPD	DIR	EOD	EOdds	Accuracy	AUC
RF baseline	-0.184	0.773	-0.105	0.132	0.78	0.805
RF + CalibratedEqOdds	-0.175	0.784	-0.105	0.107	0.775	0.805
RF + RejectOption	-0.166	0.795	-0.076	0.132	0.79	0.805

Εικόνα 23. Εφαρμογή post processing τεχνικών για αντιμετώπιση μεροληψίας στην ηλικία

Έπειτα εφαρμόστηκαν τεχνικές post-processing για την αντιμετώπιση μεροληψίας στη μεταβλητή της ηλικίας η οποία εμφάνισε μεγαλύτερη μεροληψία. Για την εφαρμογή έγινε μετατροπή της μεταβλητής σε δυαδική μεταβλητή με όριο τα 36 έτη (<36 και ≥ 36) προκειμένου να είναι συμβατές με τις μεθόδους παρέμβασης. Εφαρμόστηκε το Random Forest το οποίο δείχνει να είναι μεροληπτικό προς την προστατευόμενη ομάδα όπως φαίνεται από το αρνητικό SPD (-0,184) και το DIR (0,773), τιμές που υποδεικνύουν ότι τα άτομα <36 λαμβάνουν λιγότερες θετικές προβλέψεις σε σύγκριση με την ομάδα αναφοράς. Η εφαρμογή της μεθόδου Calibrated Equalized Odds οδηγεί σε μικρή βελτίωση του SPD από -0,184 σε -0,175 και DIR από 0,773 σε 0,784 ενώ ταυτόχρονα μειώνει τα EOdds (0,132 σε 0,107) υποδεικνύοντας μια πιο ισορροπημένη συμπεριφορά όσον αφορά τα σφάλματα ταξινόμησης μεταξύ των δύο ομάδων αλλά χωρίς να αλλάζει το EOD (-0,105). Από την άλλη πλευρά η μέθοδος Reject Option Classification παρέχει την μεγαλύτερη βελτίωση όσον αφορά τις μετρικές fairness, το SPD γίνεται λιγότερο αρνητικό (-0,166), το DIR αυξάνεται περαιτέρω (0,795) και το EOD

βελτιώνεται αισθητά (-0,076), γεγονός που δείχνει μείωση της ανισότητας ευκαιριών για την προστατευόμενη ομάδα. Η απόδοση είναι η ίδια (0,805) παντού επομένως οι παρεμβάσεις δεν επηρεάζουν την ικανότητα του μοντέλου να διακρίνει, και το accuracy είναι λίγο διαφορετικό (0,775 στις CalibratedEqOdds και 0,79 για Reject Option). Τα αποτελέσματα γενικά δείχνουν ότι το post-processing μπορεί να μειώσει την μεροληψία στην ηλικία χωρίς ουσιαστική απώλεια απόδοσης.



6 *Επίλογος*

6.1 Σύνοψη και συμπεράσματα

Στη διπλωματική εργασία διερευνήθηκε η ανάλυση πιστωτικού κινδύνου μέσω αλγορίθμων μηχανικής μάθησης με έμφαση στη διαφάνεια, δικαιοσύνη και μεροληψία των αλγοριθμικών αποφάσεων. Σε αυτό το πλαίσιο υλοποιήθηκαν και συγκρίθηκαν σε περιβάλλον Python παραδοσιακά και πιο σύγχρονα μοντέλα χρησιμοποιώντας το German Credit Dataset ώστε να αξιολογηθεί η προβλεπτική τους απόδοση και η επεξηγηματικότητα των αποτελεσμάτων.

Στην τελική αξιολόγηση το μοντέλο Random Forest, ένα σύγχρονο μοντέλο μηχανικής μάθησης, πέτυχε υψηλότερη απόδοση στο σύνολο ελέγχου από το μοντέλο Logistic Regression. Στο συγκεκριμένο πρόβλημα το αποτέλεσμα δείχνει ότι το μοντέλο που βασίζονται σε δέντρα και ensemble μεθόδους είχε καλύτερη απόδοση με σύνθετες δομές δεδομένων. Η Logistic Regression είναι ένα απλό και κατανοητό μοντέλο που εξακολουθεί να έχει καλή απόδοση, αλλά είχε λίγο χαμηλότερη στο συγκεκριμένη περίπτωση. Τα σύγχρονα μοντέλα ML έχουν ξεπεράσει τα παραδοσιακά στατιστικά μοντέλα, κάτι που έχει ήδη αποδειχθεί από άλλες μελέτες.

Η ανάλυση εκτός από την ακρίβεια της πρόβλεψης επικεντρώθηκε στο ερώτημα κατά πόσον μια απόφαση μοντέλου θα μπορούσε να γίνει κατανοητή από έναν άνθρωπο. Αυτό είναι ένα κρίσιμο ζήτημα για εφαρμογές πιστωτικού κινδύνου υπό το ισχύον κανονιστικό πλαίσιο. Σύμφωνα με τα ευρήματα, η Logistic Regression είναι ένα απλό μοντέλο στην εξήγηση λόγω της γραμμικής του φύσης και τα σύνθετα μοντέλα μπορούν επίσης να γίνουν διαφανή χρησιμοποιώντας μεθόδους Explainable AI όπως το SHAP και το LIME. Με τις τεχνικές αυτές έγινε ανάλυση της συνολικής

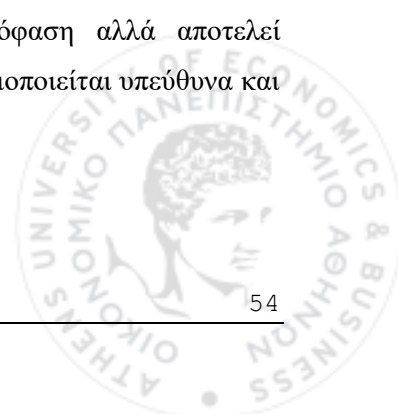
συμπεριφοράς του μοντέλου και των επιμέρους προβλέψεων, επιτρέποντας την μείωση του φαινομένου black box.

Η ανάλυση δικαιοσύνης που έγινε εμφάνισε ύπαρξη μεροληψίας στα δεδομένα κυρίως ως προς την ηλικία και πολύ λιγότερο ως προς το φύλο, για το οποίο οι μετρικές δικαιοσύνης ήταν στα αποδεκτά όρια. Για αυτό το λόγο εφαρμόστηκε τεχνική μετριασμού μεροληψίας μόνο ως προς την ηλικία στο στάδιο post-processing για τη μείωση των ανισοτήτων που παρατηρήθηκαν στις τελικές προβλέψεις. Μετά την εφαρμογή οι μετρικές δικαιοσύνης βελτιώθηκαν χωρίς αντίκτυπο στην προβλεπτική απόδοση και επιτεύχθηκε μια πιο δίκαιη συμπεριφορά. Η μεροληψία μειώθηκε αλλά δεν εξαλείφθηκε που μπορεί να σημαίνει ότι η ανισότητα οφείλεται σε έμμεσες συσχετίσεις στα δεδομένα και σε ιστορικά μοτίβα στα χαρακτηριστικά πράγμα που αναδεικνύει την ανάγκη συνεχούς ελέγχου της δικαιοσύνης και διάκρισης.

Η μελέτη σχετικά με τη σημασία των χαρακτηριστικών έδειξε ότι τα μοντέλα λαμβάνουν αποφάσεις με βάση οικονομικές και πιστωτικές μεταβλητές, συμπεριλαμβανομένης της κατάστασης του λογαριασμού και του πιστωτικού ιστορικού, του ποσού του δανείου και της διάρκειας του δανείου. Οι δημογραφικές μεταβλητές, συμπεριλαμβανομένης της ηλικίας και του φύλου είχαν πολύ μικρή επίδραση στις προβλέψεις των μοντέλων και, ως εκ τούτου, δεν αποτελούν κυρίαρχους παράγοντες λήψης αποφάσεων. Τα αποτελέσματα είναι σημαντικά για την πτυχή της δικαιοσύνης και της μη διάκρισης και είναι συνεπή με άλλες μελέτες που δείχνουν ότι η πολυπλοκότητα του μοντέλου δεν καθορίζει την προκατάληψη, αλλά καθορίζεται από τα δεδομένα και τα μέσα αξιολόγησης. (Bono et al., 2021).

Τα αποτελέσματα της διπλωματικής δείχνουν ότι σε πρακτικό επίπεδο τα μοντέλα υψηλής απόδοσης μπορούν να συνδυαστούν με τεχνικές Explainable AI για την επίτευξη ισορροπίας μεταξύ ακρίβειας και διαφάνειας. Αυτή η προσέγγιση είναι σύμφωνη με το ευρωπαϊκό κανονιστικό πλαίσιο που αναπτύσσεται συνεχώς σχετικά με τη χρήση αλγοριθμικών συστημάτων στον χρηματοπιστωτικό τομέα όπου απαιτούνται βασικές απαιτήσεις επεξηγησιμότητα και διαφάνειας, καθώς και λογοδοσίας για εφαρμογές όπως η αξιολόγηση πιστοληπτικής ικανότητας. (Langenbacher, 2022).

Τα σύγχρονα μοντέλα μηχανικής μάθησης μπορούν να θεωρηθούν αξιόπιστα για την αξιολόγηση του πιστωτικού κινδύνου υπό την προϋπόθεση ότι ερμηνεύονται και ελέγχονται σωστά. Η επεξηγηματικότητα στην αξιολόγηση δεν είναι απλώς μια τεχνική απόφαση αλλά αποτελεί απαραίτητη προϋπόθεση για να διασφαλιστεί ότι η τεχνητή νοημοσύνη χρησιμοποιείται υπεύθυνα και δίκαια στον χρηματοπιστωτικό τομέα.



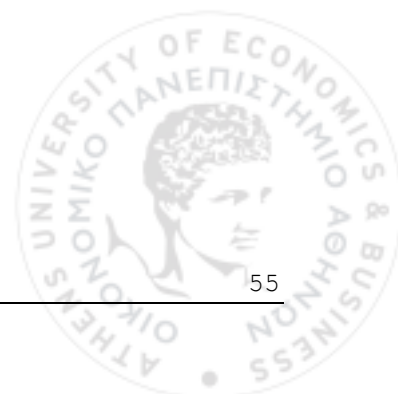
6.2 Μελλοντικές επεκτάσεις

Παρότι η παρούσα διπλωματική εργασία κατέληξε σε σημαντικά ευρήματα, υπάρχουν ορισμένοι περιορισμοί που θα πρέπει να ληφθούν υπόψη. Η ανάλυση βασίστηκε αποκλειστικά στο German Credit Dataset το οποίο αποτελεί ένα σχετικά μικρό και παλαιότερο σύνολο δεδομένων. Αυτό περιορίζει τη δυνατότητα γενίκευσης των αποτελεσμάτων σε σύγχρονα τραπεζικά περιβάλλοντα. Επιπλέον η αξιολόγηση της δικαιοσύνης πραγματοποιήθηκε με συγκεκριμένους στατιστικούς δείκτες fairness, οι οποίοι όπως έχει αναφερθεί στη βιβλιογραφία δεν μπορούν να αποτυπώσουν όλες τις πιθανές μορφές αλγοριθμικής μεροληψίας.

Οι περιορισμοί αυτοί αναδεικνύουν ταυτόχρονα και τις βασικές κατευθύνσεις για μελλοντική έρευνα. Αν και η παρούσα μελέτη παρείχε μια εκτενή ανάλυση της αξιολόγησης πιστωτικού κινδύνου με αλγορίθμους μηχανικής μάθησης και τεχνικές Explainable AI, υπάρχουν αρκετά σημεία στα οποία η έρευνα θα μπορούσε να επεκταθεί.

Μια μελλοντική κατεύθυνση θα μπορούσε να περιλαμβάνει την εφαρμογή τεχνικών αντιμετώπισης μεροληψίας (bias mitigation) όχι μόνο στο στάδιο post-processing, το οποίο επηρεάζει το τελικό αποτέλεσμα των μοντέλων αλλά και στο στάδιο pre-processing και in-processing. Η συγκριτική ανάλυση διαφόρων μεθόδων θα παρείχε μια πιο ολοκληρωμένη εικόνα του βαθμού μεροληψίας που μπορεί να μειωθεί και την ανάλυση των πιθανών αντισταθμίσεων μεταξύ δικαιοσύνης, προβλεπτικής απόδοσης και ερμηνευσιμότητας. Η επόμενη ερευνητική κατεύθυνση θα μπορούσε να είναι να κάνει την ανάλυση της ερμηνευσιμότητας να υπερβαίνει τον εντοπισμό χαρακτηριστικών που επηρεάζουν τις προβλέψεις και να δημιουργήσει πιο πρακτικές ενεργές εξηγήσεις ελέγχοντας πώς ρεαλιστικές αλλαγές στα χαρακτηριστικά ενός συγκεκριμένου δανειολήπτη μπορούν να οδηγήσουν σε αντιστροφή της πρόβλεψης, π.χ. από υψηλό κίνδυνο σε χαμηλό κίνδυνο και ως εκ τούτου παρέχοντας πληροφορίες όχι μόνο για το γιατί ελήφθη μια απόφαση αλλά και για το τι μπορεί να αλλάξει για να βελτιωθεί το τελικό αποτέλεσμα.

Τέλος, περαιτέρω έρευνα θα μπορούσε να επικεντρωθεί στη μεταφορά των μοντέλων σε ένα πραγματικό επιχειρηματικό περιβάλλον, εξετάζοντας και τη συμμόρφωση τους με τις απαιτήσεις των κανονιστικών πλαισίων. Ο συνδυασμός της τεχνικής ανάλυσης με τις νομικές ανησυχίες θα ενίσχυε περαιτέρω τη χρησιμότητα των αποτελεσμάτων.



7

Βιβλιογραφία

- André Aoun Montevechi, de Carvalho Miranda, R., André Luiz Medeiros, & Arnaldo Barra Montevechi, J. (2024). Advancing credit risk modelling with Machine Learning: A comprehensive review of the state-of-the-art. *Engineering Applications of Artificial Intelligence*, 137, 109082–109082. <https://doi.org/10.1016/j.engappai.2024.109082>
- Ariza-Garzon, M. J., Arroyo, J., Caparrini, A., & Segovia-Vargas, M.-J. (2020). Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending. *IEEE Access*, 8, 64873–64890. <https://doi.org/10.1109/access.2020.2984412>
- Aza, A. (2024). Scores as Decisions? Article 22 GDPR and the Judgment of the CJEU in *SCHUFA Holding (Scoring)* in the Labour Context. *Industrial Law Journal*, 53(4), 840–858. <https://doi.org/10.1093/indlaw/dwae035>
- Barra, C., Papaccio, A., & Ruggiero, N. (2022). Basel accords and banking inefficiency: Evidence from the Italian local market. *International Journal of Finance & Economics*. <https://doi.org/10.1002/ijfe.2637>
- Bono, T., Croxson, K., & Giles, A. (2021). Algorithmic fairness in credit scoring. *Oxford Review of Economic Policy*, 37(3), 585–617. <https://doi.org/10.1093/oxrep/grab020>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
- Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable Machine Learning in Credit Risk Management. *Computational Economics*, 57.
- Chang, V., Xu, Q. A., Akinloye, S. H., Benson, V., & Hall, K. (2024). Prediction of bank credit worthiness through credit risk analysis: an explainable machine learning study. *Annals of Operation Research/Annals of Operations Research*. <https://doi.org/10.1007/s10479-024-06134-x>
- Chen, T., & Guestrin, C. (2016). XGBoost: a Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 1(1), 785–794. <https://doi.org/10.1145/2939672.2939785>

- Chen, Y., Giudici, P., Liu, K., & Emanuela Raffinetti. (2024). Measuring fairness in credit ratings. *Expert Systems with Applications*, 125184–125184. <https://doi.org/10.1016/j.eswa.2024.125184>
- Dastile, X., & Celik, T. (2024). Counterfactual Explanations With Multiple Properties in Credit Scoring. *IEEE Access*, 12, 110713–110728. <https://doi.org/10.1109/access.2024.3441037>
- Dastile, X., Celik, T., & Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91, 106263. <https://doi.org/10.1016/j.asoc.2020.106263>
- European Commission. (2025, August 1). *AI Act*. European Commission. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Gero Szepannaek, & Karsten Lübke. (2023). How much do we see? On the explainability of partial dependence plots for credit risk scoring. *Argumenta Oeconomica*, 2023(2), 137–150. <https://doi.org/10.15611/aoe.2023.1.07>
- Goodness, S., Shan, A., Oladele, S., & Stark, B. (2025, March 25). *AI and Credit Scoring: Assessing the Fairness and Transparency of Machine Learning Models in Lending Decisions*. ResearchGate; unknown. https://www.researchgate.net/publication/390172601_AI_and_Credit_Scoring_Assessing_the_Fairness_and_Transparency_of_Machine_Learning_Models_in_Lending_Decisions
- Hardt, M., Price, E., & Srebro, N. (2016, October 7). *Equality of Opportunity in Supervised Learning*. ArXiv.org. <https://doi.org/10.48550/arXiv.1610.02413>
- Hurlin, C., Perignon, C., & Saurin, S. (2021). The Fairness of Credit Scoring Models. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3785882>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An Introduction to Statistical Learning*. Springer.
- Johnson, K. D., Foster, D. P., & Stine, R. A. (2016). *Impartial Predictive Modeling and the Use of Proxy Variables*. ArXiv.org. <https://arxiv.org/abs/1608.00528>
- K., N., Cherukuri, R. K., Bollapalli, N. L., Gurrum, R., Kandula, B. M., & Addagarla, L. S. (2025). Fairness-Aware Machine Learning for Credit Scoring: An Empirical Study Using Mitigation Techniques. *International Journal of Recent Advances in Engineering and Technology*, 14(01, 2025).
- Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling* (pp. 333–338, 343–350). Springer New York.
- Liang, Z., Rewolinski, Z. T., Agarwal, A., Tang, T. M., & Yu, B. (2025). Local MDI+: Local Feature Importances for Tree-Based Models. *ArXiv (Cornell University)*. <https://doi.org/10.48550/arxiv.2506.08928>

- Lui, A. T., Lamb, G., & Durodola, L. (2025). A right to explanation for algorithmic credit decisions in the UK. *Law, Innovation and Technology*, 1–29. <https://doi.org/10.1080/17579961.2025.2469352>
- Mariscal, C., Yustiawan, Y., Rochim, F. C., & Tanuar, E. (2024). Implementing and analyzing fairness in banking credit scoring. *Procedia Computer Science*, 234, 1492–1499. <https://doi.org/10.1016/j.procs.2024.03.150>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021a). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 1–35. <https://doi.org/10.1145/3457607>
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021b). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 1–35. <https://doi.org/10.1145/3457607>
- Moldovan, D. (2023). Algorithmic Decision Making Methods for Fair Credit Scoring. *IEEE Access*, 11, 59729–59743. <https://doi.org/10.1109/access.2023.3286018>
- Moscato, V., Picariello, A., & Sperlí, G. (2021). A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*, 165, 113986. <https://doi.org/10.1016/j.eswa.2020.113986>
- Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, J., & Ryu, K. (2019). An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments. *Sustainability*, 11(3), 699. <https://doi.org/10.3390/su11030699>
- Nallakaruppan, M. K., Balusamy, B., Shri, M. L., Malathi, V., & Bhattacharyya, S. (2024). An Explainable AI framework for credit evaluation and analysis. *Applied Soft Computing*, 153, 111307. <https://doi.org/10.1016/j.asoc.2024.111307>
- Pavón Pérez, Á., Fernandez, M., Al-Madfai, H., Burel, G., & Alani, H. (2023). Tracking Machine Learning Bias Creep in Traditional and Online Lending Systems with Covariance Analysis. *Proceedings of the 15th ACM Web Science Conference 2023*. <https://doi.org/10.1145/3578503.3583605>
- Ridzuan, N. N., Masri, M., Anshari, M., Fitriyani, N. L., & Syafrudin, M. (2024). AI in the Financial Sector: The Line between Innovation, Regulation and Ethical Responsibility. *Information*, 15(8), 432. <https://www.mdpi.com/2078-2489/15/8/432>
- Škorjanc, Ž. (2025). The Right to Explanation of a Credit Score: A Holistic Approach under the GDPR, AI Act, and Directive (EU) 2023/2225 on Credit Agreements for Consumers. *Global Privacy Law Review*, 6(Issue 3), 91–106. <https://doi.org/10.54648/gplr2025022>
- Sokolova, M., & Lapalme, G. (2009). A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing & Management*, 45(4), 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
- Solon Barocas, Moritz Hardt, & Arvind Narayanan. (2019). *Fairness and machine learning*.

Fairmlbook.org. <https://fairmlbook.org/>

- Tripathi, D., Edla, D. R., Bablani, A., Shukla, A. K., & Reddy, B. R. (2021). Experimental analysis of machine learning methods for credit score classification. *Progress in Artificial Intelligence*. <https://doi.org/10.1007/s13748-021-00238-2>
- Verma, S., Boonsanong, V., Hoang, M., Hines, K. E., Dickerson, J. P., & Shah, C. (2020). *Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review*. <https://doi.org/10.48550/arxiv.2010.10596>
- Verma, S., & Rubin, J. (2018). Fairness definitions explained. *Proceedings of the International Workshop on Software Fairness - FairWare '18*. <https://doi.org/10.1145/3194770.3194776>
- Zhou, N., Zhang, Z., Nair, V. N., Singhal, H., & Chen, J. (2022). Bias, Fairness and Accountability with Artificial Intelligence and Machine Learning Algorithms. *International Statistical Review*, 90(3). <https://doi.org/10.1111/insr.12492>

