

UNIVERSITÀ DEGLI STUDI DI PAVIA
Dipartimento di Scienze Economiche e Aziendali



Laurea Magistralis in Finance

**Can Shapley – Value – Based Explainability Be Used To Assess
Fairness In Credit Risk Models using the HMDA Dataset?**

Supervisor:
Paolo Guidici PhD

Experimental Thesis by:
Jana Zakaria

Academic Year 2025/2026

Table of Contents

1.	Chapter 1: Introduction	4
1.1.	Credit Risk Management	4
1.2.	Credit Decision Making	5
1.2.1.	Fairness in Credit Decision Making	6
1.2.2.	Growing Role of Machine Learning in Lending	8
1.3.	Algorithmic Bias in Credit Scoring Models	8
1.4.	Contribution to Responsible AI, Fair Lending, and Model Transparency	9
2.	Chapter 2: Literature Review	11
2.1.	Credit Risk Modeling	11
2.1.1.	Overview of Credit Scoring	12
2.1.2.	Traditional Credit Scoring Models	12
2.1.3.	Modern Machine Learning Models	14
2.2.	Fairness in Machine Learning	17
2.2.1.	Definition of Fairness	17
2.2.2.	Prior Studies on Bias in Credit Scoring and Fairness Metrics Used	19
2.3.	Explainable AI	19
2.3.1.	Model Interpretability	20
2.3.2.	Role of Shapley Values as an Explainability Tool	21
2.3.3.	Contribution of Explainability to Fairness Evaluation	22
2.4.	Research Gaps	22
2.4.1.	Limitations of Group-Based Fairness Metrics	22
2.4.2.	Integrating Explainability Methods	24
3.	Chapter 3: Data Description	25
3.1.	Overview of the HMDA Dataset	25
3.2.	Variables Description	26
3.2.1.	Dependent Variable: Loan Approved Status	27
3.2.2.	Protected Variable: Applicant Race	27
3.2.3.	Control Variables	27
3.3.	Data Cleaning and Preparation	28
3.3.1.	Data Preprocessing	28
3.3.2.	Handling Class Imbalance Through Under-Sampling	30
3.3.3.	Training and Testing Data Split	31

3.4. Descriptive Statistics and Correlations	32
4. Chapter 4: Methodology	34
4.1. Research Framework and Analytical Approach	34
4.2. Model Development.....	36
4.2.1. Logistic Regression Model	36
4.2.2. Random Forest	42
4.2.3. Neural Network.....	47
4.3. Shapley Value Computation	52
4.3.1. Concept and Theory of Shapley Values	53
4.3.1.1. Shapley Values for Logistic Regression.....	54
4.3.1.2. Shapley Values for Random Forest.....	56
4.3.2. Global Shapley Values (GSVs)	57
4.4. Fairness Assessment	60
4.5. Bias Mitigation	62
4.5.1. Propensity Score Matching (PSM)	63
4.5.2. Construction of a Balanced Dataset	64
4.5.3. Fairness Evaluation After Mitigation	64
4.6. Statistical Testing	65
5. Chapter 5: Results	66
5.1. Statistical Testing and Initial Data Patterns	66
5.2. Model Performance Evaluation using ROC-AUC	67
5.3. Explainability Results (Shapley Values).....	69
5.3.1. Feature Importance Rankings	69
5.3.2. Race Contribution Analysis	71
5.3.3. Cross Model Comparison.....	71
5.4. Subgroup Analysis	73
5.5. Bias Mitigation Results (PSM Analysis).....	75
5.6. Interpretation Findings.....	77
6. Chapter 6: Discussion	80
6.1. Key Insights on Fairness and Explainability	80
6.2. Implications for Banking and Regulation	82
7. Chapter 7: Conclusion and Recommendations.....	85

1. Chapter 1: Introduction

1.1. Credit Risk Management

Credit risk management is the process of identifying, evaluating, monitoring, and controlling the risks associated with borrowers defaulting on their debt payment obligations as per their agreement. In the context of loans, credit risk emerges if an individual or organization is either incapable of paying back a loan as per the terms of agreement or simply unwilling to do so. Since loan-making forms one of the main income sources for banks and other financial institutions, credit risk management becomes imperative (Zou & Li, 2014).

The persistence of financial problems around the globe, digitalization, technological advancements, and application of artificial intelligence in the financial sector have ensured that credit risk management continues to be in the spotlight. This is because the regulators still want transparency and other advanced functionalities from this domain. The regulators want the assurance that the banks have full knowledge about their clients and the credit risk involved with them.

Credit risk management is not only relevant to the lending decision-making process itself. Poor credit risk measurement can lead to higher default rates, financial losses, loss of capital, and even bankruptcy in extreme cases. Financial catastrophes from the past have demonstrated the adverse outcomes of poor risk management, reinforcing the requirement for sound risk measurement techniques (Oyetola et al., 2023). This is the reason why banks, as well as regulatory bodies, focus so much attention on devising effective approaches for credit risk assessment. In addition to the predictive accuracy, the modern credit risk management should take into account problems of equity, transparency, and accountability.

1.2. Credit Decision Making

Credit decisioning refers to the process by which a financial institution evaluates the creditworthiness of a borrower. Such credit evaluation determines whether the individual is eligible to receive credit, how much credit he or she can receive, and under what conditions or terms such credit will be available. The main aim of credit decisioning is to reduce risks while promoting responsible lending (Cancino & Cancino, 2019).

With the credit scores/ratings obtained, the financial institution can then determine whether to lend money to the borrower or not and under what terms and conditions. It may include determining the interest rate, the credit limit, and repayment terms. Credit decisioning plays a very important role for both the borrower and the lender since it makes sure that credit is issued responsibly (Tigges et al., 2024).

Given the financial and social implications involved in lending to borrowers and lenders, it is critical to ensure that not only is the process accurate, but it is also consistent and fair.

Credit decisions are usually made based on what is known as the "5 Cs of Credit," a well-known model that uses certain elements to determine the risk in giving credit. The "5 Cs" can be illustrated using Figure 1 as follows: Character, Capacity, Capital, Collateral, and Conditions (Gupta, 2026).



Figure 1: 5Cs of Credit Decision (Gupta, 2026)

1. Character is defined as the ability of the borrower to repay the debt obligation and it is determined by credit history, payment performance, and financial reputation.
2. The Capacity of the borrower determines their ability to pay off the loan using elements like income level, stability of employment, and outstanding loans.
3. Capital is the financial strength of the individual, comprising of the savings, asset base, or equity of the borrower in the deal.
4. Collateral is defined as the assets put up by the borrower as security against the loan and acts as a cushion for the lender in case of default.
5. Conditions include the reason for the loan, prevailing economic conditions, and interest rates among others.

1.2.1. Fairness in Credit Decision Making

The question of how to guarantee the fairness of the credit risk model and avoid discrimination is a constant problem and potential for development on the part of the lender. (Thompson, 2025).

Being aware of increased regulatory and societal scrutiny of lending practices, lenders have understood the necessity of preventing the application of biased models which may put some demographic groups and people with certain traits at a disadvantage regarding the availability of

credit (Kozodoi et al., 2022). Thus, lending decisions need to be made considering the credit worthiness of individuals and not using criteria related to personal features which are not at all related to the repayment of a loan.

Credit scoring models developed using the historical data of lending will definitely include discrimination.

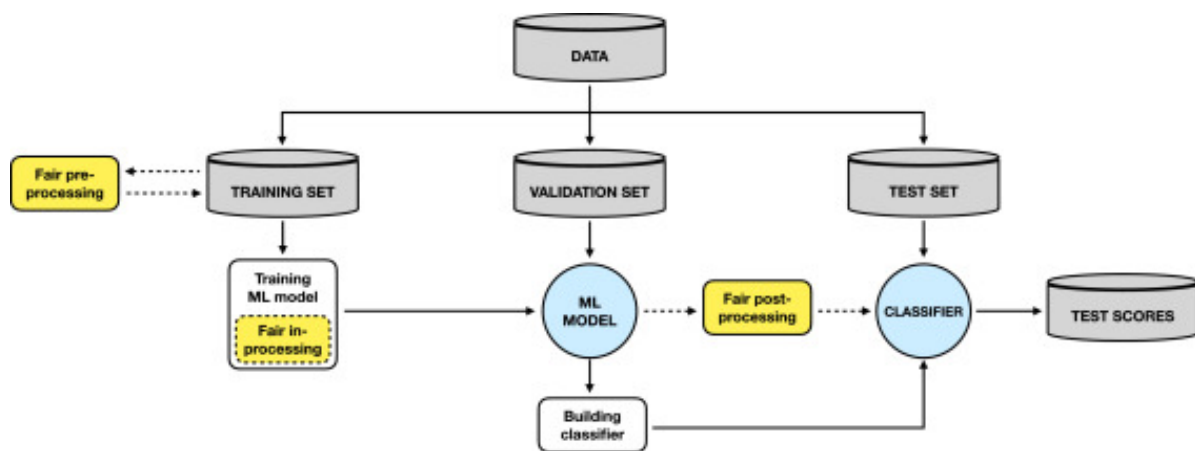


Figure 2: Fairness Intervention Points in the Machine Learning Pipeline (Kozodoi et al., 2022)

Figure 2 shows that fairness in machine learning does not happen at just one point of the model development process but can be introduced all the way through the analytical process. The methods used to mitigate bias include pre-processing, in-processing, and post-processing methods. In the former case, the bias in the data set is reduced prior to model development, in the second case, fairness constraints are integrated in the learning process, and in the latter case, post-model development adjustments are done to increase fairness. When discussing credit decision-making, these approaches are used as different ways of optimizing both prediction accuracy and fairness in dealing with loan applicants. In this research, a pre-processing approach

will be used with Propensity Score Matching (PSM), using explainability based on the Shapley value.

1.2.2. Growing Role of Machine Learning in Lending

Traditional lending has been revolutionized by machine learning in terms of increased accuracy and speed of decision-making with respect to credit assessment, due to the analysis of trends and patterns in large datasets without concentrating only on individual borrower features. Traditional scoring models were used to assess credit risk, and this technique was highly appreciated because of its simplicity and transparency. Scorecards were built through the use of theoretical knowledge of economy as well as practical experience, which made them very robust and easy to regulate. Nevertheless, traditional scoring models are still not capable of processing large amounts of data (“HOW MACHINE LEARNING IS USED IN THE LENDING INDUSTRY,” 2020)

Artificial Intelligence (AI) and Machine Learning (ML) have revolutionized the lending business by helping lending institutions evaluate the credit worthiness of customers in a fast and precise manner. AI and ML also enable inclusion of alternate sources of data in the credit worthiness model to eliminate any sort of bias that might exist in traditional credit scoring mechanisms, resulting in an increase in accuracy and inclusiveness of the lending process. In addition, AI and ML allow lenders to detect credit-worthy people who have been denied loans owing to lack of credit history.

1.3. Algorithmic Bias in Credit Scoring Models

Through the use of artificial intelligence (AI), machine learning (ML), and regression analysis, the decision making regarding lending becomes much faster and more accurate through predictive algorithms than by means of human judgment. A well-developed algorithm is able to

detect complex relationships between many factors, thus being able to give better predictions.

But, on the other hand, a poorly developed algorithm, or one which is based on wrong data, can give wrong predictions and even make the situation worse (Fernandez Vidal & Menajovsky, 2019).

Biased training data could result in discriminatory machine learning models. There are two forms of bias in training data. The first one is having a collection of prejudiced examples that the model learns from or those under-represented groups that receive a wrong evaluation. The second form of bias is when there is no training data.

It is hard to find algorithmic biases since machine learning algorithms can uncover quite complicated correlations between data. In addition, just because the predictions are accurate, it doesn't mean that decisions will be fair. Moreover, complex models like Random Forests and Neural Networks work as black-boxes, which complicates the explanation of the reason behind the prediction. This naturally leads to explainability. (Hanna et al., 2025)

Since algorithmic discrimination is frequently embedded in highly complex machine learning models, there has been an increase in the use of explainable AI approaches to determine if protected attributes affect predictions made by such models.

1.4. Contribution to Responsible AI, Fair Lending, and Model Transparency

This research contributes to the emerging topic of Responsible AI by assessing the possibility that explainability via Shapley values can help improve the fairness assessment process in credit risk models. Previous studies have mostly concentrated on the use of group-based performance metrics to measure fairness. However, such approaches may be insufficient in providing a deeper understanding of the reasons behind discriminatory outcomes. In this research, Shapley values are used to investigate not only the existence of discrimination between various groups but also

what features cause it. Moreover, the effect of using the Propensity Score Matching method as one of the techniques of reducing discrimination is assessed, thus allowing for the comparison of model behavior before and after the intervention. Using Logistic Regression, Random Forest, and Neural Network models on the HMDA data set, this research contributes to the overall goals of responsible AI by increasing transparency, promoting fair lending practices, and other considerations.

2. Chapter 2: Literature Review

2.1. Credit Risk Modeling

Credit risk modeling involves evaluating the probability of default of a borrower and the estimated loss arising from the default. Through credit risk models, the complicated information on the borrower and the finances involved can be transformed into measurable risks, which helps in making credit decisions.

Credit risk modeling serves mainly for the purpose of assisting in making intelligent decisions about lending, enhancing the precision of the risk assessment process, achieving optimal pricing of loans, managing credit portfolios, and estimating the capital requirements to cushion against any possible loss (Chatterjee, 2015). Apart from these, credit risk models allow banks to consolidate risks and make measurements in relation to all business segments, products, and regions.

For the banking industry, credit risk models are extremely important and critical elements in the context of risk management and decision-making processes. Credit risk models are utilized in the bank's operations to analyze credit scoring, loan approval, client profitability, risk-based pricing, portfolio management, and performance evaluation. In addition, credit risk models help banks to establish the amount of capital needed and to comply with the regulatory requirements (*CREDIT RISK MODELLING: CURRENT PRACTICES AND APPLICATIONS*, 1999). Despite the development of the credit risk models and their significant contribution to the internal risk management process of a bank, regulators need them to be conceptually justified, empirically tested, and consistent enough to use the credit risk models for regulatory capital calculations. In comparison with market risk models, credit risk models face some difficulties.

2.1.1. Overview of Credit Scoring

Credit scoring is the method that lenders use to evaluate the creditworthiness of borrowers and make the decision whether they will qualify for credit. It is carried out through scoring of the applicants using several variables including credit history, level of outstanding debts, credit history duration and credit activity among others. The resulting score indicates the credit risk levels of the applicants and enables lenders to make appropriate decisions (Brock, 2026).

In credit scoring, the probability of the borrower to default on the loan is usually estimated through a numeric credit score; the higher the score is, the less probable is the default on the loan. These scores are computed based on the variables reflecting the financial behavior of the borrower. Despite the common use of some variables, such as payment history and existing debts for all types of loans, the choice of the variables to estimate the credit score depends on the type of the loan being requested. For instance, credit card scoring includes more weight of credit utilization and account history, whereas mortgage scoring takes into account down payment size, employment stability, and loan size among other variables (Markov et al., 2022). With time, the methodology for developing credit scoring models has changed considerably and moved from conventional statistical modeling towards machine learning that can process more sophisticated data.

2.1.2. Traditional Credit Scoring Models

In traditional credit scoring models, statistical techniques are mainly utilized in order to estimate the probability of a person's defaulting. Among all these techniques, **Logistic Regression** has been the most popular due to its simplicity, comprehensibility, and effectiveness in predicting credit risks. Through estimation of the probability of default on the basis of characteristics of the borrowers, logistic regression gives banks an opportunity to have a clear tool for their decisions

in regard to the credit (Bolton, 2010). Despite its widespread use in the financial sphere, the limitations of logistic regression in detecting non-linear dependencies in huge data sets have stimulated the development of machine learning models.

Logistic regression is one of the most popular methods used in credit scoring to predict loan default. It utilizes probability estimates and logistic transformations to classify customers based on their risk (M. Vasuki, Mbonigaba Celestin, A. Dinesh Kumar & Paul Johnson Asamoah., 2025).

Another traditional scoring model for assessing the creditworthiness of individuals is the **Linear Discriminant Analysis** (LDA). LDA is a supervised classification algorithm, which is designed to differentiate between various risk groups of borrowers (Reza et al., 2024). The main idea behind LDA is the dimension reduction of the dataset through selecting the linear combinations of predictor variables, which maximize separation among classes. By performing the dimension reduction, LDA resolves problems related to high dimensional data, including overfitting, increased computational costs, and sparseness.

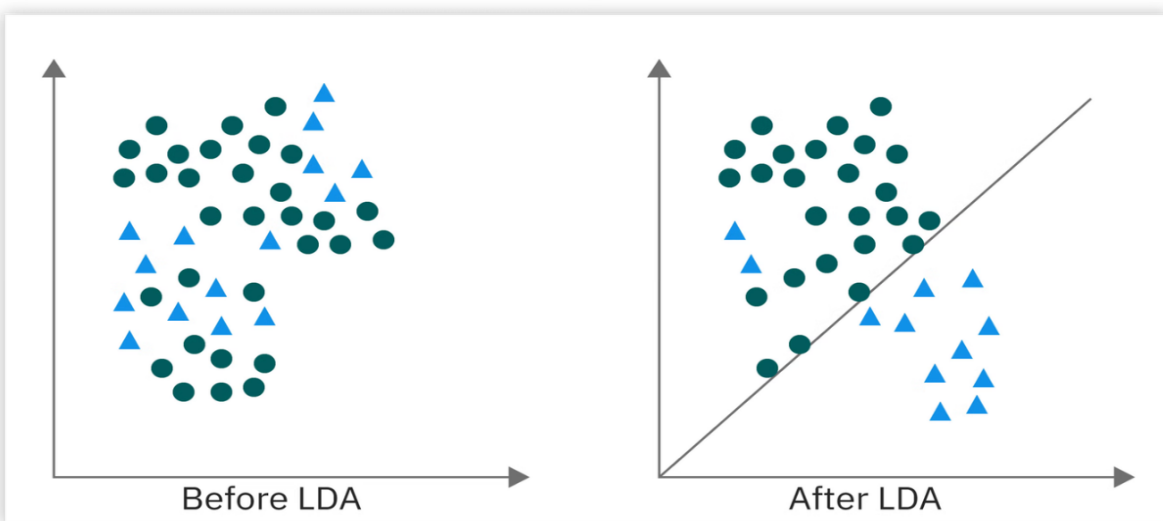


Figure 3: *Illustration of Class Separation Before and After Linear Discriminant Analysis (LDA) (What Is Linear Discriminant Analysis (LDA)?, 2026)*

Figure 3 demonstrates the impact of Linear Discriminant Analysis (LDA) in data classification. Prior to LDA (left side), the classes show substantial overlapping, making it hard to differentiate between them. After LDA is applied (right side), the classes are transformed into a lower dimensional space that separates the two classes from one another. This makes the differentiation between the two classes easier as a result of finding the best possible linear boundary for them. This flexibility makes LDA suitable for multi-class classification problems and different from logistic regression in that the latter can only perform binary classifications. Therefore, LDA is commonly used to improve the performance of other machine learning algorithms for classification such as decision trees and random forests.

2.1.3. Modern Machine Learning Models

With the growing availability of large and sophisticated databases, machine learning algorithms have found their place in the field of credit scoring. As opposed to traditional statistical methods, which usually imply certain assumptions concerning the nature of relationships between different parameters, machine learning allows for automatic detection of complicated and nonlinear relations within a dataset. Such an approach contributes to increased accuracy of prediction and helps to detect creditworthy customers (Demir, 2025). Some of the most common machine learning techniques applied to credit scoring include **Random Forest** and **Neural Networks**. These algorithms outperform many traditional approaches, especially when working with high-dimensional and large-scale databases.

The Random Forest algorithm is an ensemble technique that improves the predictive accuracy of algorithms by using several decision trees together while preventing overfitting. A conventional

RF algorithm involves X decision trees, all of which vote for the outcome and the final output is usually the average (in case of regression) or the mode (in case of classification). This ensures that there is a low bias while utilizing the power of ensemble (Agiza et al., 2025).

Random Forest

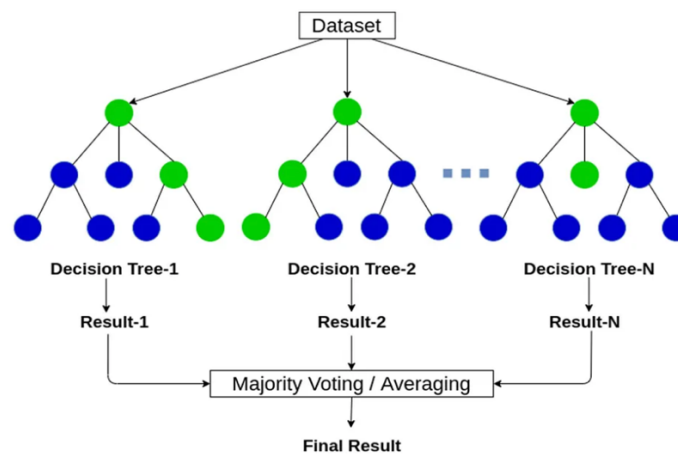


Figure 4: Random Forest – Workflow (Jain, 2024)

As shown in Figure 4, in every node of the decision tree, rather than using all the features, random selection of features is done. The randomness further adds up, and the trees become uncorrelated and thus, the model becomes more generalizable. This technique addresses the problem of overfitting in case of decision trees.

The next Modern Machine Learning Model is Neural Networks. Neural Networks improve credit rating systems by making use of huge and complicated databases to predict the probability of repayment of a loan by the borrower. Neural networks search for any hidden nonlinear relationships within the data, which cannot be seen by conventional modeling techniques like linear regression (Vladimir K Zhirov, 2021).

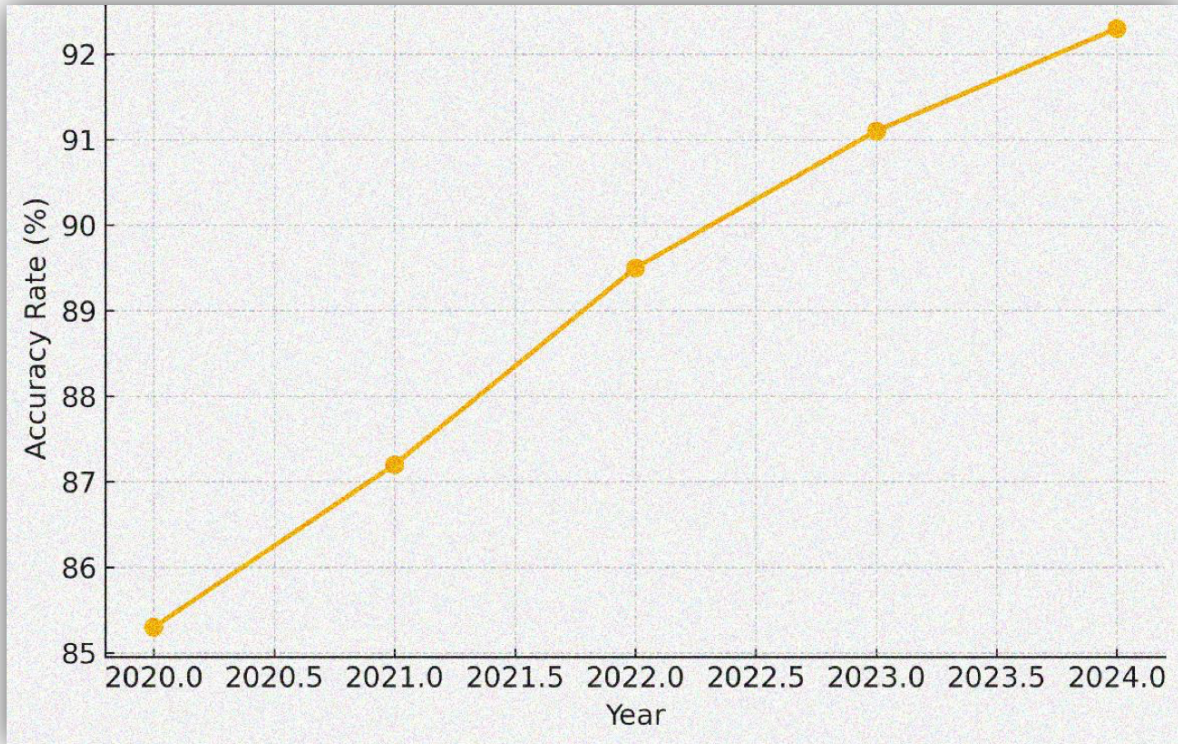


Figure 5: *Progressive Improvement in the Accuracy of Credit Scoring Models Over the Years (M. Vasuki, Mbonigaba Celestin, A. Dinesh Kumar & Paul Johnson Asamoah., 2025).*

Now How Has the Accuracy of Credit Scoring Models Improved with Modern Machine Learning Techniques? As can be seen from Figure 5, the accuracy of credit scoring models has been increasing due to the introduction of new machine learning algorithms. The application of logistic regression resulted in 85.3% accuracy in 2020 and increased to 87.2% after the introduction of the Random Forest models in 2021. The use of the Gradient Boosting model allowed to increase the prediction accuracy to 89.5% in 2022 and, by 2023, Artificial Neural Networks achieved 91.1%. In 2024, the highest accuracy was achieved with the use of Support Vector Machines (SVMs) – 92.3%. (M. Vasuki, Mbonigaba Celestin, A. Dinesh Kumar & Paul Johnson Asamoah., 2025).

2.2. Fairness in Machine Learning

Many financial organizations have turned to machine learning (ML) in the decision-making process. The growing trend of algorithmic decision-making has resulted in the emergence of numerous studies focused on fair machine learning. ML algorithms are used in the development of risk scorecards and are applied in the credit-related decision-making process by financial organizations. Nevertheless, there is a lack of literature regarding fair ML in credit scoring (Kozodoi et al., 2021).

Machine learning algorithms are capable of processing huge amounts of information in an efficient way, which helps lenders estimate the creditworthiness more precisely and make loan decisions with less probability of default. On the other hand, machine learning algorithms used in the assessment of credit risks can cause certain issues such as unintentional discrimination among sensitive social groups (Liu Xuemeng, 2024). Many researchers have tried to solve the problem in different ways, but choosing the most efficient mitigation strategy remains a problem.

2.2.1. Definition of Fairness

The role of machine learning in the credit market involves distinguishing applicants who have a high potential of paying back the loan and those who have a higher probability of defaulting. In this process, machine learning models can also discriminate against applicants who belong to protected classes such as race, gender, age, and ethnicity. This form of discrimination is unintentional and occurs due to biases in the training data set or biases in the machine learning algorithm itself (Christophe Hurlin et al., 2024). As a result of this process, the credit score models used for evaluating the creditworthiness of loan applicants can discriminate against particular demographic groups by making it less likely for them to get approved or increasing their credit costs. The concept of fairness in machine learning thus ensures that the basis for

lending decisions is creditworthiness and not any protected attribute or indirect proxy of it (Christophe Hurlin et al., 2024).

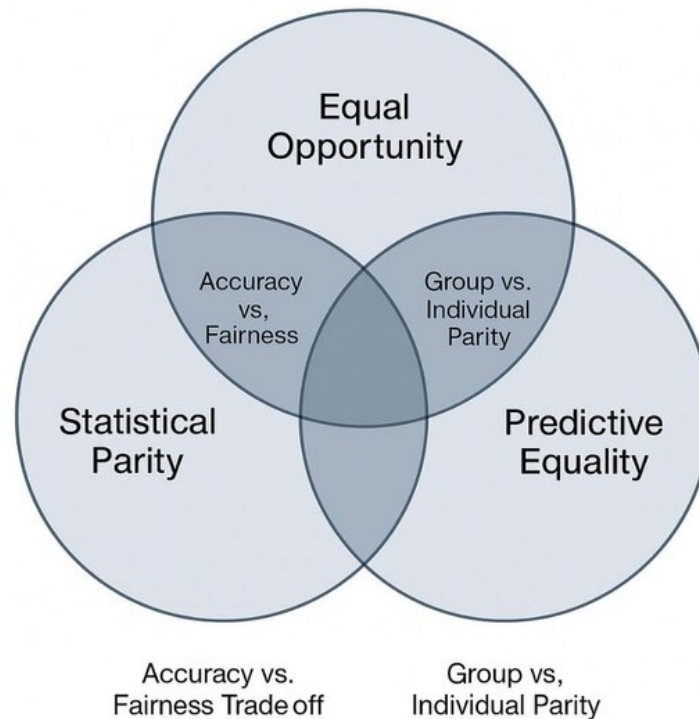


Figure 6: Venn Diagram of Overlapping Fairness Definitions (Statistical Parity, Equal Opportunity, Predictive Equality) (Clement et al., 2025)

Figure 6 shows that it is not possible to define the concept of fairness in machine learning using a single metric. In this regard, there have been a number of concepts of fairness which stress different aspects of equity of decisions. Statistical parity is concerned with achieving similar probabilities of positive outcomes, like approvals of loan applications, irrespective of the membership of particular demographic groups. On the other hand, the concept of equal opportunity involves equal chances of obtaining a positive outcome for applicants who possess similar levels of credit worthiness. Finally, predictive equality is concerned with ensuring similar error rates, especially false positives, among protected groups (Clement et al., 2025). From the

above figure, it can be seen that the fairness metrics used above have a lot of similarities, but they also entail important trade-offs. For instance, increased fairness may entail decreased predictive accuracy and the satisfaction of one fairness measure may entail difficulties in meeting another at the same time. This explains why there is no universal definition of fairness.

2.2.2. Prior Studies on Bias in Credit Scoring and Fairness Metrics Used

According to recent research, algorithmic bias is still a major problem in credit scoring algorithms driven by artificial intelligence. For instance, the article "Algorithmic Bias and Fairness in AI Credit Scoring: Evidence, Mechanisms, and Governance Responses" assessed evidence drawn from 15 empirical studies carefully selected from an initially screened database of 55 peer-reviewed papers written between 2018 and 2025. The study found that bias in credit scoring algorithms based on AI technology is seldom caused by technical issues. On the contrary, the bias appears mostly due to historical biases present in training data sets, the presence of proxy variables associated with protected attributes, and the lack of institutional disincentives for discrimination where regulatory oversight is weak (Sajid & Amjad, 2026).

2.3. Explainable AI

Explanation of AI (XAI) in credit scoring implies methods of making machine learning more transparent through explanations about why some predictions were made. Unlike mere prediction about approving or rejecting loans, XAI helps to understand which features of applicants like their income, credit score, debts, or job position influenced the outcome of the prediction the most. XAI is especially significant since many advanced credit scoring systems such as Random Forests and Neural Networks are black-box models that are hard to interpret.

Use of AI for better predictive power along with XAI for sufficient explanations can help to improve existing credit scoring systems (de Lange et al., 2022). Yet one issue in using credit scoring as a classification problem is that only a small number of people among all customers tend to default, meaning that the dataset is imbalanced. The SHAP values are identified as the most reliable and effective ones for explanations about the importance of different features of the model (de Lange et al., 2022).

2.3.1. Model Interpretability

The inclusion of artificial intelligence (AI) in credit scoring and loan underwriting systems has enhanced the accuracy of the predictions as opposed to the use of conventional methods of credit evaluation (Ravi et al., 2025). The problem is that most AI algorithms rely on the principles of black box and hence, there is limited transparency in how loan underwriting decisions are made since the workings inside the algorithm cannot be easily comprehended. This has been a cause of concern in respect of issues such as fairness, accountability, and regulation. Therefore, Explainable Artificial Intelligence (XAI) has become increasingly relevant as a result (Ravi et al., 2025).

Credit scoring problems can be solved by black box algorithms like Random Forests, Neural Networks, and certain boosting methods. These black box models have great predictive accuracy; however, due to their opaque nature, it becomes very hard to understand how different features affect prediction results. Therefore, it becomes crucial to use Explainable AI approaches such as Shapley values.

Lack of explainability in AI makes financial decisions vulnerable to loss of trust, lack of compliance, and customer disengagement. Explainable AI is required in financial organizations since it helps create transparency and accountability as well as responsible decision-making. This

will improve customer engagement, innovation compliance, and business decision-making (Kelly Bailey, 2025).

2.3.2. Role of Shapley Values as an Explainability Tool

One of the major hurdles in the adoption of sophisticated machine learning algorithms in credit scoring is their poor interpretability. While such machine learning algorithms may deliver high accuracy in prediction, they often fail to generate an explanation of predictions that humans can understand (Rivalani Hlongwane et al., 2024). The inability to explain predictions generated by these algorithms has become a cause for worry for regulators, especially in the banking industry.

The SHapley Additive exPlanations (SHAP) framework was created to overcome this difficulty using ideas from cooperative game theory. SHAP calculates Shapley values, which quantify the contribution of each input characteristic to an individual model prediction. It was first developed to estimate the equitable distribution of rewards among players in a cooperative game. These figures explain how many factors affect credit choices in an understandable and straightforward manner (Rivalani Hlongwane et al., 2024).

SHAP values are predicated on the notion that each potential feature combination's output should be taken into account in order to assess its significance. This implies that, using the same hyperparameters and training data, a different model should be developed for every possible combination of the available features (Chiusano, 2022). After obtaining all of these models, SHAP values may be calculated by comparing the model outputs when each particular feature is used and when it is not.

SHAP has emerged as one of the most popular explainability methods for interpreting complex machine learning models in recent years. It improves model transparency, boosts confidence in

automated credit scoring systems, and supports fairness evaluation in credit risk modelling by allowing academics and practitioners to comprehend the reasons influencing loan acceptance or rejection decisions (Rivalani Hlongwane et al., 2024)

2.3.3. Contribution of Explainability to Fairness Evaluation

Fairness metrics are numerical criteria that allow assessing whether the results generated by a machine learning model have any kind of biases among various demographic groups. Instead of providing a rationale for those biases, fairness metrics determine whether they exist or not by comparing the model outputs or errors among protected and unprotected groups. Therefore, fairness metrics can be effectively utilized to assess credit scoring models for biases toward particular races, genders, or ages of applicants (Zhou et al., 2022).

2.4. Research Gaps

Although there has been considerable research done on issues of fairness in machine learning and the use of techniques for explainable AI, some research gaps still exist. While research has shown some success in quantifying biases in algorithms and increasing the interpretability of machine learning models, there has been little consideration given to integrating these methods in evaluating fairness in credit risk models.

2.4.1. Limitations of Group-Based Fairness Metrics

Fairness metrics, such as Demographic Parity, Equal Opportunity, and Equalized Odds, identify unfair discrimination based on comparison of how the model's positive predictions, accuracy rates, and error rates vary in different demographic subgroups.

Group fairness means that outcomes must be distributed equally between groups. There are many types of group fairness criteria, one of which is called demographic parity. It identifies whether the rates of positive predictions are equal for all groups. If we consider the case of binary

classification, then demographic parity implies that selection rates among groups are the same, which means that the probability of getting positive predictions does not depend on the group (MacPherson, 2026).

Demographic Parity could help achieve fair outcomes based on equal access if the historical categorizations of people have been done using partial information and under biased conditions. However, since demographic parity is based on selection rates and not on error rates, it cannot provide any indication regarding whether the algorithm misses qualified individuals and directs too many people from one demographic category for screening, or performs poorly in case of a demographic category which had lower representation in the training data set (MacPherson, 2026). When independent validation is possible for an outcome, demographic parity must be used alongside other metrics.

A fairness metric may show that loan approval rates vary among demographic groups in the context of credit risk modelling, but it is unable to ascertain whether these differences are caused by particular model characteristics, such as income, employment status, loan amount, or other financial variables. As a result, fairness measurements by themselves are unable to shed light on how intricate machine learning algorithms make decisions (Carey, 2020). This restriction emphasises the significance of explainability methods, including Shapley values, which may determine how each attribute contributes to each prediction. Shapley-based explanations offer a fuller understanding of potential sources of unfairness in credit scoring models and can assist determine whether particular variables disproportionately influence forecasts for different groups by looking at feature-level contributions.

2.4.2. Integrating Explainability Methods

Prior studies have primarily looked at model explainability and fairness evaluation as two separate fields. Fairness measures can show whether a model generates different results for different groups, but they don't give much information on the characteristics that cause these differences. Similarly, explainability techniques are not always used to look at fairness issues, but they can show how individual qualities affect predictions. In order to close this gap, this work uses the HMDA dataset to analyze the drivers of potential bias in credit risk models by combining fairness assessment with explanations based on Shapley values.

3. Chapter 3: Data Description

3.1. Overview of the HMDA Dataset

One of the most publicly accessible datasets for mortgage lending analysis in the US, the Home Mortgage Disclosure Act (HMDA) database, is used in this study. Under the Home Mortgage Disclosure Act, the Consumer Financial Protection Bureau (CFPB) is responsible for gathering and maintaining the HMDA dataset. Its major goals are to monitor lending practices, encourage openness in mortgage lending, and assist in enforcing fair lending laws. The dataset is a useful tool for study on credit risk, discrimination, and algorithmic fairness since it includes comprehensive data on mortgage loan applications, borrower attributes, and lending decisions.

One reason why the HMDA data is suitable for the current research is that the dataset includes both financial features and demography data such as the race of an applicant and thus allows for the measurement of any possible biases present in the decision-making process. In contrast to many other credit datasets, HMDA contains protected attributes and thus helps researchers analyze whether there are differences in model outcomes among different groups of people while accounting for their financial features.

Similarly to the methodological approach used in the reference study, the current research is based on 157,269 loan applications submitted in 2017 in the state of New York. The dataset was processed, and all unnecessary variables were removed.

Table 1 displays the variables employed in this study. Declined Loan is the dependent variable.

It is a binary variable that shows whether a loan application was accepted (0) or denied (1).

Applicant Race, which is recorded as White (0) and African American (1), is the protected

variable that is used to assess possible racial bias. Lien Status, Loan Type, Loan Amount, Loan Purpose, Applicant Gender, and Applicant Income are the explanatory (control) variables. Credit risk modelling frequently makes use of these variables, which reflect important financial and demographic traits that impact lending decisions.

Variable definitions.

Variable	Definition
<u>Dependent variable</u>	
Declined Loan	The variable takes the value 1 if the lender declines a loan. Specifically, if a loan application initially satisfies the approval requirements of Government-sponsored enterprise (GSEs) or Federal Housing Administration (FHA), but it fails to satisfy the lenders' requirements. It takes the value 0 if the lender approves the loan.
<u>Protected variable</u>	
Applicant race	The variable takes the value 1 if the applicant is African American and 0 if White.
<u>Explanatory variables</u>	
Applicant gender	The variable takes the value 1 if the applicant is male and 0 if female.
Applicant income	Natural logarithm of the applicant's gross annual income (in thousands of dollars) the lender relies on when making the credit decision.
Loan amount	Natural logarithm of the amount (in thousands of dollars) of the covered loan, or the amount applied for.
Loan purpose	The variable takes the value 1 if the purpose of seeking a loan was for refinancing the mortgage and 0 if purchasing a home.
Lien status	This variable takes the value 1 if the loan application is secured by 'first lien' and 0 for a sub-ordinate lien. A first lien level of security indicates that the lender is the first to be paid when a borrower defaults and the property or asset is used as collateral for the debt.
Loan type	The variable takes the value 1 if the loan was insured by the FHA and 0 if insured by a GSE.

Table 1: Variable Definitions (Babaei et al., 2025)

3.2. Variables Description

The variables used in the empirical analysis are explained in this section. The factors were chosen because they were relevant to the assessment of credit risk and fairness. They are divided into three groups: the dependent variable, which stands for the loan decision; the protected variable, which is used to evaluate potential bias across demographic groups; and the control variables, which record the loan and financial characteristics of applicants that affect credit decisions.

3.2.1. Dependent Variable: Loan Approved Status

Declined Loan, which reflects the result of every home loan application, is the dependent variable employed in this study. This binary variable is coded as 0 if the loan application was accepted and 1 if it was rejected, in accordance with **Table 1**. It acts as the target variable in the predictive models, enabling the study to determine the likelihood that a loan application would be denied in light of the loan, financial, and demographic factors of the applicant. This variable is crucial to the fairness assessment carried out in this study since it forecasts loan approval results and serves as the foundation for assessing whether the models generate different decisions across ethnic groupings.

3.2.2. Protected Variable: Applicant Race

The protected variable in the present research is "**Applicant Race**," and it is the one on the grounds of which the possible presence of racial discrimination in decision making regarding credit lending will be assessed. According to the classification provided in **Table 1**, the race will be defined by means of coding as a binary variable where the African American will be equal to 1 and White will be equal to 0. The variable does not influence the prediction of an individual's credit worthiness, but it helps to identify whether there are any differences between the predictions of protected and non-protected classes in the models. The study assesses whether racial differences exist in loan acceptance decisions and assesses the fairness of the established credit risk models by comparing model predictions across different groups.

3.2.3. Control Variables

This study comprises a number of control factors that reflect the loan, financial, and demographic aspects of the applicants in addition to the dependent and protected variables.

Because financial institutions frequently take these factors into account when evaluating creditworthiness, they are included in the predictive models. **Applicant Gender, Applicant Income, Loan Amount, Loan Purpose, Lien Status, and Loan Type** are the control variables shown in **Table 1**. While Applicant Income represents the borrower's stated yearly income, Applicant Gender makes a distinction between male and female applicants. Loan Purpose specifies the intended use of the loan, such as home purchase or refinancing, while Loan Amount indicates the entire amount sought by the borrower. Loan Type describes the kind of mortgage product, while Lien Status shows the priority of the lender's claim on the property. These factors collectively represent the loan-specific and financial traits that affect lending decisions and serve as the foundation for the creation of the credit risk models employed in this study.

3.3. Data Cleaning and Preparation

The HMDA dataset was subjected to a number of data preparation procedures prior to the development of the credit risk models in order to guarantee its quality and suitability for analysis. These procedures included splitting the dataset into training and testing subsets, preprocessing the data, and using undersampling to correct class imbalance. To enhance model performance and guarantee trustworthy and objective findings, proper data preparation is crucial.

3.3.1. Data Preprocessing

The HMDA dataset underwent a number of preprocessing procedures prior to model creation to make sure it was appropriate for machine learning analysis as shown in **Figure 7**. First, the dependent variable (Declined Loan), the protected variable (Applicant Race), and the chosen control variables were the only variables kept that were relevant to the study's goals. To enhance the quality of the data, observations with inconsistent or missing values were eliminated. A binary result, with 1 denoting a rejected loan application and 0 denoting an approved application,

was used to encode the dependent variable (**Table 2**). The remaining categorical variables, such as Applicant Gender, Loan Purpose, Loan Type, and Lien Status, were turned into numerical representations that were compatible with the machine learning methods, while Applicant Race was transformed into a binary protected attribute. Following these preparatory stages, the dataset was ready for model training and class balance.

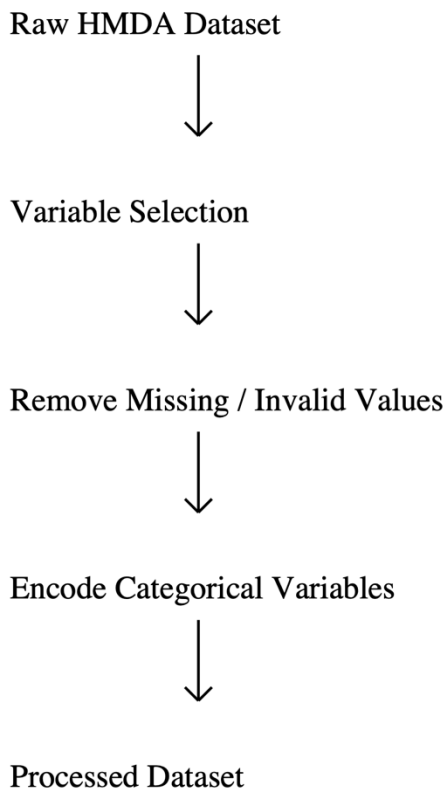


Figure 7: *Data Preprocessing Workflow*

Variable	Original Values	Encoded Values
Declined Loan	Approved / Declined	0 / 1
Applicant Race	White / African American	0 / 1
Applicant Gender	Female / Male	0 / 1

Table 2: Variable Encoding Table

3.3.2. Handling Class Imbalance Through Under-Sampling

Filtering of HMDA data set was initially done such that it included only the White and Black or African Americans applicants, considering the objective of this research paper. The data set obtained had 157,269 cases, 145,055 White applicants (92.2%) and 12,214 Black applicants (7.8%). Outcome variable was modified into binary variable where 1 denotes originated loan and 0 means application approved but not originated. The data thus obtained had 147,255 originated loans (93.6%) and 10,014 approved but not originated applications (6.4%).

In addition, a final selection of 15 variables was made for the analysis, which included the dependent variable (loan outcome), the protected race variable, and other relevant applicant, loan, and census tract attributes. Any variable that might lead to data leakage, including denial reason, purchaser type, rate spread, and anything beyond the decision point, was discarded. The race variable was further coded separately as the protected attribute with 1 standing for Blacks and 0 for Whites.

The missing value analysis revealed that there were 88 records without any information on the census tract. Since they constituted a very small percentage of the whole data set, these were dropped to leave a data set of 157,181 records without any missing values. These consisted of

147,249 originated loans and 9,932 approved but not accepted applications with 144,969 White and 12,212 Black applicants.

Dataset	Loan Originated (1)	Approved But Not Accepted (0)	Total
Before Under-sampling	103,073 (93.7%)	6,951 (6.3%)	110,024
After Under-sampling	6,951 (50%)	6,951 (50%)	13,902

Table 3.1: Class Distribution Before and After Under-Sampling

3.3.3. Training and Testing Data Split

In the second step, the data was split into 70% training data and 30% testing data through stratified sampling by taking both loan outcome and race into consideration. Accordingly, 110,024 training and 47,157 testing observations were obtained. The initial number of training observations comprised 103,073 originated loans and 6,951 approved but not accepted loans. In order to resolve the issue of class imbalance in loan outcome, under-sampling technique was utilized only for the training data set and reduced the majority class to 6,951 observations. As a result of this process, the balanced training data set consisted of 13,902 observations, of which 6,951 observations belong to each class (50% each) (**Table 3.1**).

3.4. Descriptive Statistics and Correlations

Variable	Correlation with Black
Loan originated	-0.045
Loan type	0.156
Applicant sex	-0.095
Applicant income	-0.035
Minority population	0.461
Tract-to-MSA/MD income	-0.124
Owner-occupied units	-0.168
1-4 family units	-0.121

Note: Bold values denote significant correlations ($p < 0.05$).
Correlation with Black is Pearson's r .

Table 3.2: Correlations Between Race and Selected Variables

Race	Approved But Not Accepted (%)	Loan Originated (%)
White	6.00%	94.00%
Black	10.06%	89.94%

Table 3.3: Loan Outcome by Race

There are some significant differences between the statistics related to black and white applicants. While the loan origination rate was 94.00% for white applicants, it was 89.94% for black applicants, which is a difference of about 4.06% (**Table 3.3**). Also, the black applicants' income was lower (\$107.76 thousand compared with \$153.10 thousand for white applicants), and black applicants were linked to tracts with substantially higher minority population shares (66.29% compared with 22.17%). The results from the initial correlation analysis also demonstrated a weak negative correlation between race and loan origination ($r = -0.045$) and a strong positive correlation between race and minority population ($r = 0.461$) (**Table 3.2**). While these statistics do not prove discrimination on their own, they provide the grounds for differences to be investigated using machine learning and fairness methods.

4. Chapter 4: Methodology

The methodological approach for examining racial fairness in machine learning-based credit risk assessment is presented in this chapter. It outlines the analytical techniques used on the HMDA dataset, including post-mitigation analysis, Shapley-value-based explainability, model evaluation, fairness assessment, and data preprocessing and model creation. In order to determine whether such inequalities may be found and minimized, the chapter also describes how Propensity Score Matching (PSM) and statistical testing work. When combined, these techniques offer an organized strategy for answering the study question and evaluating the chosen credit risk models' prediction performance and fairness.

4.1. Research Framework and Analytical Approach

The methodology of this research is built in such a way that it will be possible to measure the racial fairness of the machine learning approach to credit risk assessment as well as to determine how the techniques for building the explanation could help identify possible discriminatory practices. The analysis is conducted on the Home Mortgage Disclosure Act (HMDA) data and involves sequential steps of prediction, explainable artificial intelligence, fairness assessment, and post-mitigation evaluation.

The research methodology starts with data preprocessing, which involves choosing relevant variables, handling missing values, developing the loan outcome variable in binary form, and defining race as the protected attribute. The processed data is then partitioned into the training and testing datasets with a ratio of 70/30 depending on loan outcomes and race through stratification. To deal with the loan imbalance in the training dataset, under-sampling is used, while the test data set remains the same.

Next, three machine learning algorithms: Logistic Regression, Random Forest, and Neural Network, are created utilizing the training dataset. The performance of each algorithm is measured by ROC-AUC on the testing dataset. This enables the study to determine whether there is enough discriminatory power provided by these models prior to examining the fairness implications of the study. The prediction of each model is further examined by applying Shapley value that explains the contribution of the predictors into the model prediction. This enables the study to go beyond understanding whether disparities exist between the model outputs to find out what variables make up these differences.

Post the explainability analysis, the next step involves a group-based fairness analysis that involves comparing model predictions and associated fairness metrics for Black and White applicants. In cases where differences are noted, Propensity Score Matching (PSM) is performed to mitigate any differences in terms of creating groups that are more comparable through observed features. The next step involves re-analyzing the models and fairness metrics post-mitigation in order to see whether the noted differences can be mitigated without compromising on the predictive performance of the model. Lastly, statistical testing is implemented to assess the statistical significance of observed differences before and after mitigation.

The whole analytical process can thus be represented as follows:

HMDA data -> Preprocessing -> Train/Test (70/30 Stratified) -> Under-sampling -> Model building -> Model Evaluation (using ROC-AUC metric) -> Shapley Values explanation -> Fairness check -> Propensity Score Matching -> Fairness evaluation after mitigation -> Statistical significance test.

The framework provides an answer to the research question by incorporating explainability in the process of assessing fairness. The methodology does not rely on group-level indicators for measuring fairness but employs Shapley values to pinpoint the predictors behind each prediction made and to determine whether there are any differences between races in the contributions of those features. Further PSM and statistical analysis enable the examination of whether or not those differences can be overcome.

4.2. Model Development

In this section, I will discuss the processes that have been adopted to clean the HMDA dataset before running the machine learning algorithm. The processes include variable selection, cleaning the dataset, dealing with missing values, creation of the target and the protected variables, addressing the problem of class imbalance using under-sampling technique, and the process of stratified split of the dataset into training and testing datasets.

4.2.1. Logistic Regression Model

The Logistic Regression model was selected as the first machine learning algorithm for prediction of the result of a mortgage loan application process. The selection was made on the grounds that the dependent variable is of binary type, therefore the algorithm can be applied to assess the probability that an observation falls into either of two outcome categories. As far as this research is concerned, it is required to predict whether the application results in the origination of a loan or not. Moreover, the application of Logistic Regression in this study is particularly relevant since the algorithm allows for interpreting the contribution of each particular factor to the result of prediction.

Logistic Regression model was built based on the cleaned HMDA data as presented in Section 4.1. All the procedures mentioned in the above section relating to variable selection, handling of

missing observations, creation of the target and protected variables, class balancing by under-sampling, and stratified split of the data set were performed prior to building the model. This means that the Logistic Regression model was built on the preprocessed data and not the raw HMDA data.

Dependent variable was represented as binary response for the loan application results.

Independent variables were comprised of characteristics of the applicant and the loan which were selected in the data preprocessing phase. These variables were employed in Logistic Regression model to predict the probability of loan outcome.

Logistic Regression generates probabilities and not actual outcomes. A threshold was thus set to classify the probability into an outcome. In this case, a threshold of 10% was applied to the generated probability. The same threshold was maintained for the generation of all the predicted outcomes that would be evaluated in the later stages of the analysis.

	Application approved but not accepted	Loan originated
Application approved but not accepted	2326	12946
Loan originated	7688	134309

***Table 4.1:** Confusion Matrix for Logistic Regression Model Before Splitting*

LR is used here to predict that a loan will be originated vs not using a 10% threshold to classify a default. The model predicts whether a loan is approved or rejected using this threshold. The confusion matrix (**Table 4.1**) gave an accuracy of 87%, an error rate of 13%, and a false positive

rate of 8.79%. Here I am not evaluating accuracy or fairness metrics yet, I am only understanding how the model behaves.

After going through the data preparation process, the Logistic Regression model was evaluated using the training and testing approach. For fairness analysis, I split the full data set into 70% for training and 30% for testing unseen data in order to prevent overfitting and because Logistic Regression Model done above overestimates performance. This was done to check whether the model is capable of generalizing itself on the data that was not used to train the model. The best model approves and rejects both cases and do not collapse into one decision.

Fairness issues in credit scoring involve issues of asymmetric errors, especially false positives and false negatives. The model may appear to be accurate in the training phase but prove detrimental to some groups when applied to new applicants. Thus, both the performance metrics and fairness metrics are evaluated using the test data only.

Thus the goal is to: Train on known data -> Test on unseen data -> Evaluate accuracy.

	Application approved but not accepted	Loan originated
Application approved but not accepted	780	4405
Loan originated	2224	39771

Table 4.2: Confusion Matrix for Logistic Regression Model After Splitting Data into 70/30

The test predictions were assessed using a confusion matrix, accuracy, error rate, FPR, and AUC.

The use of these metrics was motivated by the fact that accuracy might not be an adequate

measure of a classifier's performance because there are different types of classification errors in the case of the credit risk scenario.

The confusion matrix generated using the Logistic Regression model on the test set is illustrated in **Table 4.2**. This matrix includes the number of observations that have been accurately and inaccurately classified by the model in the two outcome groups.

It is clear that the model has been able to classify 780 observations in the first group and 39,771 observations in the loan originated group. On the other hand, 4,405 observations have been classified in the first group despite the fact that their outcome was loan originated, while 2,224 observations were classified as loan originated when the actual outcome belonged to the first category.

Thus, the confusion matrix offers a more comprehensive approach to evaluating the model performance in comparison to accuracy. For example, with the use of the confusion matrix, it is possible to detect different types of classification errors that can then be used in the calculation of error rates such as the false positive and false negative rates. This point is particularly relevant to the current research as these metrics are further utilized to determine whether the model results vary among different protected groups.

Table 4.3 displays the overall performance metrics derived from the Logistic Regression model.

Performance Measure	Logistic Regression
Accuracy	85.95%
Error Rate	14.05%
False Positive Rate	9.97%
AUC	0.6344

Table 4.3: Performance of the Logistic Regression Model on the Test Dataset

With an accuracy of 85.95%, the Logistic Regression model successfully identified almost 86% of the test observations. 14.05% was the comparable error rate. The false positive rate was 9.97%, meaning that the classification framework used in this investigation mistakenly classified about 10% of the observations that belonged to the appropriate negative class as positive.

While the accuracy provides a general measure for classifying performance, it does not give any insight into how well the model can discriminate between the two classes at various probability levels. Thus, ROC analysis was also performed.

To measure the effectiveness of the Logistic Regression classifier in separating the two output classes based on various classification thresholds, an ROC curve was generated. The ROC curve illustrates the correlation between the true positive rate and the false positive rate at different classification thresholds.

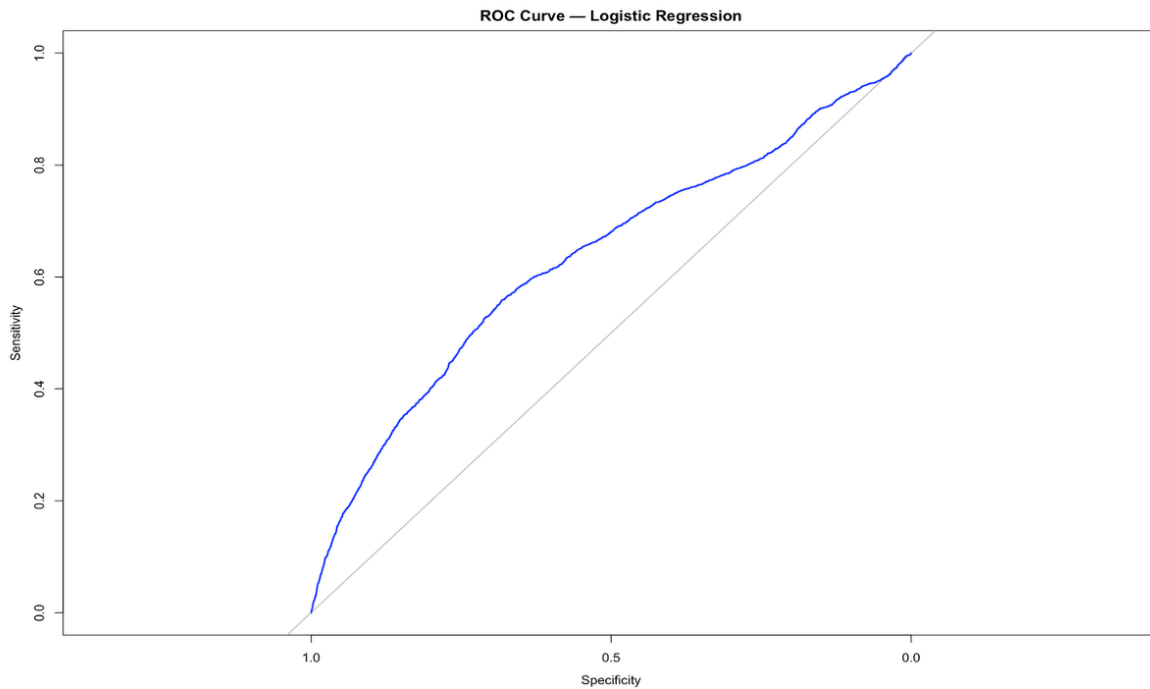


Figure 8: ROC Curve for the Logistic Regression Model

The AUC of the Logistic Regression model was roughly 0.6344 (**Figure 8**). AUC is a threshold-independent indicator of the model's capacity to distinguish between the two outcome classes.

The outcome shows that while the model's discriminatory performance is not very powerful, it can differentiate between the two classes to some extent. The ROC curve seems to be not very close to 1 which means that it is fine, but not very good. (the further from the vertical line the better). This outcome is crucial for understanding the model since the comparatively high classification accuracy shouldn't be taken as proof of strong discriminatory ability on its own.

Therefore, the ROC graph is a supplement to the confusion matrix and the accuracy-related measurements. Whereas the confusion matrix assesses the predictions based on the chosen 10%

threshold, the ROC graph assesses the behaviour of the model in classification for different thresholds.

4.2.2. Random Forest

After the development and analysis of the Logistic Regression model, the Random Forest model was used to find out whether a non-linear machine learning algorithm would offer better predictive performance than the logistic regression model in terms of classifying the outcome of loans. The reason behind the choice of the Random Forest algorithm was that it is able to capture complicated and non-linear relationships between predictor and outcome variables.

The Random Forest model was generated from the processed HMDA data set discussed in Section 4.1. Consequently, the selection of features, data cleansing, handling of missing values, formation of the target and protected attributes, balancing of classes using under-sampling technique, and splitting of the data set were all done prior to training the Random Forest model.

To predict the binary loan outcome, the Random Forest model was trained using the explanatory factors that were kept throughout the preprocessing phase. Random Forest may identify non-linear correlations between the predictor and target variables, in contrast to a linear classification model. In order to find more intricate patterns in the data, the model combines the predictions of several decision trees.

The training data was used to fit the model, and the test data was then used to assess it. The Random Forest model was evaluated using the same general framework as Logistic Regression. This made it possible to compare the models using reliable performance metrics.

Using the test dataset, the Random Forest model generated the following confusion matrix:

	Application approved but not accepted	Loan originated
Application approved but not accepted	20	28
Loan originated	2984	44148

Table 4.4: Confusion Matrix for Random Forest Model on the Test Dataset

According to the confusion matrix (**Table 4.4**), 20 data from the first outcome category and 44,148 observations from the loan-originated category were correctly identified by the Random Forest model. Nevertheless, 2,984 observations that belonged to the loan-originated category were classified into the alternative result, whereas 28 observations were mistakenly put into the loan-originated category.

Performance Measure	Logistic Regression
Accuracy	93.62%
Error Rate	6.38%
False Positive Rate	0.0006%
AUC	0.569

Table 4.5: *Performance of the Logistic Regression Model on the Test Dataset*

The accuracy attained by the Random Forest algorithm was 93.62%, which is more accurate than the accuracy attained by the logistic regression model. This model generated an error rate of 6.38%. Another performance measure that this algorithm has is the false positive rate of 0.0006% (**Table 4.5**).

The confusion matrix offers a crucial warning that the high accuracy should be read cautiously, even though these results seem favorable <<when accuracy and false positive rate are considered alone. Only a very tiny number of observations were classified into the alternative category by the model, whereas a relatively large number of observations were classified into the loan-originated group. As a result, the Random Forest model's high accuracy does not always mean that it offers a fair or usable classification of both outcome classes.

The ROC curve was also used to evaluate the discriminatory ability of the Random Forest classifier for various thresholds.

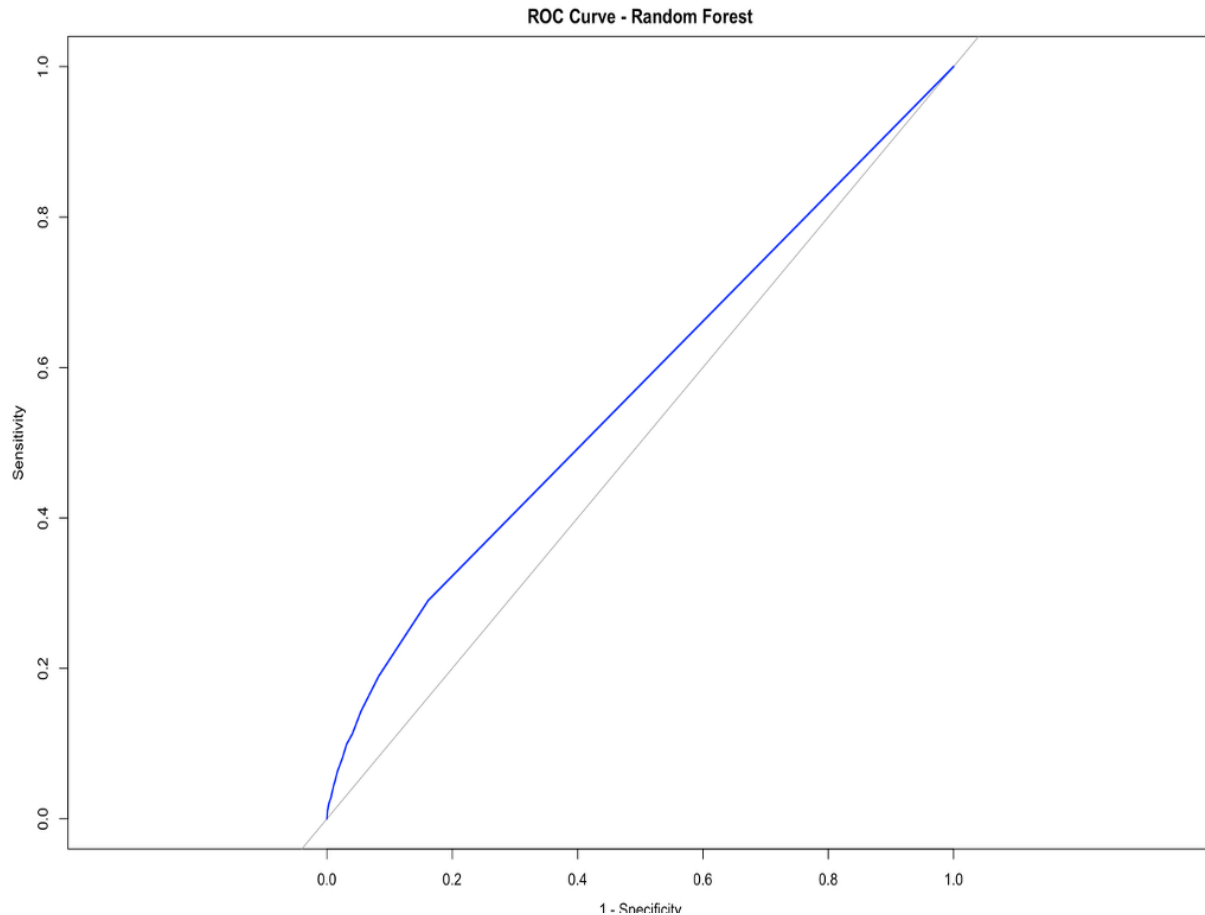


Figure 9: ROC Curve for the Random Forest Model

Despite obtaining a very high accuracy level of 93.62%, the AUC of the model is approximately 0.569. This indicates why accuracy is not adequate to judge the performance of the model. The confusion matrix also suggests that there are many predictions clustered under one category.

This clearly means that the classification performance of the Random Forest Model is not necessarily better due to high accuracy level.

The Random Forest algorithm is a type of ensemble learning which uses several Decision Trees in order to generate the result. Hence, a good understanding of the working of the Decision Tree

algorithm will be a good starting point to understand the Random Forest model used in this research paper. In order to understand the working of Random Forest, a decision tree was taken into consideration first.

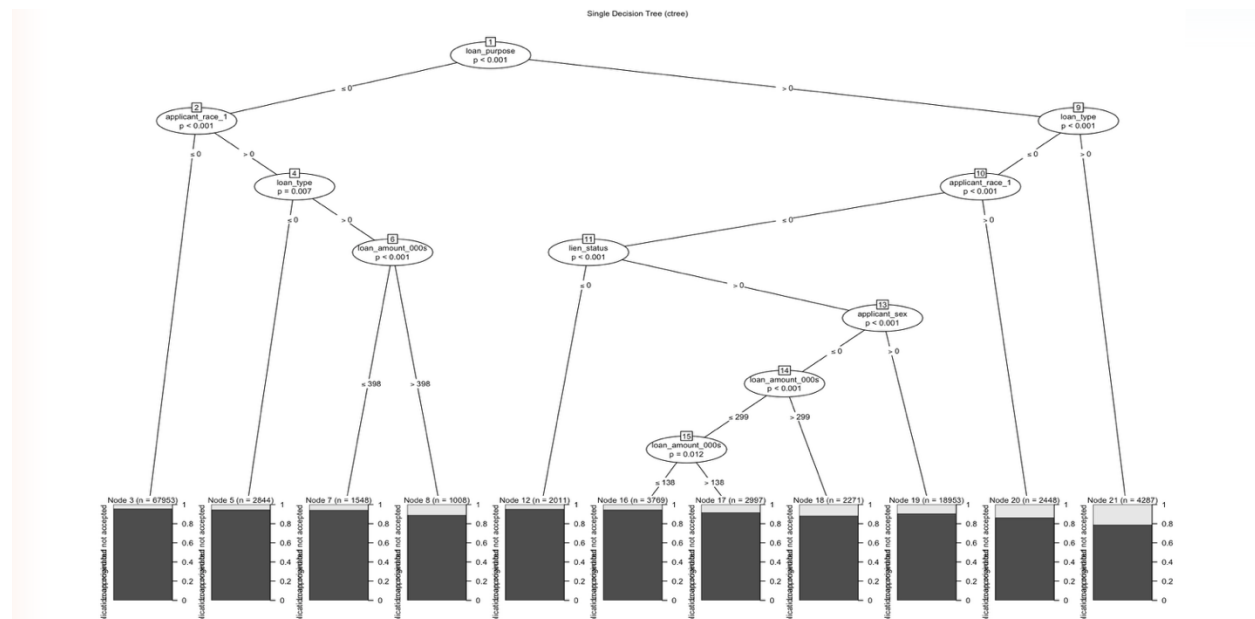


Figure 10: Decision Tree Structure for the Classification of Loan Application Outcomes

As indicated in Figure 10 above, the Decision Tree is rooted on loan purpose. From this, we can conclude that, in the framework of the fitted tree, the loan purpose variable is responsible for the first division of the applications into outcome categories. Other branches include variables such as race of the applicant, type of loan, and loan amount. This is an indication that, based on the loan purpose, variables such as those concerning the applicant and loan are used to further classify the outcomes.

The terminal nodes of the Decision Tree consist of the last groups of decisions made during the classification process. Bars shown in each terminal node reflect the proportion of observations which belong to each of the output classes. Many of the terminal nodes consist mostly of non-

accepted applications. This is due to the class imbalance detected during the data pre-processing phase and illustrates the impact of the target variable distribution on the classification performance of the single Decision Tree.

However, despite having a relatively clear illustration of the classification process, one decision tree model can be prone to changes in the structure of the tree depending on the specific observations and data split that occur in the process of building the model. This means that any alterations in the training dataset can lead to alterations in the tree's structure and its final outcomes.

For this purpose, the research uses a Random Forest, which is an extension of the Decision Tree method based on a combination of many trees. Random Forest seeks to increase prediction stability and lower the danger of overfitting associated with a single Decision Tree by combining the predictions of several trees.

But the application of ensemble methods alone cannot solve any problems with potential unfairness. Despite being better for predictions than Random Forest, these predictions need to be analyzed against the applicants' groups that were analyzed in the present research. That is why Random Forest models are assessed not only based on their ability to predict well, but also based on classification results and fairness issues.

4.2.3. Neural Network

After building the Logistic Regression and Random Forest models, another model – the Neural Network model – was built to test whether a model that is able to capture complex and non-linear relationships could offer a different means of predicting the loan outcome.

Neural networks can be applied in modelling complex relationships between predictors and

outcomes and hence were considered as part of models to compare to the traditional approaches employed in this analysis.

The Neural Network was trained on the explanatory variables that had survived the preprocessing process. The aim of the model was to learn the connection between the predictors and the binary response variable. As opposed to Logistic Regression where there is an assumption of a linear relationship between the predictors and the log-odds of the response variable, a Neural Network can capture the complicated and non-linear connections between the predictors and the response variable.

The model was fit on the training data and thereafter applied to predict on the test set. The predictions made were assessed through a confusion matrix as well as different performance metrics. The same evaluation approach was used across the different machine learning algorithms to enable the comparison of their results.

The output produced by the neural network was measured through the confusion matrix for the test data set. Table 4.6 shows the confusion matrix that was obtained.

	Application approved but not accepted	Loan originated
Application approved but not accepted	3004	44176
Loan originated	0	0

Table 4.6: Confusion Matrix for the Neural Network Model on the Test Dataset

The confusion matrix demonstrates that every observation was categorised into the same projected outcome category by the neural network (**Table 4.6**). In particular, no observations were assigned to the alternative predicted class, whereas 3,004 observations from the first actual outcome category and 44,176 observations from the second actual outcome category were assigned to the same predicted category.

This finding is significant because it shows that a large percentage of correctly categorised observations does not always mean that the model has successfully learned to differentiate between the two outcome classes. Rather, all of the model's predictions fall into a single class.

All the performance metrics achieved using the Neural Network are shown in **Table 4.7** below.

Performance Measure	Logistic Regression
Accuracy	6.37%
Error Rate	93.63%
False Positive Rate	1%
AUC	0.6703

Table 4.7: Performance of the Neural Network Model on the Test Dataset

The Neural Network showed 6.37% accuracy with 93.63% error. This poor performance is in agreement with the confusion matrix, according to which the model could not classify any observations correctly in both outcome groups.

It was claimed that the false positive rate for the model was 1%. Nevertheless, it should be pointed out that the use of this parameter should be analyzed in connection with the confusion matrix since the model made its predictions only for one class of outcomes. In such a case, the single error-based metric does not give a comprehensive idea of model performance.

It means that the multiple evaluation of the Neural Network with different performance metrics should be used. In particular, it is necessary to take into account the confusion matrix as it gives some additional information about model predictions.

Another test used on the ROC curve is to determine whether or not the Neural Network can differentiate between the two classes at varying classification cutoff points.

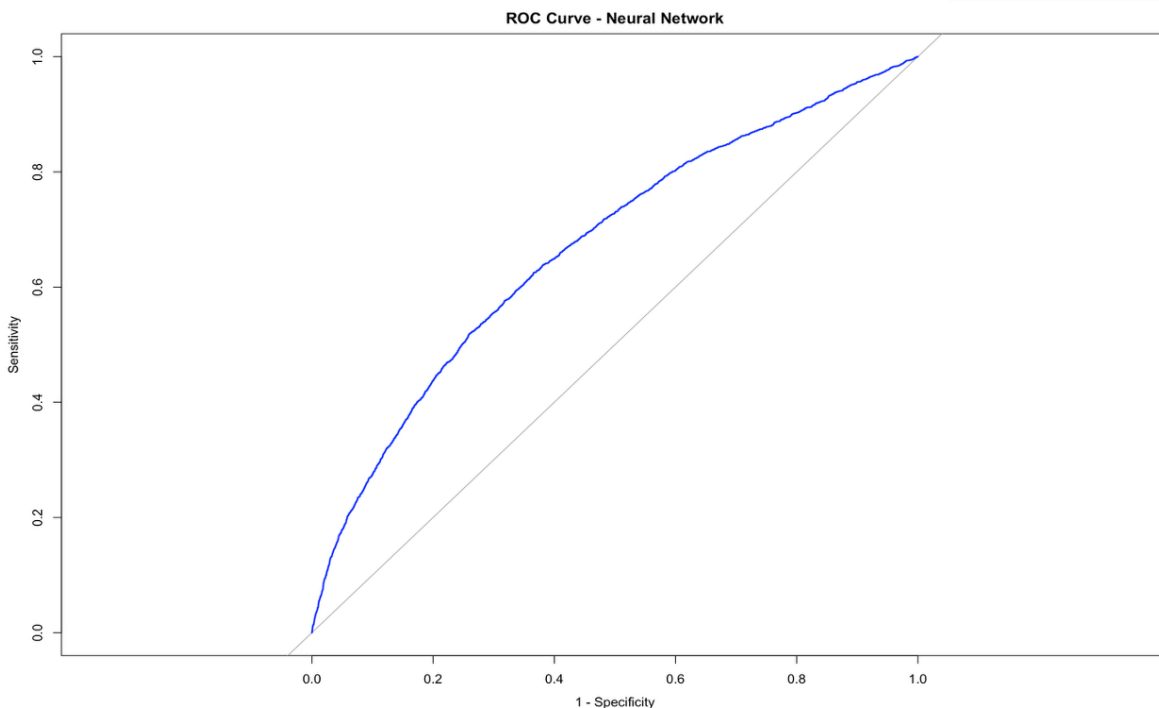


Figure 11: ROC Curve for the Neural Network Model

The Neural Network yielded the value of AUC equal to 0.6703 approximately (**Figure 11**).

Despite the fact that the value of AUC is greater than that of AUC for the Random Forest, the analysis of the confusion matrix indicates that the chosen classification method led to the situation where all test data points were classified into one category. Thus, the value of AUC cannot be used on its own when assessing whether the model yields practically valuable classification.

The discrepancy between the AUC result and classification results illustrates the importance of considering multiple angles of the model analysis. In other words, a model can have some capacity to order or separate observations according to the probabilities, but at the same time it can generate inadequate categorical results.

From the confusion matrix, it is evident that the Neural Network did not create any balanced predictions for the two outcomes. This means that the model generated predictions only for one category for all test cases. In such a way, the model was not able to make predictions for cases in the other outcome category in the classification process in this study.

Such performance led to an accuracy of 6.37% and an error rate of 93.63%. So, in spite of the fact that Neural Networks are suitable models for modelling non-linear relationships, the implementation of this model in the current study did not show good classification results.

This situation shows why accuracy and other performance indicators should be analyzed in conjunction with the confusion matrix, since the latter provides clear proof whether the model makes predictions for different outcome categories or not. While the ROC/AUC analysis offers

information regarding discriminating across thresholds, the confusion matrix directly demonstrates if the model is truly making predictions across the various outcome categories.

In summary, the performance of the Neural Network in terms of classification was not satisfactory with regard to the test data used in this study. These findings suggest that even though the Neural Network has the capability of modeling complex and non-linear relationships, it may not necessarily lead to satisfactory classification performance with respect to the dataset and methodology applied in this study. The findings have shown the importance of including various metrics in model evaluation such as the confusion matrix, accuracy, error rate, false positive rate, and the area under curve.

The Neural Network is therefore evaluated along with the Logistic Regression and the Random Forest models in the coming section of the model comparison.

4.3. Shapley Value Computation

The Shapley values technique was applied to increase the explainability of the machine learning models by determining the individual variable contribution to the model output. For this study, the mean absolute SHAP value has been chosen as an indicator of the importance of a particular feature.

The analysis was carried out for the Logistic Regression and Random Forest models. Applicant race and sex are of special interest since their contribution to the model outputs will be further analyzed in terms of fairness. The results of each model are provided below.

4.3.1. Concept and Theory of Shapley Values

The Shapley values were employed in this research to determine the contribution of each predictor to the forecasts produced by the machine learning algorithms. For each sample, the Shapley value denotes the contribution of a specific variable to the forecast made by the model. Hence, variables with higher Shapley values are more influential on the forecast made by the model, while those with lower Shapley values have relatively less influence.

Because the Shapley value can be either positive or negative depending on whether the variable increases or reduces the prediction made by the model, the absolute Shapley value is employed when quantifying the overall contribution of the variable to the prediction made by the model. The mean absolute SHAP value is obtained from the average of the absolute Shapley values for all the samples.

For this study, mean absolute SHAP values were used in ranking the features in terms of importance in both Logistic Regression and Random Forest models. A higher mean absolute SHAP value means that the feature has more average impact on the predictions of the model, while a lower mean absolute SHAP value implies that the feature has less average impact on the model.

This approach of ranking features using mean absolute SHAP values helps in quantifying the importance of each feature and enables comparisons between contributions of different features. This is very important for this study since it will enable the comparison between the impact of characteristics of applicants and other loan-based features. The Global Shapley Values from this analysis are discussed in the following subsection.

4.3.1.1. *Shapley Values for Logistic Regression*

Apart from measuring the effectiveness of prediction, the study analyzed the factors that influenced the predictions made by the Logistic Regression model the most. Shapley values were employed for this analysis. Mean Absolute SHAP value is the measure that has been used to capture the magnitude of the effect of each factor on the model prediction, without taking into account its direction.

The results of Mean Absolute SHAP value of the variables analyzed by the Logistic Regression model are presented in Table 4.8 below.

With a mean absolute SHAP value of 0.02850, the results show that loan purpose contributed the most on average to the model predictions. With a mean absolute SHAP value of 0.00784, loan type was the second most significant factor. The mean absolute SHAP value for the demographic data was 0.00460 for applicant race and 0.00229 for applicant sex.

The mean absolute SHAP values were lower for loan amount, lien status, and applicant income. Overall, the findings show that among the most significant factors influencing the Logistic Regression predictions were loan-related traits. In addition, the inclusion of applicant sex and race among the significant variables is pertinent to the study's goal of fairness.

Feature	Mean_abs_shap
Loan_purpose	0.0285
Loan_type	0.00784
Applicant_race_1	0.00460
Applicant_sex	0.00229
Loan_amount_000s	0.00174
Lien_status	0.00155
Applicant_income_000s	0.000613

Table 4.8: Mean Absolute SHAP Values for the Logistic Regression Model

For the Logistic Regression model, which offered the most insightful interpretation of the factors influencing loan result predictions, the Shapley value analysis was very crucial. The findings indicated that while applicant race and sex also influenced the model's predictions, loan type and purpose were the most significant factors.

Because of its interpretability, the Logistic Regression model was very helpful for this study's goals because it allowed for the identification and subsequent examination of the role of demographic factors in the fairness analysis. Therefore, before determining whether these affects

are linked to differences across protected groups, the Shapley values serve as a foundation for understanding which variables affect the model.

4.3.1.2. Shapley Values for Random Forest

SHAP values were also used to determine the comparative contribution of the predictors to the forecasts produced using the Random Forest algorithm. Mean absolute SHAP values were utilized to measure the contribution of each predictor to the Random Forest forecasts.

Results of the SHAP analysis of the Random Forest algorithm are shown in Table 4.9.

Feature	Mean abs shap
Loan_amount_000s	0.00248
Applicant_income_00s	0.00200
Loan_type	0.00166
Loan_purpose	0.00131
Applicant_race_1	0.00104
Applicant_sex	0.000503
Lien_status	0.000489

Table 4.9: Mean Absolute SHAP Values for the Random Forest Model

The findings indicate that the variable loan amount was the most important one in predicting outcomes using the Random Forest method, with a mean absolute SHAP value of 0.00248.

Applicant income was the second most important variable, with a mean absolute SHAP value of

0.00200. The third and fourth important variables were loan type and loan purpose, with mean absolute SHAP values of 0.00166 and 0.00131, respectively.

Race and sex of the applicant were also important predictors in the model, although they had lower mean absolute SHAP values compared to loan amount, applicant income, loan type, and loan purpose. Race had a mean absolute SHAP value of 0.00104, while the mean absolute SHAP value for applicant sex was 0.000503.

Overall, the Random Forest model gave financial and loan-related traits more weight than demographic factors. The biggest contributions came from loan amount and applicant income, indicating that these factors were more important in the model's predictions. Even though the mean absolute SHAP values for applicant race and sex were lower, their inclusion in the contributing variables is still important for the fairness study. This suggests that although demographic factors were taken into account during the model's decision-making process, they were not the main factors influencing the Random Forest forecasts. Therefore, rather than being regarded as proof of discrimination based only on their SHAP values, their potential impact should be investigated further through the fairness metrics.

4.3.2. Global Shapley Values (GSVs)

To determine the overall contribution of each feature to the model, Global Shapley Values (GSVs) were computed by adding together the individual SHAP values from all of the observations. The global contribution of each variable in this study was represented by the mean absolute SHAP value. The variables that had the biggest overall impact on the model predictions were determined by ranking the features using the obtained values.

Table 4.8 displays the Global Shapley Values for the **Logistic Regression Model**. According to the findings, `Loan_purpose` and `Loan_type` made the most worldwide contributions. The variables with comparatively higher contributions among the applicant characteristics were `Applicant_race_1` and `Applicant_sex`. The mean absolute SHAP values for the remaining variables were significantly lower.

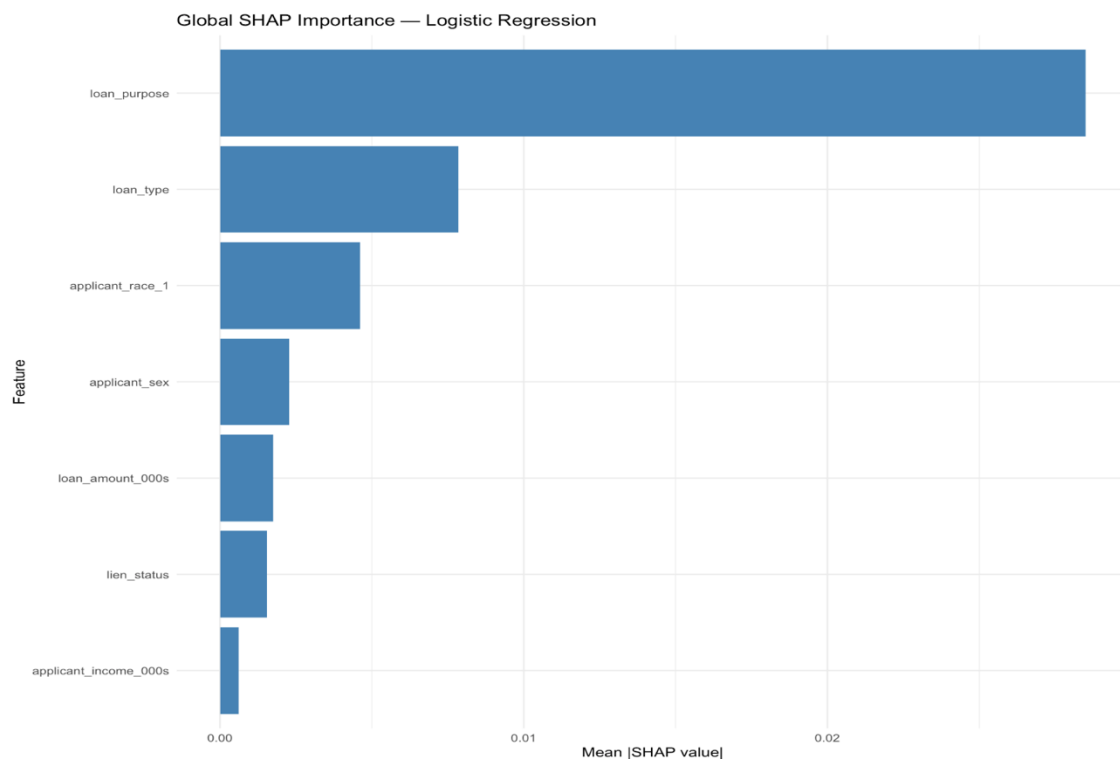


Figure 12: Global SHAP Importance for the Logistic Regression Model

A horizontal bar chart should ideally be used to visually represent the SHAP results. The features ought to be arranged in descending order of mean absolute SHAP value. Compared to the numerical table alone, this makes it easier to compare the relative contributions of the factors as shown in **Figure 12**.

According to the SHAP analysis, the model's predictions are influenced by demographic factors. However, the existence of sex and race as significant factors shouldn't be taken as proof of discriminatory behaviour on its own. Instead, it shows that these factors influence the model's predictions and, as a result, merit more investigation from a fairness standpoint. As a result, the fairness assessment that follows in this chapter takes into account the examination of demographic discrepancies independently.

SHAP analysis demonstrates that applicant race and sex have an impact on the model's projections, even if their significance is less than that of loan-related factors, and reveals that the logistic regression model is mostly driven by loan characteristics rather than applicant financial indicators. This demonstrates that there is no concrete proof of prejudice. This indicates that the model uses these hidden characteristics to forecast loan outcomes.

To conclude, this analysis indicated that applicant race and sex contribute to loan outcome predictions, suggesting the presence of potential bias patterns (not direct evidence).

Figure 4.9 displays the Global Shapley Values for the **Random Forest Model**. The Random Forest model ranked Loan_amount_000s as the most significant feature, followed by Applicant_income_000s, in contrast to Logistic Regression. Applicant_race_1 and Applicant_sex had smaller worldwide contributions, followed by Loan_type and Loan_purpose.

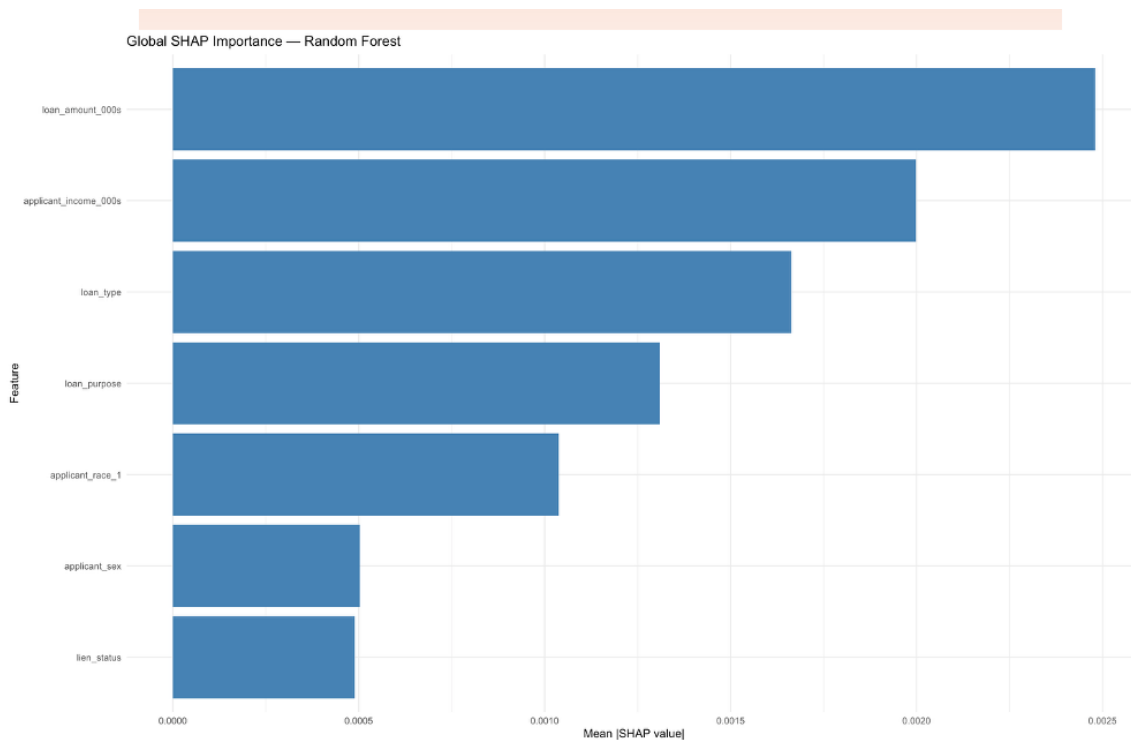


Figure 12: Global SHAP Importance for the Logistic Regression Model

The variables that contributed most significantly to predictions varied depending on the modelling approach, as seen by the comparison of the two models. While Random Forest provided more weight to loan amount and applicant income, Logistic Regression gave more weight to loan type and purpose. Despite these variations, both models' predictions were influenced by the race and sex of the applicants. In order to ascertain whether there are variations in model outcomes among the protected groups, these results are taken into account in the fairness analysis that follows.

4.4. Fairness Assessment

After analyzing model performance and contributions from various features, the next step was to analyze model fairness in order to check whether models generate different outputs for different groups of applicants. In particular, race was considered as the feature of interest since the

research objectives were related to it and it was included into model predictions based on the Shapley values analysis.

In order to conduct model fairness analysis, it is important to detect differences in the outputs and error rates for different racial groups. First, model fairness analysis is performed by taking into account race as a possible source of bias, and then it is expanded into subgroup analysis.

4.4.1 Race as an Indicator of Potential Bias

Race was examined as a sign of possible bias in the model's judgements given the applicant race's contribution to the Logistic Regression predictions found through the Shapley value study. The **approval rate, false positive rate (FPR), and false negative rate (FNR)** variations between the two racial groups taken into consideration in the study were the main focus of the fairness assessment.

Applicant Race	Approval Rate	False Positive Rate	False Negative Rate
Group 0 (White)	8.9%	8.1%	78.5%
Group 1 (African American)	35.5%	33.0%	42.0%

Table 4.10: Fairness Measures by Applicant race

The two racial groups differ significantly, according to the findings (**Table 4.10**). White applicants had an approval rate of 8.9%, while African American applicants had an approval rate of 35.5%, a difference of 26.6 percentage points. This shows that a significantly higher

percentage of African American applicants were categorised by the algorithm as having a good outcome.

The false positive rate also showed a significant variation. There was a 24.9% point difference in the FPR between African American applicants (33.0%) and White applicants (8.1%). This suggests that African American applicants had a significantly greater risk of erroneous positive classifications.

The misleading negative rate displayed the reverse trend. White applicants had a FNR of 78.5%, whereas African American applicants had a FNR of 42.0%, a difference of 36.5 percentage points. African American applicants therefore had a significantly lower FNR than White applicants, despite having a significantly higher approval rate and FPR.

When combined, these findings show that the Logistic Regression model does not yield results that are comparable for the two racial groups. The variations in FPR, FNR, and approval rate point to possible racial bias tendencies in the model's predictions. However, as inequalities may also reflect variations in the underlying characteristics of the applicant groups, these differences by themselves do not prove that the model is discriminatory.

The results hence motivate a need to investigate further within subgroups. The following section will test if the race gap found above persists within different subgroups with emphasis on the risk of Simpson's Paradox.

4.5. Bias Mitigation

After the fairness assessment, bias mitigation technique was conducted to minimize the influence of disparity in terms of composition of the racial groups. The selected method for achieving that objective was Propensity Score Matching (PSM). This method will be used to achieve greater

comparability between White and Black/African-American applicants on the basis of the characteristics of the applicants and loans. There were three steps involved in the bias mitigation process; propensity score matching, balance test, and fairness measures after matching.

4.5.1. Propensity Score Matching (PSM)

Propensity Score Matching was used to find applicants from both racial groups having similar observed characteristics. In this analysis, race of the applicant was the indicator of the protected group, where white applicants formed the control group while black or African American applicants comprised the treated group.

Propensity score was calculated by using the logistic regression model which considered the variables such as income of the applicant, loan amount, loan purpose, loan type, applicant's sex and lien status. These variables were considered to include the observable difference between the racial groups influencing the applicant groups. The matching was carried out by applying the nearest neighbor matching with 1:1 ratio based on the propensity score estimation.

Matching Variable	Before PSM Absolute SMD	After PSM Absolute SMD	Balance After PSM
Propensity score distance	0.6020	0.0001	✓ Balanced
Applicant income	0.2987	0.0051	✓ Balanced
Loan type	0.4569	0.0283	✓ Balanced
Applicant sex	0.3376	0.0144	✓ Balanced
Loan purpose	0.0771	0.0158	✓ Balanced
Loan amount	0.0250	0.0575	✓ Balanced
Lien status	0.0674	0.0214	✓ Balanced

Table 4.11: Comparison of covariate balance before and after propensity score matching (PSM).

Table 4.11 illustrates how Propensity Score Matching (PSM) improved covariate balancing. Prior to matching, there were comparatively significant disparities between the ethnic groups for factors like loan type, applicant sex, and applicant income. Following matching, all variables' absolute standardized mean differences (SMDs) dropped below the 0.10 cutoff, suggesting a satisfactory level of balance. This implies that in terms of the observed loan and applicant characteristics, the matched White and Black or African American application groups became significantly more comparable.

The goal of matching was to form pairs of applicants from both racial groups having similar observed characteristics. Thus, further fair comparison could be conducted for this sample instead of the original groups which differed greatly in size and composition.

4.5.2. Construction of a Balanced Dataset

To create the balanced dataset, the matched observations were extracted using the PSM process. Each application in the protected group was matched to one applicant from the control group, resulting in an equal number of White and Black or African American candidates in the final dataset.

For the fairness study that followed, the balanced dataset was kept. The matched dataset offered a more suitable foundation for investigating whether racial disparities in model outcomes remained after adjusting for observable differences between the groups by equalizing the representation of the two racial groups and enhancing the comparability of the observed characteristics.

4.5.3. Fairness Evaluation After Mitigation

To evaluate the impact of the mitigation process, the fairness measures were recalculated after the balanced dataset was created. To guarantee comparability between the pre- and post-

mitigation studies, the same fairness metrics from the original assessment were kept.

Approval rates, false positive rates (FPR), and false negative rates (FNR) for the two racial groups were among these metrics.

The equivalent measures acquired prior to matching were compared with the post-mitigation measures. A decrease in the racial group disparities would suggest that variations in the groups' makeup contributed to some of the observed gap. Chapter 5 presents the comparison's numerical results.

4.6. Statistical Testing

To determine whether loan outcomes varied significantly between racial groups, a statistical test was performed. The association between the reported loan outcome and the race of the applicant was evaluated using a Chi-square test of independence. In accordance with the protected-group concept applied throughout the fairness analysis, White and Black or African American applicants represented applicant race.

The alternative hypothesis suggested that there was a statistically significant correlation between the two variables, whereas the null hypothesis suggested that applicant race and loan outcome were independent. At the 5% level ($\alpha = 0.05$), statistical significance was assessed. In order to confirm that the Chi-square test's assumptions were met, the expected frequencies were also looked at.

By evaluating whether variations in loan outcomes among ethnic groups were statistically significant, the statistical test supplemented the model-based fairness assessment. Statistical significance by itself, however, was not seen as proof of discrimination. Rather, the test was evaluated in conjunction with bias mitigation analysis, Shapley value analysis, and subgroup fairness measures. Chapter 5 presents the statistical test's numerical results.

5. Chapter 5: Results

In this chapter, the empirical results of this research based on the statistical, predictive, explainability and fairness analyses carried out in this research will be discussed. The empirical results will be presented according to the methodology outlined in Chapter 4 of this research, with emphasis on the differences in the models' performance and possible disparities among different racial groups.

First of all, this chapter will start with the results of statistical testing and the observed initial patterns in the data, after which the predictive model performance will be evaluated using the ROC-AUC metric. Then, the explainability analysis will be presented with Shapley values used to determine the variables that make the biggest contribution to the model predictions, paying special attention to the effect of the applicant race. Next, the subgroup analysis will be carried out to assess the differences in the performance of the models among different racial groups. Finally, the Propensity Score Matching results will be analyzed to determine whether there are any differences left after the comparable groups of applicants have been formed.

5.1. Statistical Testing and Initial Data Patterns

The initial analysis looked for a statistically significant correlation between the recorded loan outcome and the race of the applicant. There were disparities between White and Black or African American applicants in the results distribution. Outcome 0 (Loan originated) was assigned to 93.94% of White applicants, while outcome 1 (Application approved but not accepted) was assigned to 6.06% (**Table 5.1**). The equivalent percentages among Black and African American applicants were 10.06% and 89.94%, respectively. Therefore, among Black or

African American candidates, Outcome 1 was almost 4.01 percentage points more common (10.06 – 6.06).

Applicant Race	Outcome 0	Outcome 1
White	93.94%	6.06%
Black or African American	89.94%	10.06%

Table 5.1: Loan Outcome Distribution by Applicant Race

To determine if the observed differences were statistically significant, a Pearson Chi-square test was performed. With $p < 2.2 \times 10^{-11}$, the test yielded a Chi-square result of $\chi^2(1) = 302.55$. The null hypothesis of independence was rejected since the p-value was significantly below the 5% significance level. This suggests a statistically significant correlation between the recorded loan outcome and the race of the applicant.

The Chi-square test's validity was supported by the predicted frequencies for every cell being significantly higher than five. The statistically substantial correlation, however, should not be taken as proof of discriminatory lending practices on its own. Instead, it proves that racial disparities exist in the observed loan results and serves as a foundation for the model explainability and fairness evaluations that follow.

5.2. Model Performance Evaluation using ROC-AUC

The Area Under the Receiver Operating Characteristic Curve (ROC-AUC) was used to assess the models' prediction performance. The ability of each model to differentiate between the two loan outcome classes across various classification thresholds was evaluated using this metric.

With an AUC of roughly 0.6344, the Logistic Regression model demonstrated a moderate capacity to differentiate between the two outcome classes. Despite its comparatively high classification accuracy, the Random Forest model yielded a lower AUC of roughly 0.569, indicating inferior discriminatory performance. At roughly 0.6703, the Neural Network had the greatest AUC (**Table 5.2**).

Model	ROC-AUC	Overall Assessment
Logistic Regression	0.6344	Moderate discrimination
Random Forest	0.5690	Weak discrimination
Neural Network	0.6703	Highest AUC

Note: Higher ROC-AUC indicates stronger ability to distinguish between the two loan outcome classes.

Table 5.2: ROC AUC Performance Comparison

Despite having the best ROC-AUC, the Neural Network's confusion matrix revealed that it collapsed into a single prediction class, leading to poor classification performance in real-world scenarios. Similar to this, the Random Forest's high accuracy was deceptive due to the high concentration of its predictions in a single class. On the other hand, Logistic Regression maintained an AUC of 0.6344 when generating predictions for both outcome classes. As a result, the maximum ROC-AUC or accuracy were not the only factors considered while choosing a model. Because logistic regression allowed for the examination of differences between applicant groups and produced more relevant categorization behavior, it was chosen as the best model for the ensuing fairness and explainability analyses. This is in line with the

analysis's model comparison, which found that Logistic Regression was the best model for identifying significant disparities in fairness.

5.3. Explainability Results (Shapley Values)

After predictive performance was assessed, Shapley values were analyzed to determine the contribution of each variable to the model's predictions. Random Forest Shapley values were used for comparison, and the analysis mostly concentrated on Logistic Regression, which was chosen as the best model for the fairness assessment.

Three evaluations of the explainability outcomes were conducted. To determine which factors contributed most to model predictions, the overall feature importance rankings were first looked at. Second, as the protected trait of interest, the applicant's race was given special consideration. Lastly, the Shapley results of Random Forest and Logistic Regression were compared to see if the various modelling techniques showed comparable patterns of feature effect.

5.3.1. Feature Importance Rankings

The Logistic Regression model's Shapley value analysis showed distinct variations in the predictor variables' relative contributions. The most significant variable in the model was loan purpose, which had the highest mean absolute Shapley value of 0.0285. The loan type (0.00784) and applicant race (0.00460) came next. The predictions were also influenced by applicant sex, which had a mean absolute Shapley value of 0.00229 (**Table 5.3**).

The loan amount (0.00174), applicant income (0.000613), and lien status (0.00155) all displayed relatively small contributions (**Table 5.3**). Overall, the findings show that the Logistic Regression predictions were most strongly influenced by loan-related factors. Nonetheless, the

third most significant factor was applicant race, indicating that the model's predictions were also influenced by a protected trait.

Rank	Feature	Mean Absolute SHAP
1	Loan purpose	0.02850
2	Loan type	0.00784
3	Applicant race	0.00460
4	Applicant sex	0.00229
5	Loan amount	0.00174
6	Lien status	0.00155
7	Applicant income	0.000613

Note: Higher mean absolute SHAP values indicate greater overall contribution to model predictions.

Table 5.3: Logistic Regression Feature Importance Based on Mean Absolute Shapley Values

Because applicant race is one of the three most important factors, the ranking is very important to the fairness analysis. However, the inclusion of race in the feature importance ranking does not prove discriminatory behaviour on its own. Rather, it suggests that race influenced the model's predictions, which calls for additional research using the racial and subgroup fairness analyses shown in the sections that follow.

5.3.2. Race Contribution Analysis

To determine how applicant race affected the model's predictions, a closer look at the Shapley value analysis was conducted. After loan type and loan purpose, applicant race ranked third in the Logistic Regression model with a mean absolute Shapley value of 0.00460.

The comparatively high ranking of applicant race in the feature ranking suggests that differences in the predictions produced by the Logistic Regression model were influenced by race. It was more than the contributions of applicant sex, loan amount, lien status, and applicant income, even if it was much less than that of loan purpose.

This result is significant from a fairness standpoint as the protected trait this study examined is applicant race. Its non-zero Shapley contribution implies that racial group membership has some impact on the model's predictions. However, the contribution's amount rather than its direction is measured by the mean absolute Shapley value. Therefore, this conclusion does not directly demonstrate discrimination or determine whether race raised or decreased the likelihood of a specific loan outcome.

Rather, the outcome highlights applicant race as a characteristic that needs more fairness evaluation. Therefore, in the analysis that follows, the Shapley evidence is taken into account together with the racial subgroup variations in approval rates, false positive rates, and false negative rates.

5.3.3. Cross Model Comparison

The results of the Shapley value varied across the Random Forest and Logistic Regression models, suggesting that the modelling strategy affected the relative significance of the predictor variables. The most significant factor in the Logistic Regression model was loan purpose, which

was followed by loan type and applicant race. With a mean absolute Shapley value of 0.00460, the applicant race came in third (**Table 5.4**).

The Random Forest model, on the other hand, gave financial and loan-related traits more weight. The loan amount (0.00248) had the greatest mean absolute Shapley value, followed by loan type (0.00166), loan purpose (0.00131), and applicant income (0.00200). With a mean absolute Shapley value of 0.00104, the applicant race came in fifth (**Table 5.4**).

Feature	LR SHAP	LR Rank	RF SHAP	RF Rank
Loan purpose	0.02850	1	0.00131	4
Loan type	0.00784	2	0.00166	3
Applicant race	0.00460	3	0.00104	5
Applicant sex	0.00229	4	0.000503	6
Loan amount	0.00174	5	0.00248	1
Lien status	0.00155	6	0.000489	7
Applicant income	0.000613	7	0.00200	2

Note: SHAP values represent mean absolute Shapley values. Higher values indicate greater overall contribution.

Table 5.4: Comparison of Feature Importance across Logistic Regression and Random Forest

The comparison reveals significant variations in the relative significance of a number of variables. In the Random Forest model, loan amount and applicant income were more significant than loan purpose in the Logistic Regression model. The most significant finding for the fairness analysis was that applicant race ranked third in Logistic Regression as opposed to fifth in Random Forest. However, in both models, race had a non-zero influence.

These results show that model dependence affects feature importance. Given the classification behavior found in the model performance analysis, the Random Forest model's lower ranking of race does not necessarily mean that it is more equitable. Additionally, it is important to use caution when interpreting variations in the absolute magnitude of SHAP values among various model types. Because of this, the cross-model comparison benefits more from the relative feature ranks and the inclusion of race as a contributing variable than from the numerical SHAP values alone.

As a result, rather than being utilized separately to make judgements about model fairness, the Shapley results were taken into account in conjunction with the subgroup fairness scores.

5.4. Subgroup Analysis

The purpose of the subgroup study was to determine whether the racial disparities found in the sample as a whole remained when applicants were divided based on significant loan attributes. Since both loan type and loan purpose were found to be significant predictors in the Shapley value analysis, they were used for this research. Within each grouping, the analysis contrasted the reported loan outcome distribution between White and Black or African American applicants.

Significant discrepancies between the two racial groups were found in the Logistic Regression model's initial fairness study. Black or African American applicants had an approval rate of 35.5%, while white applicants had an approval rate of 8.9%. Additionally, variations were noted in the false negative rate (FNR) at 78.5% and 42.0%, respectively, and the false positive rate (FPR) at 8.1% and 33.0%, respectively (**Table 5.5**).

Subgroup	White (%)	Black or African American (%)	Difference (pp)	χ^2	p-value
Loan type 0	5.77	9.46	3.69	171.34	$< 2.2 \times 10^{-16}$
Loan type 1	7.54	11.03	3.49	62.78	2.31×10^{-15}
Loan purpose 0	4.17	6.58	2.41	99.13	$< 2.2 \times 10^{-16}$
Loan purpose 1	9.86	16.03	6.17	167.54	$< 2.2 \times 10^{-16}$

*Note: Percentages represent the proportion of Outcome 1 within each racial group and subgroup.
pp = percentage points.*

Table 5.5: Outcome 1 by Race within Loan Type and Loan Purpose Subgroups

The subgroup results demonstrate that even after controlling for loan type and loan purpose independently, the racial difference persisted. Outcome 1 continued to be more common among Black or African American applicants than among White applicants in each of the four subgroups, and each correlation continued to be statistically significant. Loan factors influenced the extent of the observed differential but did not eradicate it, as evidenced by the variation in the difference's magnitude between subgroups, which ranged from roughly 2.41 to 6.17 percentage points (**Table 5.5**).

Crucially, these results do not support Simpson's Paradox for either loan type or loan purpose. After stratification, a Simpson's Paradox pattern would require the general racial association to vanish or reverse. Rather, within each subgroup, the direction of the racial difference did not change. Thus, the findings imply that the overall racial connection found in the dataset cannot be explained by variations in the distribution of loan type and loan purpose alone.

The Propensity Score Matching methodology described in Section 5.5, which takes into account several observed loan and applicant variables concurrently to create more comparable ethnic groups, is further supported by these findings.

5.5. Bias Mitigation Results (PSM Analysis)

A significantly more equal sample of White and Black or African American candidates was generated via the Propensity Score Matching (PSM) process. The dataset showed a significant disparity in group size prior to matching, with 145,055 White applicants and 12,214 Black or African American applicants. 12,214 candidates from each racial group were kept after 1:1 nearest-neighbor matching, yielding a matched sample of 24,428 observations.

Standardized mean differences (SMDs) were used to evaluate the matching process's quality. Significant imbalances were found for a number of factors prior to matching. The propensity score distance (SMD = 0.6020), loan type (0.4569), applicant sex (0.3376), and applicant income (0.2987) showed the biggest variations (**Table 5.6**). These numbers show that before matching, there were significant differences between the two racial groups in a number of observed variables.

The discrepancies between the groups significantly decreased after PSM. The propensity score's SMD dropped from 0.6020 to 0.0001, applicant income from 0.2987 to 0.0051, loan type from 0.4569 to 0.0283, and applicant sex from 0.3376 to 0.0144. Small post-matching differences were also seen for the remaining factors (**Table 5.6**).

Covariate	SMD Before PSM	SMD After PSM
Propensity score distance	0.6020	0.0001
Applicant income	0.2987	0.0051
Loan type	0.4569	0.0283
Applicant sex	0.3376	0.0144
Loan purpose	0.0771	0.0158
Loan amount	0.0250	0.0575
Lien status	0.0674	0.0214

Note: SMD = standardized mean difference. Absolute SMD values below 0.10 indicate satisfactory covariate balance.

Table 5.6: Covariate Balance Before and After Propensity Score Matching

Importantly, the post-matching absolute SMD for all the variables considered was less than the commonly used cut-off of 0.10, indicating that the measured variables were relatively well balanced. Therefore, the results indicate that the PSM technique has been effective in creating two groups whose members were more alike regarding the loan and applicant variables considered during the matching process.

Loan outcomes were compared between the two racial groups in the matched sample after the observed factors were successfully balanced. Following matching, 7.23% of White candidates and 10.06% of Black or African American applicants experienced Outcome 1, resulting in a racial disparity of almost 2.83 percentage points. Prior to matching, the corresponding rates were almost 4.01 percentage points apart, at 6.06% and 10.06%, respectively (**Table 5.7**).

Sample	White Outcome 1 (%)	Black or African American Outcome 1 (%)	Racial Gap (pp)
Before PSM	6.06	10.06	4.01
After PSM	7.23	10.06	2.83

Note: pp = percentage points. Percentages represent the proportion of Outcome 1 within each racial group.

Table 5.7: *Racial Outcome Differences Before and After Propensity Score Matching*

As a result, the racial outcome disparity shrank by roughly 1.17 percentage points, or roughly 29% of the initial margin. This suggests that while the racial discrepancy was not completely removed, it was lessened by balancing the observed loan and application attributes.

The association between applicant race and loan outcome remained statistically significant, according to a Chi-square test on the matched sample ($\chi^2(1) = 61.69$, $p = 4.02 \times 10^{-11}$).

Therefore, even after balancing the observable factors used in the matching process, a significant racial disparity persisted.

5.6. Interpretation Findings

The findings of the research offer the proof of racial differences being observed in the loan outcomes and being relevant during the prediction modelling and the fairness tests. At the very beginning of the process, statistical testing proved that Outcome 1 is observed in 6.06% of White applicants and 10.06% of Black or African American applicants. The Chi-square test proved that there was a statistically significant relationship between the race of the applicant and the observed loan outcome, $\chi^2(1) = 302.55$, $p < 2.2 \times 10^{-16}$.

Conventional measures of model performance used in the process of predictive modelling have been demonstrated to be relevant to be evaluated critically. Though, the Neural Network had the highest value of ROC-AUC (0.6703), and the Random Forest had the highest overall accuracy, these models had an obvious flaw of the classification, which was focused on only one outcome class. Logistic Regression (ROC-AUC = 0.6344) produced much more useful predictions for both outcome classes and was used for the subsequent fair tests.

The explainability analysis shed light on the variables that influenced the predictions of models. In the case of the Logistic Regression model, loan purpose and loan type were the two most important variables, whereas applicant race came third, with an average Shapley value of 0.00460. Applicant race influenced the Random Forest model as well but was of lesser importance in this model. This means that the protected characteristic influenced the predictions of various models, showing that the influence of certain variables depended on the model.

The subgroup analysis showed that the differences in races were not caused by differences in loan type or loan purpose. Outcome 1 was still more likely to occur in the case of Black or African American applicants in each subgroup, and the association was statistically significant in all the cases. This shows that the association was consistent in all subgroups, which means that Simpson's Paradox could not be observed for loan type and loan purpose.

Lastly, the PSM analysis showed that variations in observed loan and application attributes were linked to a portion of the initial racial disparity. Covariate balance was significantly improved by matching, and all post-matching standardised mean differences were less than 0.10. After matching, the racial disparity in Outcome 1 shrank by around 29%, from roughly 4.01

percentage points to 2.83 percentage points. Nonetheless, $\chi^2(1) = 61.69$, $p = 4.02 \times 10^{-1}$ —the post-matching link between race and outcome continued to be statistically significant.

When combined, the results show that observable variations in loan and applicant attributes account for some, but not all, of the racial discrepancy found in the data. The persistence of a racial difference following subgroup analysis and PSM indicates that the observed discrepancy could not be eliminated by simply balancing the measured parameters. However, these results shouldn't be taken as proof of discriminatory lending practices or as proof that race has a causal impact. Instead, they point out a recurring statistical fairness issue that should be carefully taken into account when developing, assessing, and using credit risk models.

6. Chapter 6: Discussion

The study's key conclusions on explainability and fairness are covered in this chapter, along with how they relate to credit risk models. The discussion examines how variations in predictive performance, Shapley value contributions, subgroup outcomes, and the Propensity Score Matching analysis contribute to the overall evaluation of potential racial disparities in lending decisions, building on the empirical findings reported in Chapter 5.

The relationship between model performance and fairness, as well as the significance of explainability when assessing the relevance of protected factors such as applicant race, are given special consideration. While acknowledging that the observed statistical differences should not be construed as proof of discriminatory lending practices on their own, the chapter also examines the practical consequences of the findings for financial institutions and regulatory oversight.

6.1. Key Insights on Fairness and Explainability

The results from this study have revealed that when assessing credit risk models, predictive performance and algorithmic fairness should be considered simultaneously. Even if a particular model demonstrates good performance in terms of metrics like accuracy or ROC-AUC, it can exhibit unwanted classification behavior or deliver significantly diverse results among demographic subgroups. This problem became evident in comparing three predictive models, as differences in their overall performance did not necessarily mean that one model made better or fairer predictions.

The Logistic Regression was the best choice for conducting the fairness assessment since it made meaningful predictions in both outcome classes while demonstrating mediocre discriminatory

performance. By contrast, the Random Forest showed excellent overall accuracy while making biased predictions toward one of the classes, whereas the Neural Network had poor classification performance despite being the winner in ROC-AUC. These findings highlight the drawbacks of using models for lending applications based only on one performance metric.

Furthermore, the results of explainability analysis showed the necessity of the examining of the contribution of individual features to model predictions. "Loan purpose" and "loan type" turned out to be the most influential predictors for the Logistic Regression model. This shows that loan-related features were very significant in making the decision. But applicant race was the third most influential predictor. Race influenced prediction of the Random Forest model, but its relative significance was less than that of Logistic Regression. The inclusion of race into both models is the reason why protected features need individual investigation.

However, in the context of fairness, Shapley values should be interpreted with caution. The study's mean absolute Shapley values quantify a feature's contribution, but they do not determine whether that contribution consistently favors or disadvantages a certain racial group. As a result, racial contributions cannot be seen as proof of discrimination on their own. Because explainability considers both how the model makes predictions and whether its results vary among protected groups, it offers a more insightful evaluation when combined with group-level fairness metrics.

This interpretation was reinforced by the subgroup analysis. Examining applicants separately based on loan type and loan purpose did not eliminate racial disparities, and the direction of the difference was the same for all groupings. As a result, Simpson's Paradox did not account for the observed overall relationship for the traits under investigation. This implies that the racial

discrepancy found in the sample as a whole could not be explained by variations in these loan variables alone.

A further viewpoint on this conclusion is offered by the Propensity Score Matching study. The Outcome 1 gap shrank from roughly 4.01 to 2.83 percentage points after observable loan and application factors were adjusted between the ethnic groups. This decrease suggests that the initial discrepancy was partially explained by variations in the group composition. However, the remaining difference remained statistically significant, indicating that the found racial connection was not entirely explained by balancing the measured factors.

When combined, these results show how useful it is to evaluate credit risk models using explainability strategies, statistical testing, subgroup fairness analysis, predictive performance metrics, and matching procedures. A model's fairness cannot be determined by a single metric. Rather, evaluating fairness necessitates taking into account both group-level results and model behavior while distinguishing between statistical difference and proof of discriminatory treatment.

6.2. Implications for Banking and Regulation

The results of the conducted research have significant implications for model development and governance processes in terms of credit risk models. In light of the growing use of quantitative analysis and machine learning algorithms in decision-making processes, the evaluation of models needs to be more than just their ability to predict outcomes. The research shows that high conventional measures of the model's performance do not mean the adequate behaviour of the model in relation to various customer segments.

Thus, it is important for commercial banks and other lenders to integrate the element of fairness assessment into the model validation procedure. Such metrics as approval rate, false positive rate, and false negative rate could be additional to conventional model performance metrics and indicate whether the results of model operation or its errors differ significantly between protected groups.

Effective model governance also depends on explainability. The Shapley value analysis showed that, in addition to traditional lending and financial factors, protected features like applicant race can influence model projections. Therefore, financial institutions should look at both the variables that are included in a model and the degree to which those variables affect its projections. By making model behaviour more transparent and assisting model developers, validators, and risk managers in identifying characteristics that need more research, explainability techniques can aid this process.

The findings also show how crucial it is to take applicant composition variations into account when evaluating fairness. After applicants were balanced based on measurable attributes, the Propensity Score Matching study decreased the observed racial outcome gap, suggesting that some of the initial differential was related to variations in the racial group composition. The discrepancy was not entirely eliminated, though. This implies that rather than depending solely on unadjusted comparisons between demographic groups, financial institutions should take into account pertinent applicant and loan characteristics while conducting fairness assessments.

The results encourage the use of various types of evidence when assessing automated credit judgement systems from a regulatory standpoint. Complementary information regarding a model's functionality can be obtained via predictive performance, model explainability, subgroup

outcomes, statistical testing, and adjusted comparisons. If a single fairness metric is used, significant variations in model behaviour or group-level results may be missed.

However, statistical discrepancies need to be interpreted carefully. A notable disparity between racial groupings does not prove that discrimination has taken place. Variations could be caused by observed traits, unobserved variables, model specifications, data constraints, or other aspects of the loan procedure. Therefore, rather than automatically attributing a discriminating interpretation to patterns, fairness analysis should serve as a diagnostic element of model governance, indicating patterns that call for additional research.

Overall, the study indicates that a balance between predictive performance, transparency, and fairness is necessary for responsible credit risk modelling. By assessing these factors collectively and recording how possible discrepancies are found, looked into, and resolved, financial institutions can improve model governance. While upholding the risk-management goals for which these models were created, such a strategy can encourage the more open and responsible application of predictive models in lending.

7. Chapter 7: Conclusion and Recommendations

Using data from the 2017 HMDA, this study investigated explainability and fairness in credit risk models. The study evaluated if racial discrepancies existed and whether they remained after controlling for observable applicant and loan variables by integrating predictive models with Shapley values, subgroup analysis, statistical testing, and Propensity Score Matching.

The results show that when assessing credit risk models, predictive performance by itself is insufficient. While variations in results were noted between White and Black or African American applicants, applicant race played a role in model predictions. While the matching process lessened but did not completely erase the racial result gap, subgroup analysis revealed that these inequalities persisted across loan types and lending purposes.

Recommendations:

When developing and validating credit risk models, financial institutions should use explainability strategies and fairness metrics. The impact of protected traits on predictions, as well as variations in results and error rates among protected groups, should receive special consideration.

By looking at more demographic groups, several years' worth of HMDA data, different fairness metrics, and other bias-mitigation techniques, future research could expand on this study. This would give a more comprehensive understanding of how diverse populations, models, and lending conditions affect the fairness of credit risk modelling.

References:

Agiza, L., Kombe, B., Musumba, C., & Philip, K. (2025). *CREDIT SCORING USING RANDOM FOREST*.

<https://doi.org/10.13140/RG.2.2.12084.31368>

Babaei, G., Giudici, P., & Neelakantan, P. (2025). Explainability, fairness and the Simpson's paradox in credit lending. *Physica A: Statistical Mechanics and Its Applications*, 680,

131030. <https://doi.org/10.1016/j.physa.2025.131030>

Bolton, C. (2010). *Logistic regression and its application in credit scoring*.

<https://api.semanticscholar.org/CorpusID:151212202>

Brock, T. (2026, June 17). *Credit Scoring: How It Affects Borrowing Terms*.

https://www.investopedia.com/terms/c/credit_scoring.asp#toc-what-is-credit-scoring

Cancino, S., & Cancino, G. (2019). Credit decision-making and information requirements.

Dimensión Empresarial, 18. [https://doi.org/10.15665/dem.v18i\(1\).2140](https://doi.org/10.15665/dem.v18i(1).2140)

Carey, V. (2020, December 2). *Fairness Metrics Won't Save You from Stereotyping*.

<https://towardsdatascience.com/fairness-metrics-wont-save-you-from-stereotyping-27127e220cac/>

Chatterjee, S. (2015). *Centre for Central Banking Studies*. Bank of England.

<https://www.bankofengland.co.uk/-/media/boe/files/ccbs/resources/modelling-credit-risk>

Chiusano, F. (2022, February 18). *Two minutes NLP — Explain predictions with SHAP values*.

Medium. <https://medium.com/nlplanet/two-minutes-nlp-explain-predictions-with-shap-values-2a0e34219177>

Christophe Hurlin, Christophe Pérignon, & Sébastien Saurin. (2024, February 9). *The Fairness of Credit Scoring Models*. arXiv. <https://arxiv.org/pdf/2205.10200>

- Clement, T., Onayemi, H., & Gaffari, M. (2025). Ethical and Bias-Aware Data Science: Quantifying and Mitigating Algorithmic Inequality. *International Journal of Scientific Research and Management (IJSRM)*, 13, 2642–2663. <https://doi.org/10.18535/ijssrm/v13i11.ec02>
- CREDIT RISK MODELLING: CURRENT PRACTICES AND APPLICATIONS*. (1999, April). Basle. <https://www.bis.org/publ/bcbs49.pdf>
- de Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for Credit Assessment in Banks. *Journal of Risk and Financial Management*, 15(12). <https://doi.org/10.3390/jrfm15120556>
- Demir, C. (2025). *Traditional Logistic Regression vs. Modern Machine Learning in Credit Scoring: A Practical Overview*. <https://pub.towardsai.net/traditional-logistic-regression-vs-modern-machine-learning-in-credit-scoring-a-practical-overview-ca3d2008bd57>
- Fernandez Vidal, M., & Menajovsky, J. (2019, September 5). *Algorithm Bias in Credit Scoring: What's Inside the Black Box?* <https://www.cgap.org/blog/algorithm-bias-in-credit-scoring-whats-inside-black-box>
- Gupta, P. (2026, May 26). *What is Credit Decisioning?* <https://www.nected.ai/blog/what-is-credit-decisioning>
- Hanna, M. G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., & Rashidi, H. H. (2025). Ethical and Bias Considerations in Artificial Intelligence/Machine Learning. *Modern Pathology*, 38(3), 100686. <https://doi.org/10.1016/j.modpat.2024.100686>
- HOW MACHINE LEARNING IS USED IN THE LENDING INDUSTRY. (2020, February 19). [TurnKey Lender]. *TurnKey Lender Blog*. <https://www.turnkey-lender.com/blog/how-machine-learning-is-used-in-the-lending-industry/>

Jain, A. (2024). *Everything about Random Forest*.

<https://medium.com/@abhishekjainindore24/everything-about-random-forest-90c106d63989>

Kelly Bailey. (2025). Why Explainable AI is Critical for Financial Decision Making. *CFI*.

<https://corporatefinanceinstitute.com/resources/artificial-intelligence-ai/why-explainable-ai-matters-finance/>

Kozodoi, N., Jacob, J., & Lessmann, S. (2021). Fairness in Credit Scoring: Assessment, Implementation and Profit Implications. *European Journal of Operational Research*, 297.

<https://doi.org/10.1016/j.ejor.2021.06.023>

Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. *European Journal of Operational Research*, 297(3),

1083–1094. <https://doi.org/10.1016/j.ejor.2021.06.023>

Liu Xuemeng. (2024, June 24). *Fairness In Credit Risk Prediction Using Machine Learning*. Tilburg University. <https://arno.uvt.nl/show.cgi?fid=181194>

M. Vasuki, Mbonigaba Celestin, A. Dinesh Kumar & Paul Johnson Asamoah. (2025). *EVALUATING CREDIT SCORING MODELS WITH STATISTICAL REGRESSION TECHNIQUES IN CONSUMER FINANCE*. Vol. 9, Number 2, pp. 180–190. <https://brainajournal.com/manuscripts/180-190.pdf>

MacPherson, L. (2026). *AI Fairness: Principles, Metrics and How to Operationalize*. Snowflake.

<https://www.snowflake.com/en/artificial-intelligence/ai-governance/ai-fairness/>

Markov, A., Seleznyova, Z., & Lapshin, V. (2022). Credit scoring methods: Latest trends and points to consider. *The Journal of Finance and Data Science*, 8, 180–201.

<https://doi.org/10.1016/j.jfds.2022.07.002>

- Oyetola, O., Madugba, J., Oladipo, A., Omolara, A., & Godwin, A. (2023). Credit Risk and its Management in the Banks: A Conceptual Review. *JOURNAL OF INTERNATIONAL MONEY BANKING AND FINANCE*, 4, 23–32. <https://doi.org/10.47509/JIMBF.2023>
- Ravi, V., Srivastava, V., Singh, M. P., Burila, R. K., Kassetty, N., Vardhineedi, P., Pasam, V., Prova, N., Prova, I., & De, I. (2025). *Explainable AI (XAI) for Credit Scoring and Loan Approvals*. <https://doi.org/10.13140/RG.2.2.35960.15368>
- Reza, M., Mahmud, M., Abeer, I. A., & Ahmed, N. (2024). *Linear Discriminant Analysis in Credit Scoring: A Transparent Hybrid Model Approach*. <https://doi.org/10.48550/arXiv.2412.04183>
- Rivalani Hlongwane, Kutlwano Ramabao, & Wilson Mongwe. (2024). A novel framework for enhancing transparency in credit scoring: Leveraging Shapley values for interpretable credit scorecards. *PLOS*. <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0308718>
- Sajid, M., & Amjad, F. (2026). *Algorithmic Bias and Fairness in AI Credit Scoring: Evidence, Mechanisms, and Governance Responses*. 5, 1546–1558.
- Thompson, M. (2025, July 21). *Fairness and Bias in Credit Risk Models*. <https://www.credit-scoring.co.uk/blog/fairnessandbias>
- Tigges, M., Mestwerdt, S., Tschirner, S., & Mauer, R. (2024). Who gets the money? A qualitative analysis of fintech lending and credit scoring through the adoption of AI and alternative data. *Technological Forecasting and Social Change*, 205, 123491. <https://doi.org/10.1016/j.techfore.2024.123491>
- Vladimir K Zhironov. (2021). *Neural network as a tool to solve the problem of credit scoring*. <https://iopscience.iop.org/article/10.1088/1742-6596/2032/1/012120/pdf>
- What is linear discriminant analysis (LDA)?* (2026). IBM. <https://www.ibm.com/think/topics/linear-discriminant-analysis>

Zhou, J., Chen, F., & Holzinger, A. (2022). *Towards Explainability for AI Fairness* (pp. 375–386).

https://doi.org/10.1007/978-3-031-04083-2_18

Zou, Y., & Li, F. (2014). *The Impact of Credit Risk Management on Profitability of Commercial Banks:*

A Study of Europe. <https://api.semanticscholar.org/CorpusID:28566351>