Declaration on the Use of Artificial Intelligence (AI) Tools

The undersigned Aysel Abdullayeva, student ID number 907448, enrolled at Ca' Foscari in the Bachelor's/Master's Degree Programme in Data Analytics for Business and Society, author of the final paper/degree thesis (Credit Risk Prediction Using Machine Learning: A SHAP-Based Interpretability and Age Fairness Analysis Across US and European Borrower Data), thesis supervisor (Cosmo Luca), aware of the penalties provided for in Article 76 of the Consolidated Act set out in Presidential Decree no. 445 of 28 December 2000, and of the loss of benefits provided for in Article 75 of the same Consolidated Act in the event of false or misleading statements, under his/her own responsibility pursuant to Articles 46 and 47 of Presidential Decree no. 445 of 28 December 2000,

DECLARES

0 that he/she has **NOT** used generative artificial intelligence technologies to support various aspects of the research and writing of the final paper/degree thesis;

0 that he/she **HAS** used generative artificial intelligence technologies to support various aspects of the research and writing of his/her final paper/degree thesis, in particular (for each point, write in full the requested information):

- indicate the name of the tool (e.g., "ChatGPT" if an application or "GPT" if a model), with reference to the specific version (for example, "ChatGPT-4, version released in March 2024");
- if more than one tool has been used, specify each of them by indicating its name and version; - Claude (Anthropic).
- provide information on how the tool was used:

  ☑ revision and optimisation of texts to improve clarity and coherence;
  ☐ generation of charts and diagrams for data presentation;
  ☑ support for the automatic drafting of sentences;
  ☐ production of illustrative images to complement the methodological section;
  ☐ other (please specify);

- indicate the sections or chapters concerned;
- indicate how the accuracy, quality and reliability of any information or text that may have been generated were verified.

DECLARES FURTHER

that all material generated by AI has been carefully checked, reviewed and reworked by the author, who assumes full responsibility for the quality, accuracy and scientific integrity of the work.

Ca' Foscari
University
of Venice

Master's Degree

Data Analytics for Business and Society

Final Thesis

Credit Risk Prediction Using Machine Learning: A SHAP-Based Interpretability and Age Fairness Analysis Across US and European Borrower Data

Academic Year: 2025/2026

Supervisor:

Cosmo Luca

Graduand:

Aysel Abdullayeva

Matriculation Number: 907848

Contents

Contents 2
Chapter 1: Introduction 5
1.1 Background and Motivation 5
1.2 Problem Statement 8
1.3 Research Objectives 11
1.4 Research Questions 15
Chapter 2: Literature Review 19
2.1 Credit Risk Assessment in Banking: Foundations and Evolution 20
2.2 Machine Learning in Credit Risk Assessment 22
2.2.1 Early Machine Learning Approaches: Neural Networks and Decision Trees 22
2.2.2 Ensemble Methods — Random Forest 23
2.2.3 Gradient Boosting and XGBoost 24
2.2.4 The Challenge of Imbalanced Data in Credit Scoring 25
2.3 Model Interpretability and Explainable Artificial Intelligence 26
2.3.1 The Interpretability Problem in Machine Learning 26
2.3.2 LIME — Local Interpretable Model-Agnostic Explanations 28
2.3.3 SHAP — SHapley Additive exPlanations 29
2.3.4 Regulatory Context for Interpretability in European Banking 30
2.4 Algorithmic Fairness in Lending 31
2.4.1 The Concept of Fairness in Algorithmic Decision-Making 31
2.4.2 Empirical Evidence on Bias in Machine Learning Credit Models 32
2.4.3 Fairness Metrics and Measurement 33
2.4.4 Regulatory Frameworks for Fairness in European Credit Scoring 34
2.5 Cross-Cultural Credit Risk Research 35
2.5.1 Differences in Credit Market Structure Across Countries 35
2.5.2 Comparative Studies of Credit Risk Models Across Countries 36
2.5.3 The Role of Regulatory Differences in Cross-Cultural Model Performance 37
2.5.4 Implications for This Thesis 37
2.6 Research Gap and Positioning of This Study 38
2.6.1 The Fragmentation of the Literature 38
2.6.2 The Interpretability Gap 39
2.6.3 The Fairness Gap 40
2.6.4 The Cross-Cultural Gap 40

2.6.5 Summary of Research Gaps and Thesis Positioning 40
Chapter 3: Data and Methodology 41
3.1 Research Design 41
3.2 Dataset Description 43
3.2.1 Give Me Some Credit Dataset 43
3.2.2 German Credit Data 44
3.3 Exploratory Data Analysis 46
3.4 Data Cleaning 49
3.4.1 Removal of Duplicate Records 49
3.4.2 Treatment of Age Outliers 50
3.4.3 Treatment of Revolving Utilization Outliers 51
3.4.4 Capping of Delinquency Variable Outliers 51
3.4.5 Imputation of Missing Values 52
3.4.6 Capping of Monthly Income and Debt Ratio Outliers 52
3.4.7 Summary of Data Cleaning Results 53
3.5 Feature Engineering 54
3.5.1 Weighted Delinquency Score 54
3.5.2 Age Group 55
3.5.3 Summary of Engineered Features 56
3.6 Preprocessing 57
3.6.1 Feature and Target Definition 57
3.6.2 Train-Test Split 58
3.6.3 Handling Class Imbalance with SMOTE 59
3.6.4 Feature Scaling for Logistic Regression 60
3.6.5 German Dataset Preprocessing 61
3.6.6 Summary of Preprocessing Results 62
3.7 Model Selection and Development 63
3.7.1 Logistic Regression 64
3.7.2 Random Forest 65
3.7.3 XGBoost 66
3.7.4 Model Evaluation Framework 67
3.8 SHAP Interpretability Framework 68
3.8.1 How SHAP Values Work in This Study 69
3.8.2 Global and Comparative Application 70

3.8.3 Implementation Details 71
3.9 Fairness Analysis Framework 72
3.9.1 Fairness Metrics Used in This Study 73
3.9.2 Connection Between SHAP and Fairness Analysis 74
3.9.3 Scope and Limitations of the Fairness Analysis 75
3.10 Use of Artificial Intelligence Tools 77
Chapter 4: Results 77
4.1 Exploratory Data Analysis Results 77
4.2 Model Performance Comparison 81
4.2.1 Hyperparameter Tuning 83
4.3 SHAP Feature Importance — US Data 84
4.4 Age-Based Fairness Analysis 88
4.5 German Credit Data Validation Results 90
4.6 Cross-Dataset Comparison and Summary 92
Chapter 5: Discussion 94
5.1 Answering the Primary Research Question 94
5.2 Sub-Question 1 — Did the Engineered Features Add Value? 95
5.3 Sub-Question 2 — Is the Model Fair Across Age Groups? 96
5.4 Sub-Question 3 — Does the Random Forest Approach Perform Comparably and Rely on Similar Factors Once Retrained on German Data? 97
5.5 Practical Implications 99
5.6 Limitations 99
Chapter 6: Conclusion 101
6.1 Summary of the Study 101
6.2 Answering the Research Questions 101
6.3 What This Thesis Contributes 103
6.4 Future Research Directions 103
References 104
Appendix 106

Chapter 1: Introduction

1.1 Background and Motivation

Credit risk refers to the probability when borrower fails to meet their financial obligations and results in financial loss for the lending institution. It's considered as one of the most significant and complex risks faced by banks and financial institutions. Unlike market risk or operational risk, credit risk, by nature it is tied to human behavior and economic circumstances. So, punctual assessment of credit risk is therefore not just a technical exercise. It is a basic requirement for the efficient operation of financial markets and the cautious distribution of capital across the economy.

The global financial crisis of 2007–2008, which is still the worst financial shock since the Great Depression, highlighted the importance of strong credit risk management. Financial institutions had systematically underestimated the credit risk hidden in complex mortgage-backed securities. When housing prices started to fall, defaults spread through the system very quickly. This collapse led to large-scale bank failures, government interventions amounting to trillions, and a long-term global recession that affected millions of people and businesses.

In the post-crisis period, international regulators answered with large-scale reforms aimed at strengthening risk management standards across the entire banking sector. The framework of Basel III which was introduced by the Basel Committee on Banking Supervision in 2010, established more serious capital requirements and improved risk measurement standards. It also introduced new liquidity requirements designed to strengthen the resilience of the banking sector. The developments have greatly increased the demands placed on credit risk modeling. After the crisis, regulators pushed hard for stricter, more auditable ways of assessing credit

risk. Banks could no longer lean on informal judgment calls. They had to use methods that could be validated openly and consistently.

The use of machine learning (ML) techniques for assessing credit risk has become one of the most promising and expanding fields of research in financial analytics. For many years, traditional statistical methods such as logistic regression and linear discriminant analysis have formed the basis of credit scoring systems. The main benefit of these methods is that they are interpretable. The predictions can be linked directly to borrower features and presented to both regulators and customers. Yet these methods assume a linear relationship between the input variables and the log-odds of default and may therefore be unable to account for complex, non-linear interactions among borrower characteristics, a situation that is frequently found in real-world credit data.

Machine learning algorithms, by contrast, can learn highly complex patterns from large volumes of data without imposing restrictive distributional assumptions. In credit risk applications, models like Random Forest and gradient boosting algorithms have continually shown better predictive performance than traditional methods. Which is so documented in benchmark studies such as Lessmann et al. (2015), who compared 41 classification algorithms on credit scoring datasets and found that ensemble methods significantly outperformed logistic regression baselines. Use of machine learning in regulated credit situations creates a basic contradiction, even though it has the predictive advantage. The models function as black boxes and the reasoning inside them is hard or impossible to understand without additional analytical tools. The tension between predictive power and transparency is at the base of the present discussion concerning the responsible use of artificial intelligence in the financial services sector.

Apart from accuracy and interpretability, credit risk models have significant societal implications which are gaining increasing attention from both regulators and academics. For example, a borrower who is incorrectly classified as high risk may be denied access to credit entirely or offered with high interest rates only. Results can affect their ability to purchase a home, invest in education, or manage unexpected financial emergencies. When misclassifications have an unbalanced effect on certain demographic groups, such as younger borrowers or people from lower income backgrounds. The model is effectively spreading on or increasing existing social inequalities. The Artificial Intelligence Act of the European Union, which came into force in 2024, classifies AI systems used for assessing creditworthiness or for setting up a credit score of natural persons as high-risk applications. Therefore, it requires them to meet strict obligations in the areas of risk management, transparency, data governance, accuracy and non-discrimination. The original deadline for meeting these high-risk obligations was August 2026 but has now been postponed to December 2027 as a result of a correction adopted across the EU which gives providers more time to meet the new requirements. This regulatory context means that examining fairness in machine learning credit models is not just an academic task but a practical and legal obligation for financial institutions operating in Europe.

This thesis comes about within that wider context which includes the changing demands for risk management, the growing capabilities of machine learning, the increasing regulatory requirements, and the rising focus on algorithmic fairness. Based on my professional experience in the banking industry, I witnessed the way credit evaluation decisions affected both the institution's level of risk and that of individual borrowers. This study seeks to provide

empirical evidence and analytical insights into the current debate concerning the application of machine learning to credit risk assessment.

Statement for the use of Artificial Intelligence tools during the preparation of this thesis is provided in Section 3.10.

1.2 Problem Statement

Even though there has been an increasing amount of research into credit risk modeling, there are some important limitations which still exist in the current literature, and this thesis seeks to deal with them. The limitations concern three closely related aspects: the methodological gap between predictive performance and model interpretability, the insufficient attention paid to demographic fairness in algorithmic credit decisions, and the limited cross-cultural generalizability of existing findings. All these gaps motivate the research design adopted in this study and determine the specific contributions it seeks to make.

The first limitation concerns the tension between predictive accuracy and interpretability in machine learning credit models. Over the past two decades, the field has witnessed clear progress from traditional statistical methods like logistic regression and linear discriminant analysis towards more complicated ones and gradient boosting algorithms. Studies such as Lessmann et al. (2015) demonstrate that many modern machine learning classifiers can do better than traditional scorecard approaches evaluation metrics that are commonly used. However, this gain in accuracy has been achieved at a heavy cost. The internal logic of models such as Random Forest and XGBoost is more complex and less transparent than those of traditional statistical models. Random Forest generates predictions by combining the outputs of multiple decision trees. While XGBoost uses a sequential boosting approach, each new tree is built to improve the errors of previous trees. As a result, to understand the reasons for

individual predictions, it is necessary to use further interpretability techniques such as SHAP values.

This lack of transparency causes serious practical difficulties within the field of regulated financial services. Banks operating under the European Capital Requirements Regulation and European Banking Authority guidelines are required to ensure that credit risk models are appropriately governed, validated, documented, and explainable to supervisory authorities. In addition, the General Data Protection Regulation (GDPR) creates protection for individuals affected by automated decision-making and provides certain rights regarding decisions based on automated processing under Article 22. In fact, this means that a bank cannot simply place a high-performing black-box model and rely on only predictive accuracy. It must also be able to explain why a particular borrower was classified as high or low risk, which variables drove that classification, and how sensitive the prediction is to changes in individual input values. The most accurate machine learning models may be effectively unusable in a compliant banking environment, without appropriate interpretability tools.

Methods for post-hoc model explanation have advanced significantly in recent years. It is most notable through the development of SHAP (SHapley Additive exPlanations) values by Lundberg and Lee (2017) . Their use in credit risk research has increasingly gained attention as researchers seek to improve the interpretability of complex ML models. However, many studies primarily use SHAP values to generate global feature importance rankings. They place less importance on explanations at the individual level, on comparing patterns across different models, or the connection between interpretability results and regulatory requirements. This thesis addresses this gap by positioning SHAP-based interpretability not as a supplementary analytical step but as a central research contribution.

The second important gap in the existing literature covers algorithmic fairness. Because credit scoring models are trained on historical data, they can pick up the patterns that are present in that data, including any differences in differences between groups of people. For example, younger borrowers tend to default more often, because they have shorter credit histories or less stable income. A trained model on this data can pick up on that pattern and start linking age itself to higher risk. This can be even for someone whose personal situation is actually fine. Discrimination like this can lead to worse lending outcomes for entire groups, cutting their access off to credit and extending financial inequalities that exist already.

The regulatory response to this concern has become increasingly clear. The EU's AI Act has classified credit scoring as a high-risk form of AI because of the way it can affect people's fundamental rights, and this includes the risk of discriminatory outcomes. Despite this regulatory attention, the academic literature on fairness in credit risk modelling remains fairly limited. Most existing studies focus just on predictive performance metrics such as AUC and F1 scores. And they don’t examine whether model predictions differ systematically across demographic subgroups. This thesis directly addresses this gap through a structured fairness analysis. It compares actual and predicts default rates across three borrower age groups. The groups of Young Adults, Middle-Aged, and Older Adults provide practical evidence which model produces age-differentiated predictions and discusses the implications of these differences for fair lending practice.

For the 3rd limitation of the existing literature is the geographical scope of credit risk research. Most published studies rely on data from just one country, usually the US. They reach conclusions about model performance and identify which features are most important without verifying that these conclusions are valid in other cases. The importance of this limitation

should not be underestimated since there are differences in credit markets from one country to another. They are in terms of data availability, lending practices, regulatory frameworks, borrower behavior, and the relative importance of different risk indicators. A model that works well by relying on delinquency history in the US might behave very differently in Europe. Things like savings behavior, employment stability, and account status may say more about creditworthiness for European borrowers.

This gap is not just a methodological detail but matters in practice too. Both banks that operate across different borders and the regulators who aim to establish uniform credit AI standards in Europe have a genuine interest in this issue. This thesis trains a model on the Give Me Some Credit dataset from the US, then retrains the same structure on the Statlog German Credit Data from the UCI Machine Learning Repository. Comparing how it performs and what it relies on in each country adds an international angle which is often missing from credit risk research.

1.3 Research Objectives

This thesis pursues four main objectives. Each objective targets a specific limitation and together they form the foundation of the research design.

Objective 1 — Build and compare machine learning models for credit risk prediction.

The first objective is developing and comparing three machine learning classification models for predicting if a borrower will default on their credit duties. The three models selected are Logistic Regression, Random Forest, and XGBoost. For decades Logistic Regression has been widely used in credit scoring, which is a traditional model. In this study it was included as a transparent starting point. It is a reference point against the performance of more complex models that can be measured. Random Forest builds a large number of decision trees and combines their outputs to produce a final prediction. It has strong performance on classification

tasks and ability to handle complex relationships between variables and doesn't require strict statistical assumptions. XGBoost, short for Extreme Gradient Boosting, is a more advanced algorithm. It builds trees sequentially; each new tree corrects the errors of the previous one.

All these models are evaluated using a set of standard performance metrics. The core metric is the Area Under the Receiver Operating Characteristic Curve, commonly referred to as AUC-ROC. This metric measures how well the model distinguishes borrowers who will default and who will not, with a value of 1.0 representing perfect discrimination and 0.5 representing no better than random chance. The F1 Score is also reported, as it provides a balance between precision and recall. It makes it informative for evaluating the model's ability to correctly identify defaulters and limits false predictions. The severe imbalance in the dataset is a crucial point, since only 6.7% of the borrowers defaulted.

Class imbalance is a situation where one outcome is far more common than the other poses and it is a significant challenge for credit risk modelling. If big part of borrowers in the training data are non-defaulters, a model can achieve high overall accuracy simply by predicting non-default for everyone. While it completely can fail to identify the borrowers who will actually default. This study applies to the Synthetic Minority Over-sampling Technique, known as SMOTE for solving this problem. SMOTE works by generating synthetic new examples of the minority class. This produces a more balanced training dataset and encourages the model to learn the characteristics of defaulters rather than just predicting the majority class. SMOTE is used only on the training data and not at all on the test data, ensuring that the assessment of the model mirrors the actual class distribution that happens in practice.

Objective 2 — Enrich the analysis through original feature engineering

The second objective is to contribute two original variables that are made from raw data and are not in the original dataset. Feature engineering was used creating new variables from existing ones using domain knowledge. It is an important step in machine learning that can significantly improve model performance and analytical insight.

The first variable that was engineered is Weighted Delinquency Score. The original dataset contains three separate columns recording how many times a borrower was 30–59 days late, 60–89 days late, and 90 or more days late on a payment. Instead of treating these three unrelated variables, this study joins them into a single severity-weighted score using the formula: (30–59 days late × 1) + (60–89 days late × 2) + (90+ days late × 3). This weighting follows a basic principle in credit risk management. More serious late payments higher default risk, and the score is built to reflect that directly.

The second engineered variable is an Age Group classification. Age is as a continuous numeric variable in the original dataset. In this study, age is transformed into a categorical variable with three groups: Young Adults (18–35), Middle-Aged (36–55), and Older Adults (56–90). The purpose of this variable is to enable a structured fairness analysis. Because it creates clearly defined demographic subgroups and their prediction outcomes can be directly compared. The original continuous age variable stays in the predictive model itself, so AgeGroup serves an analytical rather than predictive role.

Objective 3 — Explain the best model using SHAP values

The third objective is applying SHAP-based interpretability analysis to the best performing model. So, we can understand which variables drive their predictions. SHAP stands for SHapley Additive explanations, it is a method developed by Lundberg and Lee (2017).

SHAP analysis addresses, how much did each variable contribute to their predicted default probability? A high positive SHAP value for a particular variable means that variable pushed the prediction towards default and negative one away from default. It is possible to get an overall view to know important variables to the model and how they interact with the target outcome by combining the explanations for all the borrowers in the test set.

This level of interpretability goes far beyond what traditional features of importance measures are able to offer. It allows the bank or regulator to understand not just which variables matter, but how they matter. In this study, SHAP analysis is applied to the best performing model on the US dataset and repeated on the German dataset. A direct comparison was carried out to determine which factors influence credit risk predictions in the two different financial contexts.

Objective 4 — Assess fairness across age groups and compare model performance across US and European data

The fourth objective brings together a fairness analysis with a cross-cultural comparison approach. The fairness examination looks at whether the top-performing model consistently makes different predictions for borrowers in different age groups. Specifically, it compares the actual and predicted default rates for each of the three age groups mentioned above. Potential fairness issues may arise if the model continually generates significantly different prediction patterns across age groups, especially if certain groups constantly overestimate or underestimate default risk. There is regulatory and ethical concern regarding the use of demographic characteristics in decisions made by algorithms concerning credit. This analysis has also been influenced by the EU AI Act.

The analysis asks a more specific question: if the same Random Forest method is retrained using German data, will it achieve a similar degree of accuracy? Will it rely on the same borrower's characteristics in order to do so? The best performing model from the primary analysis is retrained on the German Credit Data dataset. It is a widely used European benchmark dataset from the UCI Machine Learning Repository and contains 1,000 borrower records. The model is checked using the same criteria as those used for the US data, and a SHAP analysis is once more carried out to get a feature importance profile for the German model. The comparison between the US and German SHAP profiles gives direct evidence of if the same borrower characteristics drive credit risk in both financial systems or different risk factors operating in the European context.

1.4 Research Questions

The four mentioned objectives give rise to the research questions that guide the analytical framework of this dissertation. These questions are organized around one primary question and three supporting sub-questions. Each of them is directly linked to one of the research objectives described above.

Primary Research Question

Which machine learning model most effectively predicts individual credit risk among borrowers, and what are the key factors driving those predictions?

This is the central question of the thesis and motivates the entire analytical pipeline. It asks not only which model performs best in a technical sense measured through AUC and F1 score, but also about borrower behaviour. The second part of the question is, what is different than a standard model comparison exercise. It is useful to know AUC of 0.82. But knowing which

specific borrower characteristics caused that model to flag a particular individual as high risk makes the analysis useful for banks, regulators, and policymakers.

Sub-question 1

Does the Weighted Delinquency Score provide meaningful predictive value and interpretability compared to the original variables, and does the Age Group variable meaningfully support the age-based fairness analysis?

This is where feature engineering contribution of the thesis addresses the question. Creatin new variables and including them in the model is not enough. Their value must be demonstrated empirically. The question is answered through two approaches. On the one hand, the SHAP feature importance analysis shows if the engineered variables appear among the best predictors in the model. On the other hand, the position of the original delinquency columns in this ranking shows whether the Weighted Delinquency Score has successfully incorporated their combined predictive power or whether the separate columns still contain information that the mixed score fails to capture.

Sub-question 2

Do model predictions differ across borrower age groups, and what are the implications of any observed differences for fairness in algorithmic lending?

This addresses the fairness dimension of the thesis. Instead of asking whether the model is accurate overall and it asks whether it is equally accurate and equally fair for borrowers of different ages. This is comparison of the actual default rate within each age group against the rate predicted by the model. A large gap between these two figures for a particular group

suggests that the model is biased in its treatment of that group. Overestimating their risk could lead to unfair credit denials, underestimating it could expose the lender to unexpected losses.

Sub-question 3

Does the Random Forest approach reach comparable predictive performance once retrained on a European credit dataset, and do the same factors drive its predictions in the US and German data?

This sub-question addresses the cross-cultural dimension of the thesis. It is what distinguishes this study from the majority of existing credit risk research. Because they usually focus exclusively on a single national context. It asks two related but different things. The first is, does the same Random Forest modelling approach achieve comparable predictive performance when it is retrained on German borrower data? The second is, even if the model performs similarly in both, does it rely on the same variables to make its predictions? By comparing the SHAP feature importance profiles from the US and German analyses side by side, this thesis provides both aspects of the of this question.

This thesis aims to contribute to research on machine learning-based credit risk assessment. These contributions are methodological, empirical, and practical and are summarized below before. Rather than working alone, these contributions complement one another since each one looks at a different aspect of the credit risk problem and provides a more detailed view than any single contribution alone would.

Methodological Contributions

The first methodological contribution of this thesis is the development and empirical evaluation of two engineered features. Both the Weighted Delinquency Score and the Age

Group variable were both designed on the knowledge of banking and on practical experience in the field of credit risk, not on automated feature selection procedures. This approach reflects a broader methodological argument that machine learning models applied to financial problems. The best results are achieved not by treating the data as a black box to be processed by algorithms, but by actively shaping the inputs to the model.

The second methodological contribution is the integration of SHAP-based interpretability analysis. Usually in the existing credit risk literature, model interpretability is treated as an afterthought. This thesis tries a different approach, using SHAP analysis at two levels: global feature importance and cross-dataset comparison. It views interpretability not as a technical tool but as a kind of research instrument in itself.

The third methodological contribution is the cross-dataset comparison framework. Instead of assessing the modeling on one dataset and making conclusions based on a single nation's situation, this thesis uses two datasets from different countries. This cross-dataset gives more rigorous comparison test than one based on a single dataset could. The models are trained and evaluated on a primary dataset and then retrained and compared on a secondary dataset using the same architecture and evaluation metrics. They might be used in future research for a systematic investigation into the cross-cultural portability of credit risk models across a wider variety of national contexts.

Empirical Contributions

The primary empirical contribution of this thesis is a detailed comparison of three machine learning models Logistic Regression, Random Forest, and XGBoost on the Give Me Some Credit dataset. Also, combination of AUC-ROC, F1 Score, Precision, and Recall metrics were used following rigorous data cleaning, feature engineering, and class imbalance treatment.

The second empirical contribution is the SHAP feature importance analysis. It reveals that the engineered Weighted Delinquency Score is the single most important predictor of credit default in the best performing model.

The third empirical contribution is the age-based fairness analysis. It provides quantitative evidence to see which model produces different predictions for borrowers in different age groups.

The fourth empirical contribution is the cross-cultural comparison between the datasets. They demonstrate in the United States the behavioral variables determine the predictions, but in Germany it is the structural variables.

All these contributions are discussed in full in Chapter 4 and Chapter 5.

This thesis may offer practical value for banks and future researchers working on credit risk, interpretability, and fairness. These implications are discussed in full in Section 6.3.

Chapter 2: Literature Review

This chapter reviews the existing academic literature that is relevant to the research questions and the aims of this thesis. The review is divided thematically into six sections, each of which deals with a separate area of knowledge that contributes to the analytical framework used in this study. The chapter begins with an examination of the foundations of credit risk assessment in banking. Then, reviews the literature on class imbalance in credit datasets, model interpretability and explainable artificial intelligence, and algorithmic fairness in lending. The chapter finishes with a review of cross-cultural credit risk research and a synthesis of the identified gaps which this thesis aims to address.

2.1 Credit Risk Assessment in Banking: Foundations and Evolution

The story of credit risk assessment can be seen in many ways as the story of efforts to achieve greater objectivity, consistency, and predictive power. It is important to understand this history not only as academic background but also to see why machine learning methods have attracted so much interest in recent decades. Also, what specific limitations of earlier approaches are intended to overcome.

In first half of the 20th century, credit decisions in banking were made mainly based on the personal judgement of loan officers. Applicants were evaluated against informal criteria related to character, capacity, capital, collateral, and conditions. This framework was known in the industry as the Five Cs of Credit. But this approach was inherently inconsistent, difficult to audit, and vulnerable to personal bias. Different loan officers at the same institution might reach different conclusions about the same applicant. The criteria applied could vary over time, or even by factors unrelated to the borrower's creditworthiness, such as officer's mood, their relationship with the applicant. As banks wanted to deal with more applications, the drawbacks of relying on judgement-based credit assessment became increasingly apparent.

One of the first formal statistical approaches to credit risk assessment emerged in the 1960s. The publication of Edward Altman's landmark 1968 paper introduced the Z-score model for predicting corporate bankruptcy. Altman's model used multiple discriminant analysis, such as statistical technique that identifies the linear combination of financial ratios aimed to separate bankrupt firms from healthy ones. By this, it was possible to produce a single composite score to classify companies as financially distressed or sound. The Z-score model was at first designed for use in corporate credit assessment, not for consumer lending. Nevertheless, it set a few principles that later shaped every quantitative approach to credit risk. These included

using multiple financial variables, applying statistics to work out the best variable weights from past data, and testing model performance on data the model hadn't seen before. Altman's work opened the door to a new era of data-driven credit assessment and demonstrated that quantitative models could do better than human judgement for predicting financial distress. (Altman, 1968).

The statistical methods of applications to consumer credit scoring followed a similar trajectory. It speeded up significantly in the 1970s and 1980s as the consumer credit market expanded and the loan applications grew beyond that manual review would not handle. For the first time, the development of the FICO score by the Fair Isaac Corporation in 1989 represented a landmark moment in this evolution. It was a standardized numerical score copied from credit bureau data that was widely adopted by lenders for consumer credit decisions. The FICO score is based on five categories of information: payment history, amounts owed, length of credit history, new credit, and types of credit used with payment history carrying the greatest weight. With it, consumer lending is transformed by making credit decisions faster, more consistent, and more auditable. It also enabled lenders to price risk more accurately by linking interest rates to creditworthiness scores.

Throughout the 1980s and 1990s, logistic regression was main statistical method used for credit scoring. Unlike linear discriminant analysis, logistic regression doesn’t make distributional assumptions. This is why it is better suited to the kinds of binary, skewed, and non-normally distributed variables which are commonly found in credit datasets. Logistic regression models the probability, by producing output values that lie between zero and one, and they can be directly interpreted as default probabilities. This makes logistic regression particularly well suited to credit scoring applications. The goal is to rank borrowers according

to their estimated probability of default rather than simply dividing them into categories of defaulters and non-defaulters so that lending decisions can be adjusted accordingly.

Logistic regression has its own limitation besides the advantages. They become apparent as the complexity and volume of credit data grew. It can only capture linear relationships between predictor variables and the log-odds of default. But the relationship between borrower characteristics and credit risk is rarely that simple. As an example, borrower's debt ratio may have little effect on default probability at low levels but above a certain threshold it may have increased effect. It is not possible to capture these non-linear relationships and interaction effects without the addition of transformed variables and interaction terms, for logistic regression. It requires significant domain knowledge and manual feature engineering effort. The motivation of the research was this limitation. Complex patterns could automatically be detected with more flexible modelling approaches in credit data without specifying them in advance.

2.2 Machine Learning in Credit Risk Assessment

2.2.1 Early Machine Learning Approaches: Neural Networks and Decision Trees

Early applications of machine learning to credit scoring came out in the late 1980s and early 1990s. Neural networks were appealing because it was possible to capture complex non-linear relationships without it being necessary to specify the functional form of those relationships in advance. For example, West (2000) compared several neural network architectures with logistic regression, linear discriminant analysis and decision trees across two real credit datasets. Even though, results showed that some neural network architectures got strong predictive performance, logistic regression stayed as the strongest among the traditional methods examined.

Despite their predictive potential, neural networks were relatively difficult to interpret. This limited their practical appeal in fields where transparency and auditability are crucial. Decision trees offered an alternative by providing clear branching rules so analysts could interpret more easily. However, individual decision trees were sensitive to instability and overfitting, since their structure could depend strongly on the training data used (Breiman et al., 1984). The limitations were motivation for later ensemble approaches that combine multiple trees in order to improve predictive stability and generalization.

2.2.2 Ensemble Methods — Random Forest

As a major step forward for machine learning, Ensemble methods emerged in the late 1990s and early 2000s, with real implications for credit risk modelling. The core idea is simple: combining several models trained on different subsets of the data. So, it can overcome the weaknesses of any single model.

Random Forest, designed by Breiman (2001), is the most widely used ensemble method in credit risk research. It builds hundreds of decision trees, and each of them trained on a random sample of the data and subset of features. Then, it combines their predictions through majority voting. After that, models are far more stable than a single tree and generalizing new data is better.

Across credit scoring studies, Random Forest has performed consistently well. Lessmann et al. (2015), found Random Forest among the top-performing methods, which clearly outperforms logistic regression on AUC across most datasets tested. It captures non-linear relationships, handles outliers well, and needs relatively little tuning. And this is what makes it both accurate and practical to apply.

Because the predictions come from hundreds of trees and there's no simple rule for any single prediction, the trade-off is interpretability. And this is a real problem in regulated banking. Because the decisions must be explainable to regulators and borrowers alike. Post-hoc tools like SHAP, (covered in Section 2.3) help address this by breaking predictions down into each variable's contribution. But the issue of the relationship between predictive power and interpretability is still the subject of ongoing debate.

2.2.3 Gradient Boosting and XGBoost

Gradient boosting is an important learning approach in credit risk research. Unlike Random Forest, gradient boosting trains trees sequentially. Each new tree is designed to reduce the errors of the current ensemble. Gradient boosting is generally more sensitive to hyperparameter settings and may be prone to overfitting if not appropriately regularized. Nevertheless, this sequential learning process can achieve strong predictive performance. Studies that have compared machine learning methods in credit scoring have found that both Random Forest and Gradient boosting show strong performance, especially when there is a class imbalance (Brown & Mues, 2012).

XGBoost (Extreme Gradient Boosting) is a scalable implementation of gradient-boosted decision trees designed by Chen and Guestrin (2016). It incorporates regularisation, a sparsity-aware learning procedure, and weighted quantile sketch for approximate tree learning. The characteristics have contributed to machine learning applications, including credit risk modelling.

Selected, Logistic Regression, Random Forest, and XGBoost methods for this thesis provide a confrontation between a relatively open to interpretation linear model and two more complex tree-based ensemble methods. Logistic Regression gives a transparent baseline but is limited

in capturing complex non-linear relationships. Through an ensemble of decision trees, Random Forest can model non-linear interactions. While XGBoost uses sequential boosting to further increase model flexibility. Three approaches allow the thesis to examine differences in predictive performance and interpretability. In addition, SHAP is used to investigate factors influencing the predictions of tree-based models.

2.2.4 The Challenge of Imbalanced Data in Credit Scoring

Credit scoring datasets frequently exhibit class imbalance. For example, when the large majority of borrowers do not default and only a small minority do. In the Give Me Some Credit dataset used in this thesis, 93.3% of borrowers are non-default and only 6.7% are default. A model that predicted no default for every borrower can have an accuracy of 93.3%, yet it could not identify any real defaulters. Thus, accuracy alone would not be a suitable criterion for evaluating it.

Brown and Mues (2012) compared several classification algorithms. This included logistic regression, decision trees, random forests, and gradient boosting. They were across five real-world credit scoring datasets under increasing class imbalance and AUC was as the performance measure. In result, random forest and gradient boosting classifiers were found coping comparatively well with severe imbalance. Meanwhile, simpler classifiers such as decision trees and k-nearest neighbors performed significantly worse. The results support to include both Random Forest and XGBoost in the selection of models that are being compared in this thesis.

For imbalanced classification, the use of AUC-ROC as a primary evaluation metric is also appropriate. Because it evaluates discriminatory ability across classification thresholds.

This thesis applies Synthetic Minority Over-sampling Technique (SMOTE), to address the imbalance in the training data itself, (introduced by Chawla et al. (2002)). Instead of simply duplicating existing minority-class observations, SMOTE generates synthetic minority-class examples by interpolating existing observations in the feature space. Although its effectiveness depends on the dataset and modelling approach, SMOTE has shown improvement to minority-class recognition in a range of imbalanced classification settings (Chawla et al., 2002).

There is more recent work to examine these questions. Schmitt (2022) found that gradient boosting, including XGBoost, achieved strong performance relative to deep-learning approaches. Also, model performance depended on the underlying dataset. It is consistent with the model comparison approach adopted in this thesis.

2.3 Model Interpretability and Explainable Artificial Intelligence

Researchers question how easily humans can understand and explain the predictions made by machine learning models. In so many areas, the opacity of complex machine learning models is an acceptable trade-off for the gains for the improvement in predictive accuracy. In financial services and in credit risk assessment interpretability is not just a desirable property. It is also a practical and regulatory necessity. This section reviews three things: what interpretability means, how researchers tried to get it and how necessary it is in European banking specifically.

2.3.1 The Interpretability Problem in Machine Learning

In ML the interpretability problem arises from a fundamental tension between predictive power and transparency. Simple models such as linear regression and logistic regression are highly interpretable. Their predictions are built through a mathematical formula. The formula's parameters can be directly examined to understand the association with the model's output.

However, as reviewed in Section 2.2, simple models are limited to capturing complex non-linear relationships that define real-world credit data. Powerful models like Random Forest, XGBoost, and deep neural networks can capture more complex non-linear relationships. But they need to compromise the transparency that makes simple models easy to understand and explain. For example, Random Forest model, produces its predictions by combining the outputs of multiple individual decision trees.

This creates real problems with banking. The problem is not that a black-box model's predictions can't be examined at all. Tools such as SHAP can assign an individual prediction to contributions from different variables. However, a model constructed from many interacting trees is considerably less transparent than a single formula. They are much more difficult to express as a simple, human-readable rule. This is important legally as well as technically. According to the US Equal Credit Opportunity Act and Regulation B, creditors are required to give specific reasons for any negative credit decision. Another difficulty is producing explanations that are both reliable and technically sound. Because it is not like the banks are free to say nothing simply because a model is complex. When machine learning has moved from research labs into real-life practice, closing this gap between predictive power and genuine explainability has become an important concern for both academics and policymakers (Rudin, 2019).

The academic literature tries to distinguish two different ways of dealing with this problem. The first one is using models that are interpretable by design. These are the ones that are simple enough for a human analyst to understand directly, without needing extra tools for explaining. Logistic regression, linear regression, single decision trees, and rule-based classifiers are all in this category. Their advantage is that transparency is built in by design. There's no gap between

what can be communicated to a stakeholder and what the model does. Their disadvantage, as discussed in Section 2.2, is that they may have lower predictive performance on tasks involving complex relationships. Rudin (2019) makes a strong case in favor of selecting interpretable models rather than black-box models coupled with additional explanations in high-stakes situations like credit scoring as well. He claimed that post-hoc explanations can be unreliable, and that the accuracy gap between simple and complex models is often smaller than commonly believed. This is an active debate position in the field.

2.3.2 LIME — Local Interpretable Model-Agnostic Explanations

One of the most frequently cited post-hoc methods is LIME, which was introduced by Ribeiro et al. (2016). It works by fitting a simple, interpretable model. The logic here is simple. Even when the overall decision boundary of a complex model is highly non-linear, the behavior in the neighborhood of any point can often be approximated by a much simpler approach. LIME builds this approximation by randomly changing or masking input features. It runs the original model on each changed version and fits a weighted linear model to the results. It gives more weight to points closer to the original observation. The obtained coefficients indicate the direction and relative importance of each feature within the framework of that local approximation, rather than the model's true internal logic.

LIME is especially useful when it comes to explaining the outcome for a particular borrower since it is more important to understand that individual case than the model as a whole. But it has real limitations. The explanations are strictly local, so two similar borrower profiles can end up with noticeably different explanations. Results depend on exactly how the perturbation is done. Chen et al. (2024) examine this directly in a credit scoring context. And the finding is that LIME and SHAP explanations both become less stable as class imbalance increases. What

is directly relevant here is, thesis's own imbalanced dataset. These limitations are part of why SHAP discussed next.

2.3.3 SHAP — SHapley Additive exPlanations

SHAP was introduced by Lundberg and Lee (2017). It's one of the most widely used post-hoc explanation frameworks in applied ML which includes credit scores as well. Unlike LIME, SHAP is based on formal mathematical theory, such as Shapley values in cooperative game theory.

Lloyd Shapley developed this idea in 1953. The original problem was different from machine learning entirely. His purpose was splitting a game's payoff among players that contributed differently. As a solution, he gives each player their average amount that they had contributed in every possible order in which the players could have joined the game. It satisfies a number of fairness criteria, such as efficiency and symmetry. Lundberg and Lee (2017) saw that explaining a model's prediction is essentially the same problem. Because each input feature can be treated as a player. The contribution depends on which other features are present. The SHAP values obtained by applying Shapley's method measure a feature's average effect on a prediction by examining it in all possible combinations of features.

This gives SHAP some real advantages over LIME. SHAP values are locally accurate. Summing up every feature's contribution, plus a baseline value, gives you exactly the model's actual prediction. Nothing is left unexplained. SHAP values are also consistent. If a feature's true importance goes up in a different version of the model, its SHAP value can't go down. And a feature that has no real effect on the prediction always gets a SHAP value of zero.

SHAP is used both globally and locally in credit risk research. Globally speaking, averaging the absolute SHAP value of each feature across every borrower gives a clear picture of what matters most overall. Locally speaking, SHAP force plots explain one prediction at a time, as kind of detail a regulator or a rejected applicant might actually need. Lundberg et al. (2020) built on this with TreeSHAP. It computes exact SHAP values for tree-based models like Random Forest and XGBoost faster than a direct calculation would do. That is what makes SHAP practical and fast for a dataset the size used in this thesis.

In credit scoring, Interpretability research has moved beyond applying these tools one case at a time. Bücker et al. (2021) developed a structured framework that made credit scoring models transparent, auditable, and explainable. They showed machine learning models can achieve a level of interpretability similar to that of traditional scorecards without having to sacrifice their predictive power. This thesis tries to take a similar approach and SHAP is not just used for explaining only individual predictions but is a central part of the whole analysis.

2.3.4 Regulatory Context for Interpretability in European Banking

The academic interest in interpretability isn't only theoretical; it is backed by real EU law. GDPR's Article 22 has provided a right to individuals object to decisions made entirely by automated systems, if automated decisions affect their legal or financial rights. The Court of Justice of the EU ruled in the SCHUFA case that the automated generation of a credit score can fall into the scope of Article 22 under certain circumstances, when the score is given a determining role in a resulting decision affecting an individual. The court also said that individuals are entitled under Article 15(1)(h) to meaningful information about the actual logic behind the score. It shouldn’t be only based on underlying formula, but an explanation of how

it was applied. This means that under the relevant conditions, some genuine account of the model's reasoning may be legally required for the banks that use ML credit scoring.

The EU AI Act that was adopted in 2024, goes further it classifies AI systems used to evaluate the creditworthiness as high-risk under Annex III, requiring documentation, transparency, and human view. Article 13 says high-risk systems have to be transparent enough for users to interpret their outputs. Article 14 even goes further and requires that human overseers be able to understand what it is producing. A specific part of Article 14(4)(c) especially names "interpretation tools and methods" as part of meeting this requirement.

This is why interpretability matters for this thesis exactly. It is not just as an academic exercise. SHAP explains the model's predictions and compares how it reasons across US and German data.

2.4 Algorithmic Fairness in Lending

Fairness in algorithmic decision-making is one of the most important and developing areas of research at the intersection of law, machine learning and social science. In credit lending, fairness concerns are important because the decisions have direct and significant consequences for individuals.

2.4.1 The Concept of Fairness in Algorithmic Decision-Making

For an algorithm, defining fairness is harder than it sounds. Accuracy and efficiency may be measured precisely. But fairness can't. Because fairness is something not easy to measure objectively. It depends on what you personally believe is treating people fairly and others can genuinely disagree with that. Researchers have proposed several formal definitions of fairness, and many of them actually contradict one another. You generally can't satisfy several fairness

criteria at once. The decision to choose one is a value judgment concerning which is most important.

Over twenty of these definitions were reviewed by Mehrabi et al. (2021) and looked at the sources of the bias. Three definitions come up most often. Demographic parity says regardless of actual outcome, positive predictions rate should be the same across groups. For equalized odds, the model's true and false positive rates should be equal across groups. And for individual fairness, similar people should get similar predictions. Each presents a different and reasonable idea of what fairness is.

2.4.2 Empirical Evidence on Bias in Machine Learning Credit Models

Academically and by regulatory interest, empirical research on bias in ML credit models has grown quickly. Fuster et al. (2022) compared logistic regression against more complex models on US mortgage data and across different demographic groups. The findings were mixed. More accurate models sometimes reduced discrimination by relying more on individual financial data rather than demographic indicators. But in other cases, they have further deepened existing inequalities by uncovering the links between demographic indicators and outcomes hidden within historical data. The main conclusion is simple but important: a more accurate model does not automatically mean a fairer model.

Kozodoi et al. (2022) tested a framework that applies fairness limitations to model training. They found a trade-off generally exists. A fairness-limited models tend to score lower on AUC But the size of that trade-off varies a lot based on the dataset and which fairness definition is used. The accuracy cost was small in some cases. This thesis doesn't impose any limit on fairness. It measures how much age-based bias exists in an unlimited model first. Which is a natural starting point before considering limited approaches like theirs.

Age is far less research attention than race or gender. And it actually dominates US fairness studies due to the Equal Credit Opportunity Act and the Fair Housing Act. Age discrimination in lending is still explicitly banned under the Equal Credit Opportunity Act. In the EU, age-based bias is covered less directly, through the AI Act's high-risk classification of credit scoring and GDPR Article 22's automative decision-making protections. This gap in the research motivates this thesis's fairness analysis, combined with age's real legal weight in both the US and EU.

Hurlin et al. (2024) recently pushed this further and developed a formal statistical test that finds exactly which variables cause a credit model to be unfair. That's a more comprehensive approach than the descriptive one and is a natural direction for extending this analysis in future work, if needed.

2.4.3 Fairness Metrics and Measurement

In credit scoring, most fairness metrics compare outcomes across demographic groups. How often do each group gets a positive prediction? How accurate do those predictions turn out to be?

A few metrics are commonly used. The difference of demographic equality is the gap in positive prediction rates between the most and least advantaged groups. Zero means perfect equality, and the bigger the number is basically inequality. Equalized odds difference works similarly, but it looks at the difference in the true positive and false positive rates. Predictive equality, closely related to calibration, asks whether a predicted risk matches with reality across groups. For example, if the model says someone has a 20% chance of default, which should match with the truth regardless of their age, gender, or any other group.

Instead of using any of these formal metrics directly, this thesis compares actual and predicted default rates within each of the three age groups. Young Adults (18–35), Middle-Aged (36–55), and Older Adults (56–90). The gap shows whether the model over- or underestimates risk for that age group, between actual and predicted rates in each group. Each group's average predicted probability adds a second, complementary way of looking at the same thing. This way is a simpler and more descriptive way than demographic equality or equalized odds. It is quite consistent with the exploratory nature of this analysis, and Chapter 6 explains more formal fairness-constrained modelling as a direction for future work.

2.4.4 Regulatory Frameworks for Fairness in European Credit Scoring

Credit scoring fairness falls within a broader legal backdrop of general anti-discrimination law in EU. 2000/43/EC and 2000/78/EC directives set equal treatment across member states, however neither applies directly to credit scoring. But they still matter, because they shape the broader legal environment that banks operate in.

The EU AI Act goes much further, treats credit scoring as high-risk, and that comes with real obligations, as discussed in Section 2.3. It should be ensured that training data actually represents the people the system will affect and doesn't quietly carry old biases. Also, the system has to keep watched once it's deployed, and act if it starts producing discriminatory outcomes. This is what this thesis carried out. Actual and predicted default rates comparison across age groups is one reasonable way of checking this problem. The Act doesn't specify any method for this, it just sets the goal, not the technique.

European supervisors have also started paying more attention to fairness in machine learning credit models, in addition to their usual focus on accuracy and capital requirements. Using

machine learning in lending leads to governance issues which are not limited to building an accurate model. This thesis tries to contribute a little to the fairness analysis.

2.5 Cross-Cultural Credit Risk Research

Credit risk research relies on data from a single country usually. And they are not compared if they hold up anywhere else. To some extent it is good, because good credit datasets are hard to find with strong credit bureau systems. But the relationship between borrower characteristics and default risk is universal enough to apply across different countries. This section questions that. It looks at research comparing credit risk modelling across countries and asks whether machine learning models can really be trusted to work the same way when tested internationally.

2.5.1 Differences in Credit Market Structure Across Countries

Across countries, credit markets differ for their institutions, data infrastructure, regulation, and cultural norms. These differences are important for how credit risk models get built and applied.

Credit bureau infrastructure is one of the important differences. Three major credit bureaus in the US hold credit information for nearly every adult in the country. That's why US models tend to be heavily on behavioral data. Payment history, credit utilization, length of credit history are important here.

Germany's SCHUFA works differently. Defaults and insolvencies are important but reporting positive payment history as fully as US bureaus is not widespread here. Instead of account balance, savings, employment, German models trust more on current financial position. The

thesis shows this in that delinquency is what drives the US model, while account status and savings are what drives the German one.

Culture plays a role too. Germany has one of the highest household savings rates in the EU. This may explain why savings variables matter there. The US has historically saved less and this is probably why savings barely show up as important in the US results.

2.5.2 Comparative Studies of Credit Risk Models Across Countries

Comparative cross-country credit risk research is surprisingly rare, considering how much it could actually matter.

Lessmann et al. (2015) is a partial exception. Their large benchmark study used credit datasets from Germany, Australia, and the US. Their main goal was finding the best-performing algorithm overall, but they also found how different methods are performed across countries. The idea was that the "best" modelling approach isn't universal. A model validated in one country shouldn't be assumed to work equally well in another without checking.

Bequé and Lessmann (2017) found something similar. When they compared extreme learning machines against methods like logistic regression and random forests, they discovered while ELMs showed promise, they were unable consistently to outperform the more established approaches. Both studies show that the best working algorithm depends on the dataset and financial environment.

Getting comparable, high-quality credit data from multiple countries is genuinely hard. It limits how much anyone can actually test if findings from one country generalise elsewhere.

2.5.3 The Role of Regulatory Differences in Cross-Cultural Model Performance

Regulation is another factor that sets countries apart, alongside structure and culture. Credit regulation is taken seriously in every developed economy. But what that regulation actually covers varies a lot by country. Which variables can lenders use? What must be disclosed about automated systems? How much capital risk can banks take on? This is what shapes how credit models get built.

Using accurate, verifiable information and avoiding irresponsible lending is required by Consumer Credit Directive and Mortgage Credit Directive in the EU. The EU AI Act goes further and says, for AI-based credit scoring, being transparence, accurate, and non-discriminative is a must in ways old, traditional statistical models were ever held to. These rules shape not only how models are built, but types of data that can be gathered to train them.

The US, governed mainly by the Equal Credit Opportunity Act and the Fair Credit Reporting Act, takes a different path. It focused on non-discrimination and data accuracy rather than AI-specific rules. Both the US and EU share the goal of accurate, fair, non-discriminatory credit decisions, but they have different specific requirements and enforcement. US credit models generally face lighter requirements around transparency and explainability, without an AI-specific framework like in the EU.

2.5.4 Implications for This Thesis

Credit markets are different from each other for their data infrastructure, culture, and regulation Those differences show up in real variation in model performance and feature importance, as Chapter 4 shows directly. US-German comparison is not just a check “does it work elsewhere

too?” It's a genuine research contribution in its own. Instead of just showing that the model behaves differently in Germany, it explains why and reasons it.

There is a simple practical takeaway. Banks and regulators can't assume a model that works well in one country, and it will behave the same way in another. It is important in practice too. European banks that increasingly deploy AI credit scoring, even though each has its own lending history and data environment.

2.6 Research Gap and Positioning of This Study

- This chapter has reviewed five thematic areas:
- The foundations of credit risk assessment
- ML methods in credit scoring
- Model interpretability and explainable AI
- Algorithmic fairness in lending
- Cross-cultural credit risk research

These define the academic landscape of this thesis. It reveals a set of gaps that shape the research design adopted here. The rest of this section identifies those gaps and positions this thesis tries to contribute to within that broader conversation.

2.6.1 The Fragmentation of the Literature

One striking pattern in this chapter's review is how rarely predictive performance, interpretability, and fairness get studied together. Papers focused on accuracy. Like the benchmark studies in Section 2.2, rarely touch interpretability or fairness. Interpretability papers using LIME or SHAP tend to treat fairness as someone else's problem. And fairness

studies like Fuster et al. (2022) and Kozodoi et al. (2022) rarely use interpretability tools to explain *why* the disparities they find actually exist.

This split makes sense. Each of these is a demanding research area on its own. But it's a real problem for banks, who can't separate the three. A model that's accurate but not interpretable fails GDPR, one that's interpretable but unfair breaks lending law, and one that's fair and interpretable but inaccurate isn't useful for managing risk in the first place. Banks need all three at once, and the literature rarely gives it to them together.

2.6.2 The Interpretability Gap

SHAP-based interpretability in credit scoring has grown a lot, as Section 2.3 covered, but a few real gaps remain.

Most studies that apply SHAP only report it at the global level. A simple ranking of mean absolute SHAP values. The beeswarm plot shows much more: not just which features matter, but which direction they push predictions and how that varies across borrowers. It's underused in literature, even though it's more informative than a bar chart. This thesis uses both.

Studies also rarely use SHAP to compare countries, something already discussed in Section 2.5.4.

Finally, SHAP findings are rarely connected to the actual regulatory context they'd need to satisfy. The fact that DelinquencyScore is the top US predictor matters practically. It's exactly what a bank would need to explain to a rejected applicant under GDPR Article 22 and the AI Act's transparency rules. This thesis makes that connection explicit, treating SHAP not just as an academic tool but as something that could actually help demonstrate regulatory compliance.

2.6.3 The Fairness Gap

As discussed in Section 2.4, age remains understudied as a fairness dimension compared to race and gender, and protection for it works differently across jurisdictions. This thesis helps fill that gap with the age-based fairness analysis in Sections 3.5.2 and 3.9.

A second, related gap is that fairness research rarely connects to interpretability. Most studies show a disparity exists without explaining why. This thesis tackles that directly, as Section 3.9.2 explains.

2.6.4 The Cross-Cultural Gap

This thesis addresses that gap through the cross-dataset SHAP comparison in Chapter 4. It looks at not just whether the model performs comparably across countries, but why connecting the feature importance differences directly to the structural differences in credit infrastructure discussed in Section 2.5. That link between empirical SHAP results and theoretical cross-cultural explanation is what sets this thesis apart from most existing cross-country studies, which typically stop at comparing performance alone.

2.6.5 Summary of Research Gaps and Thesis Positioning

This chapter has pointed to four gaps this thesis tries to address. Performance, interpretability, and fairness research rarely come together in one study. SHAP is rarely used to compare feature importance across countries, and it is applied that way, comparing US and German credit data. Age gets far less fairness research attention than race or gender, that gap was filled with a structured age-based analysis. And cross-country studies rarely explain *why* performance differs, not just that it does, the SHAP comparison has been connected to the structural differences discussed in Section 2.5.

Together, these four things place this thesis at the intersection of credit scoring, explainable AI, and fairness research, extending past the single-country focus that dominates the field. This isn't a claim to have answered every open question. The limitations, including the use of public rather than private data and a focus on just two countries, are discussed in Chapter 5. What it does offer is a grounded, practical contribution across all four areas at once. Shows how a credit model can be built to be accurate, easy to understand, fair, and properly compared across contexts. Four things' regulators and practitioners increasingly expect together, not separately.

Chapter 3: Data and Methodology

This chapter lays out exactly what data was used, what decisions were made, and how the analysis was actually carried out. It follows the same order as the real pipeline: describing the datasets, exploring them, cleaning them up, building new features, preparing the data for modelling, then training and evaluating the models. After that, it covers the SHAP interpretability work and the age-based fairness analysis.

Every decision gets explained here, not just what was done, but why. That matters, because a big part of the goal is reproducibility. Someone else should be able to pick up this chapter, use the same data, and follow the exact same steps to get the same results.

3.1 Research Design

This study uses a quantitative, supervised machine learning approach. In supervised learning, a model learns to predict a known outcome. Whether a borrower experiences serious financial distress, using a set of input variables describing each borrower. The relationship was learned by optimising on a labelled training set. Then it was evaluated on a separate test set it never saw during training, with an honest estimate to see how it would perform on new borrowers.

The primary question is: “Which model predicts credit risk best?”. This can only be answered by comparing models on real data using quantitative metrics. Feature importance, fairness, and cross-cultural comparison sub-questions need the same kind of numerical evidence, using tools like SHAP values, fairness metrics, and cross-dataset performance comparison. A qualitative approach simply couldn't produce the precise, reproducible results these questions call for.

The analysis follows seven stages: describing the datasets, exploring them, cleaning them, engineering features, preprocessing, building and evaluating the models, and finally the post-hoc analysis. Each stage builds evidence from the one before it, rather than being planned in advance before actually looking at the data. A fixed random seed (42) runs throughout, so every stochastic part of the pipeline, the train-test split, SMOTE, and training the Random Forest and XGBoost models. It gives the same results every time it's run. Everything was done in Python 3, using the libraries listed in Section 3.7, inside a Google Colab notebook that keeps a full, reproducible record of the whole process.

Two datasets are used: Give Me Some Credit from Kaggle as the primary dataset, and the German Credit Data from the UCI Machine Learning Repository as the secondary one. Using two datasets, rather than one, comes directly from the fourth research question. Whether this modelling approach performs comparably when retrained on a European dataset and reflects the cross-cultural comparison framework from Section 2.5. The full pipeline, from exploration through SHAP and fairness analysis, runs on the Give Me Some Credit data. The German data is used only for the cross-dataset comparison described in Sections 3.6.5 and 3.8, where the same Random Forest architecture is retrained and its performance and feature importance are compared against the primary dataset.

3.2 Dataset Description

3.2.1 Give Me Some Credit Dataset

The primary dataset used in this study is the Give Me Some Credit dataset (Kaggle, 2011), originally released as part of a Kaggle machine learning competition in 2011 and subsequently made publicly available through the Kaggle platform. It became one of the most widely used benchmark datasets in credit risk modelling literature. It appears in a substantial number of academic studies that use it to evaluate and compare classification algorithms for credit default prediction.

The dataset contains 150,000 anonymized borrower records from the United States. Each of them represents an individual who applied for or held a credit product during the data collection period. Each borrower has ten predictor variables and one binary target variable which is SeriousDlqin2yrs. It records whether the borrower experienced a serious delinquency event, like if the borrower fell 90 or more days behind in the following observation date. The value of one on this variable indicates that the borrower did experience serious financial distress, while a value of zero indicates that they did not. The ten predictor variables cover five broad dimensions of borrower financial behavior and characteristics credit utilisation, delinquency history, debt burden, credit exposure, and demographics and are described in detail in Table 3.1 below.

Table 3.1: Description of Variables in the Give Me Some Credit Dataset

Variable Name	Type	Description
SeriousDlqin2yrs	Binary	Target: 1 if 90+ days passed due in 2 years, 0 otherwise
RevolvingUtilizationOfUnsecuredLines	Float	Ratio of total revolving credit balance to credit limit
Age	Integer	Age of borrower in years

NumberOfTime30-59DaysPastDueNotWorse	Integer	Number of times 30-59 days late in past 2 years
DebtRatio	Float	Monthly debt payments divided by monthly income
MonthlyIncome	Float	Borrower's monthly gross income in USD
NumberOfOpenCreditLinesAndLoans	Integer	Number of open credit lines and loans
NumberOfTimes90DaysLate	Integer	Number of times 90+ days past due
NumberRealEstateLoansOrLines	Integer	Number of mortgage and real estate loans
NumberOfTime60-89DaysPastDueNotWorse	Integer	Number of times 60-89 days late in past 2 years
NumberOfDependents	Integer	Number of dependents in family excluding borrower

The dataset exhibits a number of characteristics that are typical of real-world credit data and that present specific analytical challenges. Most notably, the target variable is severely imbalanced 93.3% of borrowers in the dataset are classified as non-default while only 6.7% experienced serious delinquency within the observation period. This class imbalance reflects the real-world rarity of serious credit default events and creates a systematic bias in standard machine learning classifiers that must be addressed through appropriate preprocessing techniques, as described in Section 3.6. The dataset also contains missing values in two variables: MonthlyIncome, which is missing for 19.82% of observations, and NumberOfDependents, which is missing for 2.62% of observations and exhibits several extreme outlier values in several variables that require careful treatment during the data cleaning stage. These characteristics are discussed in detail in Sections 3.3 and 3.4.

3.2.2 German Credit Data

The secondary dataset used in this study is the German Credit Data (Hofmann, 1994), originally compiled by Professor Hans Hofmann at the University of Hamburg and made publicly available

through the UCI Machine Learning Repository. It has been used widely in the credit scoring literature as a benchmark for evaluating classification algorithms. For this study, an extended version of the dataset was obtained from an open-source repository. The dataset contains 21 variables, including the original features and a binary credit risk target. It was used for cross-cultural comparison.

The dataset contains 1,000 borrower records from a German financial institution. Each of them represents an individual who applied for credit during the data collection period. The binary target variable, credit risk takes a value of one for borrowers classified as good credit risks and zero for those classified as bad credit risks. The target variable distribution is considerably more balanced than in the Give Me Some Credit dataset. 70% of borrowers are classified as having good credit risks and 30% as bad. It is a reflection of the different lending environment and risk profile of the German dataset. The 20 predictor variables cover a broader range of borrower characteristics than the US dataset, including not only financial variables, such as account status, credit amount, loan duration, and savings level, but also demographic and social variables including age, employment duration, housing status, personal status, and the purpose of the credit application. A summary of the key variables in the German Credit Data is provided in Table 3.2 below.

Table 3.2: Key Variables in the German Credit Data

Variable Name	Type	Description
Credit_risk	Binary	Target: 1 = good credit risk, 0 = bad credit risk
Status	Categorical	Status of existing checking account
Duration	Integer	Duration of credit in months
Credit_history	Categorical	Credit history of borrower
Purpose	Categorical	Purpose of the credit application

Amount	Integer	Credit amount requested
Savings	Categorical	Savings account or bonds balance
Employment_duration	Categorical	Duration of current employment
Age	Integer	Age of borrower in years
Housing	Categorical	Housing situation of borrower
Job	Categorical	Job type and skill level
Personal_status_sex	Categorical	Personal status and sex of borrower

An important difference between the two datasets that must be acknowledged is the presence of a large number of categorical variables in the German Credit Data variables whose values are text labels rather than numbers, such as account status categories and employment duration brackets. These categorical variables cannot be used directly as inputs to machine learning classifiers, which require numerical inputs, and must be converted to numerical form through an encoding procedure before model training. The label encoding approach used in this study to handle categorical variables in the German dataset is described in Section 3.6.5. A further important difference is that the German Credit Data contains no missing values, eliminating the need for imputation procedures that are required for the US dataset.

Comparing the modelling approach on US and European data serves a purpose beyond simple robustness testing. It tests whether a model used outside the EU regulatory environment could meet the transparency, fairness, and data governance expectations that would apply if used within it.

3.3 Exploratory Data Analysis

Exploratory data analysis was conducted as the first stage of the analytical pipeline, prior to any data cleaning or transformation, in order to develop a thorough understanding of the structure,

distribution, and quality of the raw data before making any analytical decisions. The exploratory analysis focused on four key aspects of the dataset the distribution of the target variable, the distributions of the predictor variables, the presence and pattern of missing values, and the correlations between predictor variables and the target. The findings of the exploratory analysis directly informed the data cleaning decisions described in Section 3.4 and the feature engineering decisions described in Section 3.5.

Target Variable Distribution

The distribution of the target variable SeriousDlqin2yrs revealed a severe class imbalance that represents one of the most significant analytical challenges in the dataset. Of the 150,000 borrowers in the raw dataset, 139,974 or 93.3% were classified as non-default, while only 10,026 or 6.7% experienced serious financial distress within the observation period. This imbalance is consistent with the real-world rarity of serious credit default events and with the patterns reported in other studies using this dataset. As discussed in Section 2.2.4, severe class imbalance of this magnitude creates a systematic bias in standard machine learning classifiers, which tend to optimise for overall accuracy by effectively ignoring the minority class. The decision to address this imbalance through the application of SMOTE during the preprocessing stage described in Section 3.6 was directly motivated by this exploratory finding.

Predictor Variable Distributions

Examination of the distributions of the ten predictor variables revealed several important characteristics that informed subsequent data cleaning decisions. The age variable exhibited a broadly bell-shaped distribution centered around 52 years, consistent with the demographic profile of the US adult borrower population, but with a small number of extreme values, including one observation with an age of zero and 489 observations with ages above 90 that were judged to be

data quality issues rather than genuine observations and were removed during the cleaning stage. The three delinquency variables were heavily right skewed. While most of the borrowers had zero past-due events, a small number showed clearly impossible values (like being late 96 or 98) separate times within a two-year window. They were identified as data entry errors. RevolvingUtilizationOfUnsecuredLines had a similar issue. Since it measures the ratio of credit used to credit available, it should normally be between 0 and 1, but 3,321 rows showed values above 1. They were likely reflecting either data errors or rare cases where a borrower genuinely exceeded their credit limit. The MonthlyIncome variable exhibited extreme positive skewness, with a small number of borrowers recording monthly incomes more than one million dollars that were judged to be outliers requiring treatment.

Missing Value Analysis

The missing value analysis confirmed the presence of missing data in two variables. MonthlyIncome was missing for 29,731 observations representing 19.82% of the total dataset, a proportion that is substantial enough to require careful treatment but not so large as to make the variable unusable. NumberOfDependents was missing for 3,924 observations representing 2.62% of the dataset, a much smaller proportion that is more straightforwardly addressed through standard imputation. The pattern of missing values in MonthlyIncome appeared to be broadly random rather than systematically related to the default status of the affected borrowers, supporting the use of median imputation rather than more complex imputation approaches that would be warranted if the missing data mechanism were more informative.

Correlation Analysis

The correlation analysis revealed that the three delinquency variables NumberOfTime30-59DaysPastDueNotWorse, NumberOfTimes90DaysLate, and NumberOfTime60-

89DaysPastDueNotWorse exhibited the strongest positive correlations with the target variable, with correlation coefficients of 0.126, 0.117, and 0.102 respectively. This finding confirmed the well-established importance of past payment behaviour as a predictor of future credit default and provided empirical support for the decision to construct a Weighted Delinquency Score that combines all three delinquency variables into a single composite measure. Age showed the strongest negative correlation with the target variable, with a coefficient of -0.115. It confirms that older borrowers are less likely to default and support the need for age-based fairness analysis. The remaining variables showed weaker correlations with the target, though all followed the expected direction based on credit risk theory. Higher debt ratios, more dependents, and lower incomes were associated with higher default risk, while more open credit lines and higher real estate loan counts were associated with lower risk.

3.4 Data Cleaning

Data cleaning is one of the most important stages of any machine learning pipeline, as the quality of the data used for model training directly determines the reliability and validity of the resulting predictions. Raw datasets from real-world sources often contain errors, inconsistencies, and missing values that must be identified and handled before any analytical work can proceed.

This section describes the specific cleaning procedures applied to the Give Me Some Credit dataset. All cleaning procedures were applied to the raw dataset before any feature engineering or model training, and the cleaned dataset was saved as a separate file to ensure that the cleaning process could be repeated and checked independently of the later analysis.

3.4.1 Removal of Duplicate Records

The first cleaning step was the identification and removal of duplicate records observations that were identical across all variables and therefore likely represented the same borrower recorded

multiple times due to data entry errors or system processing issues. Examination of the raw dataset revealed 609 duplicate records, representing 0.41% of the total observations. All duplicate records were removed, which reduced dataset from 150,000 to 149,391 observations. While the proportion of duplicates was relatively small, their removal was nevertheless important. Duplicate records can artificially increase the effective sample size and give repeated observations more influence during model training, which may lead to overfitting. As a result, the model may learn patterns specific to repeated records rather than general patterns in the borrower population.

3.4.2 Treatment of Age Outliers

The second cleaning step addressed the extreme outlier values identified in the age variable during the exploration data analysis. As noted in Section 3.3, the age variable contained one observation with an age of zero and 489 observations with ages above 90. Age values below 18 are implausible in a consumer credit dataset no legitimate borrower below the age of majority would be expected to appear in a credit scoring dataset and were removed on the grounds that they represent data entry errors rather than genuine observations. Age values above 90 were also excluded, reflecting the judgement that borrowers of such advanced age are sufficiently rare in the consumer credit population. Including them in the training data would as noise rather than signal, and that the age outlier values of up to 109 years found in the raw data were likely caused by data entry errors rather than real observations. The removal of age values below 18 and above 90 eliminated a total of 478 observations, reducing the dataset to 148,913 records.

The decision to cap the age variable at 90 rather than at a lower threshold such as 80 was made to keep as much valid data as possible while removing only the clearly unrealistic values. The age distribution of the remaining dataset, which ranged from 21 to 90 years with a mean of about 52 years, is consistent with the expected profile of the US adult borrower population and suggests that

the cleaning threshold was appropriate. This cleaning decision is documented explicitly in the thesis because the age variable subsequently plays an important role in the feature engineering and fairness analysis stages of the pipeline, and the boundaries of the age distribution directly determine the composition of the three age groups used in the fairness analysis.

3.4.3 Treatment of Revolving Utilization Outliers

The RevolvingUtilizationOfUnsecuredLines variable which represents the ratio of total revolving credit balance to total credit limit should theoretically take values between zero and one, where zero indicates that no revolving credit is currently being used, and one indicates that all available revolving credit is fully utilised. Values above one mean that the borrower has gone over their credit limit, which is possible in practice but should be rare and not very large. The raw dataset showed 3,321 observations with RevolvingUtilizationOfUnsecuredLines values greater than one, including some values of several thousand that are clearly data errors rather than real cases of exceeding the credit limit. All observations with RevolvingUtilizationOfUnsecuredLines values greater than one were removed, which reduced the dataset to 145,592 observations. This threshold was chosen because values above one is definitionally outside the valid range of a credit utilisation ratio and cannot be interpreted as meaningful measures of credit behaviour, regardless of whether they resulted from data entry errors or system processing issues.

3.4.4 Capping of Delinquency Variable Outliers

The three delinquency count variables NumberOfTime30-59DaysPastDueNotWorse, NumberOfTime60-89DaysPastDueNotWorse, and NumberOfTimes90DaysLate contained extreme values of 96 and 98 that were identified during the exploratory data analysis as likely entry errors. Over the two-year observation period, a borrower could be late on monthly payments a maximum of approximately 24 times. Which makes values of 96 or 98 clearly unrealistic as

actual delinquency counts. Instead of removing the observations entirely, which would have resulted in the loss of many records, a capping approach was used. All delinquency count values are set above 10 to 10. This threshold was chosen because values up to 10 are possible within a two-year observation period for borrowers with multiple credit relationships. While values above this level are high enough to suggest data quality issues. The capping approach preserves the information that these borrowers have a history of serious delinquency while removing the effect of the extreme values on model training. After capping, the maximum value of each delinquency variable was 10, within the expected range of genuine delinquency counts.

3.4.5 Imputation of Missing Values

Two variables had missing values: MonthlyIncome (missing in 29,731 rows, or 19.82% of the dataset) and NumberOfDependents (missing in 3,924 rows, or 2.62%). Both were filled in using the median rather than the mean, because a handful of very high earners were pulling the average income up well beyond what a typical borrower actually earns. The median isn't affected by outliers this way, so it gives a more realistic stand-in value. After this step, the dataset had no missing values left anywhere.

This imputation was made on the full dataset before the train-test split in Section 3.6, not on the training set alone. As noted in Section 3.6.2, that's a small deviation from best practice for avoiding data leakage but with a dataset this large, the median barely would have changed if calculated on the training set alone.

3.4.6 Capping of Monthly Income and Debt Ratio Outliers

After imputation, MonthlyIncome and DebtRatio still had extreme outliers that could distort the model. Some borrowers showed monthly incomes near three million dollars, not impossible, but

far enough outside normal range that a handful of these values could push the model to learn patterns that don't actually apply to real borrowers. DebtRatio had a similar problem. Some values ran into the hundreds of thousands, which doesn't make sense for a ratio. These almost certainly came from cases where monthly income was close to zero, making the ratio blow up toward infinity.

Both variables were capped at their 99th percentile: $23,250 for MonthlyIncome and 4,999 for DebtRatio. Rather than deleting these rows, any value above the cap was simply replaced with the cap value, a technique called Winsorisation. This keeps the full sample size while still limiting how much extreme values can skew the model, which is standard practice when the number of affected rows is small.

3.4.7 Summary of Data Cleaning Results

Table 3.3 below summarizes the effect of each cleaning step on the dataset and shows the number of observations removed or modified.

Cleaning Step	Action	Observations Affected
Duplicate removal	Removed	609 deleted
Age outliers	Removed	478 deleted
Revolving utilization	Removed	3,321 deleted
Delinquency capping	Capped at 10	Values modified
Missing value imputation	Median imputation	33,655 imputed
Income and debt capping	Winsorised at 99th pct	Values modified

The final cleaned dataset contains 145,592 observations representing 97.1% of the original raw dataset and is complete across all variables with no remaining missing values. The cleaning process removed a total of 4,408 observations, or 2.9% of the original dataset, a proportion that is

sufficiently small to avoid introducing meaningful sample selection bias while effectively addressing the data quality issues identified during the exploratory analysis.

3.5 Feature Engineering

Feature engineering means building new variables from existing data, using domain knowledge to improve performance, interpretability, or to support something the raw variables can't do alone. For credit risk specifically, that means combining two things: statistical insight into which transformations might capture useful patterns, and banking knowledge about what actually drives default. This section covers the two features built for this study, the reasoning behind each, how they were constructed, and what they're expected to contribute.

Both features come entirely from variables already in the Give Me Some Credit dataset, no outside data was added. So, this isn't new information, just a different way of organizing what was already there. Both were built after the cleaning stage in Section 3.4 and before preprocessing in Section 3.6, so they're based on clean data and go through the same train-test split and SMOTE steps as everything else.

3.5.1 Weighted Delinquency Score

The first engineered feature is the Weighted Delinquency Score, a single measure combining the three separate delinquency variables already in the dataset (30-59 days late, 60-89 days late, and 90+ days late) into one severity-weighted score. Each of these three variables is predictive on its own, as Section 3.3's correlation analysis showed, but treating them as three separate, unrelated predictors misses something credit analysts already know: delinquency history isn't just about *how often* someone was late, it's about *how late*.

The formula for the score:

$$\text{DelinquencyScore} = (\text{30-59 days late} \times 1) + (\text{60-89 days late} \times 2) + (\text{90+ days late} \times 3)$$

The weight increases with severity. 90+ day delinquency counts three times more than a 30-59 day one. This is similar how traditional frameworks like FICO treat delinquency, and it also reflects my own experience in banking. A borrower who's been late once for a short period once and a borrower who's repeatedly gone 90+ days aren't just "somewhat different", they're different risk profiles.

A borrower with a clean payment history scores zero. Someone who's been 30-59 days late twice and 90+ days late once scores $(2\times1) + (1\times3) = 5$. Most borrowers in the cleaned dataset score zero, matching the low overall default rate, with a shrinking number scoring increasingly higher. The maximum possible score is 60, a borrower who reaches a cap of 10 in all three delinquency categories.

Both the composite score and the three original variables were kept in the model, instead of replacing the original variables. This makes it possible to check later. Whether the composite score actually captures something the individual variables don't or whether it's just representing the same information in another form. That comparison is made using SHAP, in Chapter 4.

3.5.2 Age Group

The second engineered feature is Age Group, a categorical variable created by binning the continuous age variable into three groups: Young Adults (18–35), Middle-Aged (36–55), and Older Adults (56–90).

This exists to make the age-based fairness analysis in Section 3.9 possible. A continuous age variable doesn't work well for group-level fairness comparisons, since there's no natural way to split "age" into discrete groups you can compare actual versus predicted default rates across. Age

Group solves that: it gives the fairness analysis clean categories to work with, so the model's predicted default rate in each age band can be compared directly against the real default rate in the test data. The continuous age variable still stays in the predictive model itself, Age Group's job here is purely analytical, not predictive.

The boundaries, 35 and 55, were chosen to keep the groups large enough for meaningful comparison, while roughly matching the life-stage differences in financial behaviour that show up in consumer finance research. These aren't fixed, universally agreed-upon cutoffs, a different choice of boundaries would produce different groups, and this is flagged in Chapter 6 as worth testing in future work.

The three groups came out reasonably well-sized: Middle-Aged borrowers were the largest group (64,702, or 44.4%), Older Adults close behind (60,608, or 41.6%), and Young Adults the smallest (20,282, or 13.9%). Young Adults being underrepresented lines up with how the US credit market actually looks. Younger borrowers tend to have shorter credit histories and hold fewer credit products and it's worth keeping in mind when interpreting the fairness results in Chapter 4.

3.5.3 Summary of Engineered Features

Table 3.4 below summarizes the two engineered features constructed for this study, providing a concise reference for the variables introduced at this stage of the analytical pipeline.

Table 3.4: Summary of Engineered Features

Feature Name	Type	Construction Method	Analytical Purpose
DelinquencyScore	Continuous	(30-59 days × 1) + (60-89 days × 2) + (90+ days × 3)	Capture severity-weighted delinquency history

AgeGroup	Categorical	Binning of age into 3 categories defined for this study	Enable demographic fairness analysis

Following the construction of the two engineered features, the dataset contained a total of 13 variables the 11 original variables from the cleaned dataset plus the two engineered features and 145,592 observations. This enriched dataset was used as the input to the preprocessing stage described in the following section.

3.6 Preprocessing

Preprocessing refers to the set of procedures applied to the cleaned and feature-engineered dataset to prepare it for machine learning model training. The preprocessing stage transforms the data into a format that is compatible with the requirements of the chosen classification algorithms, addresses the class imbalance problem identified during the exploratory data analysis, and establishes the train-test split that allows model performance to be evaluated on data that was not used during training. This section describes the three preprocessing steps applied in this study feature and target definition, train-test splitting, and class imbalance treatment through SMOTE explaining the rationale for each decision and documenting the specific parameters used.

3.6.1 Feature and Target Definition

The first preprocessing step was the definition of the feature matrix and the target vector that would be used for model training and evaluation. The target vector was defined as the SeriousDlqin2yrs column, the binary indicator of serious financial distress and was separated from the feature matrix prior to any further preprocessing. The feature matrix was defined as all remaining variables in the dataset with the exception of the AgeGroup variable, which was excluded from model training for

two reasons. First, AgeGroup is a categorical text variable whose information is already captured by the continuous age variable that remains in the feature matrix including both would introduce redundancy. Second, and more importantly, AgeGroup is used as the grouping variable for the fairness analysis described in Section 3.9, and its role in the analytical pipeline is fundamentally different from that of the predictor variables it is an analytical lens through which model predictions are examined rather than an input from which predictions are generated. The final feature matrix therefore contained 11 variables the 10 original predictor variables from the cleaned dataset plus the Weighted Delinquency Score providing a rich and balanced set of inputs that captured all major dimensions of borrower financial behaviour represented in the dataset.

3.6.2 Train-Test Split

The dataset was divided into a training set and a test set using a stratified random split, with 80% of observations allocated to training and the remaining 20% to testing. This split ratio is consistent with standard practice in machine learning research for datasets of this size the 80% training allocation provides a sufficiently large sample for model training while the 20% test allocation provides a statistically robust basis for performance evaluation. The stratified splitting procedure ensures that the proportion of defaulters and non-defaulters in both the training and test sets matches the proportion in the full dataset approximately 6.7% defaulters and 93.3% non-defaulters preventing the split from inadvertently concentrating defaulters in one set at the expense of the other and ensuring that the class distribution faced by the model during training is representative of that encountered during evaluation.

The split used a fixed random seed of 42, so that the observations allocated to the training and test sets can be reproduced in different runs of the analysis. Using a fixed random seed is standard practice in reproducible machine learning and is particularly important in this study, where the

fairness analysis requires that the AgeGroup labels for test set observations can be matched to their model predictions. The resulting training set contained 116,473 observations and the test set contained 29,119 observations both of sufficient size to support robust model training and evaluation respectively.

It is critical to emphasise that the test set was held out completely from model training. Data leakage where information from the test set inadvertently influences model training is one of the most common and consequential errors in applied machine learning, so care was taken to avoid it wherever possible. SMOTE and feature scaling, described in the following subsections, were both fitted using training data only. One caveat worth noting: the median imputation and outlier capping applied during cleaning (Section 3.4) were calculated on the full cleaned dataset before the train-test split, rather than on the training set alone. Given the size of the dataset, this is unlikely to have meaningfully shifted the median or percentile values used, but it is acknowledged here as a minor limitation of the preprocessing pipeline.

3.6.3 Handling Class Imbalance with SMOTE

As established during the exploratory data analysis and discussed in Section 2.2.4, the Give Me Some Credit dataset exhibits severe class imbalance 93.3% of observations in the training set are classified as non-default while only 6.7% are classified as default. This imbalance poses a fundamental challenge for machine learning classifiers, which tend to optimise for overall accuracy by learning to predict the majority class for all or most observations, effectively ignoring the minority class. In the context of credit risk modelling, where the primary objective is precisely to identify the minority of borrowers who will default, a classifier that fails to learn the characteristics of defaulters is of no practical value regardless of its overall accuracy.

To address this problem, the Synthetic Minority Over-sampling Technique SMOTE was applied to the training set prior to model training. As described in Section 2.2.4, SMOTE generates synthetic new observations of the minority class by creating new points between existing minority class observations in the feature space. These artificial records that expand the minority class distribution and encourage the classifier to learn more general patterns of default behavior instead of memorizing the characteristics of few genuine defaulters in the training data. The SMOTE procedure was implemented using the imbalanced-learn library in Python, with default parameters and a fixed random seed of 42 for reproducibility.

After the application of SMOTE, the training set contained 218,928 observations 109,464 non-defaulters and 109,464 defaulters representing a perfectly balanced class distribution. This balanced training set was used for all model training procedures described in Section 3.7. The test set was left in its original imbalanced state 27,367 non-defaulters and 1,752 defaulters reflecting the true class distribution in the borrower population and ensuring that model performance metrics computed on the test set accurately reflect the expected performance of the model in real-world deployment, where defaulters represent a small minority of all borrowers. The importance of this distinction between the balanced training set used for learning and the imbalanced test set used for evaluation cannot be overstated, as applying SMOTE to the test set would produce artificially inflated performance metrics that do not reflect genuine model capability.

3.6.4 Feature Scaling for Logistic Regression

An additional preprocessing step was required specifically for the Logistic Regression model the standardization of the feature matrix through z-score scaling. Unlike tree-based models such as Random Forest and XGBoost, which are not affected by the scale of input features because they make decisions based on ranked split points rather than absolute feature values, Logistic

Regression is sensitive to feature scale. Because it uses a gradient-based objective function in which features with larger absolute values can dominate the gradient updates and slow convergence. Standardization addresses this problem by changing each feature to have a mean of zero and a standard deviation of one, placing all features on a common scale that allows the gradient optimisation to converge more efficiently.

The standardization was implemented using the StandardScaler from the scikit-learn library. It is what was fitted only on the training data calculating the mean and standard deviation of each feature from the training observations and applied to both the training and test sets. This approach ensures that test-set information doesn't affect the scaling parameters. Preventing a subtle form of data leakage that that could occur if the scaler were fitted on the combined training and test data. The scaled training and test sets were used exclusively for Logistic Regression training and evaluation, while the unscaled versions were used for Random Forest and XGBoost, which do not require feature scaling.

3.6.5 German Dataset Preprocessing

The German Credit Data required a separate preprocessing procedure from the US dataset, primarily because it contains a large number of categorical predictor variables rather than the mostly numeric variables found in the Give Me Some Credit dataset. Categorical variables such as account status, savings level, and employment duration were converted to numerical representations using label encoding, which assigns each category a distinct integer value.

The target variable was provided as a binary indicator, with a value of 1 representing a good credit risk and 0 representing a bad credit risk, consistent with the binary coding convention used throughout this thesis for the US target variable. The German data contained no missing values, so no imputation step was required. The dataset was then divided into training and test sets using

an 80/20 stratified split with the same fixed random seed of 42 used throughout this study, and SMOTE was applied to the German training set only, following the same procedure and rationale described in Section 3.6.3 for the US data. The test set was left untouched to preserve its original class distribution for evaluation.

It is worth noting a limitation of this approach: label encoding assigns arbitrary numerical values to categories that have no inherent order, and applying SMOTE to label-encoded categorical variables can generate synthetic observations with interpolated values that don't correspond to any real category. A more robust approach for future work would use encoding and resampling methods designed specifically for mixed categorical and numerical data, such as one-hot encoding combined with SMOTENC.

3.6.6 Summary of Preprocessing Results

Table 3.5 below summarises the key parameters and outcomes of the preprocessing stage, providing a concise reference for the data configuration used in the model training and evaluation procedures described in the following section.

Table 3.5: Summary of Preprocessing Parameters and Outcomes

Preprocessing Step	Parameter / Outcome
Feature matrix variables	11 (10 original + DelinquencyScore)
Target variable	SeriousDlqin2yrs (binary)
Train-test split ratio	80% training / 20% test
Random seed	42
Training set size (before SMOTE)	116,473 observations
Test set size	29,119 observations
SMOTE applied to	Training set only

Training set size (after SMOTE)	218,928 observations (balanced)
Class distribution after SMOTE	50% default / 50% non-default
Feature scaling applied to	Logistic Regression only
Scaling method	Z-score standardisation (StandardScaler)

3.7 Model Selection and Development

This section describes the three machine learning classification models developed and compared in this study Logistic Regression, Random Forest, and XGBoost explaining the theoretical basis of each model, the rationale for its inclusion in the comparison, and the specific implementation parameters used in this study. The selection of these three models reflects a deliberate decision to span the full spectrum of complexity and interpretability available in the credit scoring literature, from a transparent and well-established statistical baseline through to more complex ensembles that are widely used in applied credit risk modelling. Together the three models allow the study to assess not only which approach achieves the highest predictive performance on the specific datasets used but also whether the gains in accuracy associated with increasing model complexity are sufficient to justify the associated costs in terms of interpretability and ease of deployment in regulated financial environments.

All three models were implemented using the scikit-learn and xgboost libraries in Python, with a fixed random seed of 42 applied to all models that involve stochastic elements specifically Random Forest and XGBoost to ensure reproducibility of results. The models were trained on the SMOTE-oversampled training set described in Section 3.6 and evaluated on the held-out test set, with Logistic Regression additionally using the standardised versions of the training and test sets as described in Section 3.6.4. No hyperparameter tuning was performed in this study all models were trained using default or near-default parameter settings, with the exception of the maximum

iteration limit for Logistic Regression which was increased from the default of 100 to 2,000 to ensure convergence on the scaled training data. This decision reflects the scope of the study, which priorities the comparison of model architectures and the application of interpretability and fairness analysis tools over the optimisation of individual model performance, and is consistent with the approach adopted in benchmark studies such as Lessmann et al. (2015) that use standardised parameter settings to ensure comparability across models.

3.7.1 Logistic Regression

Logistic Regression is a statistical classification method that models the probability of a binary outcome in this case, whether a borrower will experience serious financial distress as a function of a set of predictor variables using the logistic function. The logistic function, also known as the sigmoid function, maps any real-valued input to a value between zero and one, allowing the model to produce outputs that can be directly interpreted as probabilities of default. Formally, the probability of default for borrower i is modelled as:

$$P(Yi = 1 \mid Xi) = 1 / (1 + exp(-(\beta 0 + \beta 1 X1i + \beta 2 X2i + ... + \beta pXpi)))$$

where Yi is the binary default indicator for borrower i, Xi is the vector of predictor variable values for borrower i, and β0, β1, ..., βp are the model coefficients estimated from the training data by maximising the log-likelihood of the observed outcomes. Each coefficient βj represents the change in the log-odds of default linked to a one-unit increase in the corresponding predictor variable Xj, holding all other variables constant. This direct interpretability of model coefficient which allows the analyst to immediately identify the direction and relative magnitude of each variable's effect on default probability is one of the most important practical advantages of Logistic Regression in credit scoring contexts where the ability to explain model decisions is a regulatory requirement.

Logistic Regression was included in this study as a transparent baseline model against which the performance of more complex approaches could be assessed. Its long history of application in credit scoring dating back to the foundational work reviewed in Section 2.1. And its widespread use in current industry practice makes it the natural reference point for any comparative credit risk study. If more complex models such as Random Forest and XGBoost achieve substantially higher performance than Logistic Regression, this provides evidence that the non-linear patterns they are capable of capturing are genuinely present in the data and predictively relevant. If the performance differences are modest, this suggests that a simpler and more interpretable model may be preferable in practice particularly in regulated environments where the transparency advantages of Logistic Regression carry significant practical value.

The Logistic Regression model was implemented using the LogisticRegression class from scikit-learn, with the maximum number of iterations set to 2,000 to ensure convergence and all other parameters set to their default values. The model was trained on the standardised and SMOTE-oversampled training set and evaluated on the standardised test set, with predictions generated as both binary class labels and continuous predicted probabilities to enable the computation of all evaluation metrics described in Section 3.8.

3.7.2 Random Forest

Random Forest is an ensemble classification method that builds its predictions by aggregating the outputs of a large number of individual decision trees. Each is trained on a different random bootstrap sample of the training data and uses a random subset of the available features at each split point. The ensemble aggregation process combining the predictions of many individually weak learners into a single strong collective prediction dramatically reduces the variance of

predictions compared to any individual tree, producing a model that is substantially more stable and better able to generalise to new data than a single decision tree trained on the full dataset.

The Random Forest model was implemented using the RandomForestClassifier class from scikit-learn, with the number of estimators individual trees in the ensemble set to 100 and all other parameters set to their default values. The model was trained on the unscaled SMOTE-oversampled training set feature scaling is not required for tree-based models and evaluated on the unscaled test set. The Random Forest model is also the model selected for the SHAP interpretability analysis described in Section 3.8, based on its superior performance on the F1 score metric compared to the other two models a selection criterion that reflects the practical importance of correctly identifying actual defaulters in a credit risk context, as discussed in Section 3.8.

3.7.3 XGBoost

XGBoost Extreme Gradient Boosting is a highly optimised implementation of the gradient boosting framework that has achieved state-of-the-art performance across a wide range of machine learning classification tasks. Unlike Random Forest, which builds its ensemble by training trees independently in parallel on random subsets of the data, XGBoost builds trees sequentially each new tree is trained specifically to correct the prediction errors of the trees that preceded it, focusing the model's learning capacity on the observations that are most difficult to classify correctly. This sequential error-correction mechanism, combined with a regularisation term in the objective function that penalises model complexity and prevents overfitting, allows XGBoost to achieve very high predictive accuracy while maintaining reasonable generalisation to new data.

The XGBoost model was implemented using the XGBClassifier class from the xgboost library, with the evaluation metric set to log loss and all other parameters set to their default values. The

model was trained on the unscaled SMOTE-oversampled training set and evaluated on the unscaled test set, with the same random seed of 42 applied to ensure reproducibility.

3.7.4 Model Evaluation Framework

The performance of all three models was evaluated using a consistent set of metrics applied to the held-out test set, chosen to provide a comprehensive and nuanced picture of model performance that goes beyond the limitations of overall accuracy in the presence of severe class imbalance. Four evaluation metrics were computed for each model AUC-ROC, F1 Score, Precision, and Recall each of which captures a different dimension of classification performance and together provide a multi-faceted basis for model comparison.

The Area Under the Receiver Operating Characteristic Curve AUC-ROC is the primary evaluation metric used in this study. The ROC curve plots the true positive rate the proportion of actual defaulters correctly identified by the model against the false positive rate the proportion of actual non-defaulters incorrectly classified as defaulters across all possible classification thresholds. The area under this curve provides a single scalar measure of the model's ability to discriminate between defaulters and non-defaulters across the full range of possible threshold settings, with a value of one representing perfect discrimination and a value of 0.5 representing performance no better than random chance. AUC-ROC is well suited as the primary metric for this study because it is insensitive to class imbalance unlike accuracy, which can be misleadingly high when one class dominates and because it evaluates model performance across all possible threshold settings rather than at a single fixed threshold, providing a more complete picture of discriminatory power.

The F1 Score is a harmonic mean of Precision and Recall that provides a single balanced measure of model performance on the minority class defaulters that is directly relevant to the practical objective of credit risk modelling. Precision measures the proportion of borrowers predicted as

defaulters who defaulted a high precision model produces few false alarms but may miss many actual defaulters. Recall measures the proportion of actual defaulters who were correctly identified by the model a high recall model catches most defaulters but may also flag many non-defaulters incorrectly. The F1 Score balances these two competing objectives, rewarding models that achieve both high precision and high recall on the minority class. In the context of credit risk, both precision and recall are important false positives represent borrowers who are incorrectly denied credit, while false negatives represent defaulters who are incorrectly approved, and neither type of error is costless. The F1 Score provides a single metric that captures the trade-off between these two error types and is used as the secondary evaluation criterion and as the basis for selecting the best model for the SHAP interpretability analysis.

3.8 SHAP Interpretability Framework

One of the central arguments of this thesis is that predicting credit default accurately is not enough on its own. A bank that deploys a machine learning model to make lending decisions needs to be able to explain those decisions to regulators who require transparency, to auditors who need to verify that the model is operating correctly, and to borrowers who have the right to understand why their application was declined. A model that achieves an AUC of 0.82 but operates as a complete black box is of limited practical value in a regulated banking environment, regardless of how impressive its predictive performance might appear on paper. This conviction that interpretability is not an optional extra but a core requirement of responsible credit risk modelling motivated the decision to make SHAP analysis a central component of the research design rather than a supplementary step added after the main results had already been produced.

SHAP was introduced by Lundberg and Lee (2017) as a unified framework for explaining the predictions of any machine learning model. The method draws on the concept of Shapley values

from cooperative game theory, originally developed by Lloyd Shapley in 1953 as a mathematically principled solution to the problem of fairly distributing the collective payoff of a cooperative game among its individual players. The key insight that Lundberg and Lee applied to machine learning was recognising that the problem of attributing a model's prediction to its input features is structurally analogous to the problem of distributing a game's payoff among its players each feature can be thought of as a player whose contribution to the prediction varies depending on which other features are present alongside it.

In practical terms for this thesis, SHAP analysis answers the question that the model performance metrics alone cannot not just whether the Random Forest model correctly predicted that a borrower would default, but which specific variables drove that prediction and by how much. For a borrower who was assigned a high predicted default probability, SHAP tells us whether that high-risk score was driven primarily by their delinquency history, their revolving credit utilisation, their income level, or some combination of these factors. This kind of variable-level explanation is precisely what regulators under GDPR and the EU AI Act require when they mandate that individuals be given meaningful explanations of automated credit decisions that affect them.

3.8.1 How SHAP Values Work in This Study

For each borrower in the test set, SHAP computes a value for every input feature that represents that feature's contribution to pushing the predicted default probability above or below the average prediction across all borrowers. A positive SHAP value for a given feature means that the feature pushed the prediction towards default increasing the predicted probability of financial distress. A negative SHAP value means the feature pushed the prediction away from default, reducing the predicted probability. The sum of all SHAP values for a given borrower, plus the baseline prediction, equals that borrower's actual predicted probability of default.

To make this concrete in the context of this thesis consider a Young Adult borrower who was assigned a high predicted default probability by the Random Forest model. Their SHAP values might show that their high Weighted Delinquency Score contributed +0.25 to their predicted probability, their high revolving credit utilisation contributed +0.15, and their relatively low monthly income contributed +0.08, while their small number of dependents contributed -0.05. The SHAP analysis makes the model's reasoning transparent and specifically this borrower was classified as high risk primarily because of their delinquency history and credit utilisation behaviour, not because of their age or income alone. This is exactly the kind of explanation that a credit officer would need to provide to a rejected applicant and that a regulator would expect to see documented in a model governance file.

3.8.2 Global and Comparative Application

In this study SHAP analysis is applied at two levels global and comparative reflecting the two interpretability objectives of the thesis.

At the global level, SHAP values are aggregated across all 500 randomly sampled test set observations to produce a feature importance profile for the Random Forest model on the US dataset. Two visualizations are used to communicate these global findings. The beeswarm plot shows the distribution of SHAP values for each feature across all sampled observations, with each dot representing one borrower, the horizontal position showing the magnitude and direction of the SHAP value, and the color showing the raw feature value red for high values and blue for low values. This plot provides a rich picture of not only which features matter most overall but how they interact with the target outcome for example, whether high values of a feature consistently push predictions towards default or whether the relationship is more complex. The bar plot of mean absolute SHAP values provides a cleaner summary of overall feature importance, ranking features

by their average absolute contribution to predictions across all observations and making it straightforward to identify which variables the model relies on most heavily.

At the comparative level, the same SHAP analysis is repeated on the German Credit Data using the Random Forest model retrained on that dataset. The resulting feature importance profile is then compared directly with the US profile, providing empirical evidence on whether the same variables drive credit risk predictions in both financial contexts or whether fundamentally different risk factors are at work. This cross-dataset SHAP comparison is one of the distinctive elements of the analytical framework developed in this thesis and goes beyond what most credit risk studies have attempted using interpretability tools. Rather than using SHAP simply to explain one model on one dataset, this study uses it as an analytical instrument for investigating substantive differences in the drivers of credit risk across two national credit environments a use case that connects directly to the cross-cultural research gap identified in Section 2.6.

3.8.3 Implementation Details

The SHAP analysis was implemented using the shap library in Python, which provides an efficient implementation of the TreeSHAP algorithm developed by Lundberg et al. (2020) for computing exact Shapley values for tree-based ensemble models. TreeSHAP exploits the tree structure of Random Forest and XGBoost models to compute exact SHAP values in polynomial rather than exponential time, making it computationally feasible to apply SHAP analysis to large ensemble models without resorting to the approximations required by model-agnostic explanation methods such as LIME. For the US dataset, SHAP values were computed on a random sample of 500 observations from the test set a sample size that provides a representative picture of global feature importance while keeping computation time manageable within the Google Colab environment

used for the analysis. For the German dataset, SHAP values were computed on all 200 test set observations, which constituted the full test set for that dataset.

The TreeExplainer class from the Shap library was used to create the SHAP explainer object for the Random Forest model, and the resulting SHAP values were extracted for the positive class default to ensure that all SHAP values are interpreted in terms of their contribution to the probability of default rather than the probability of non-default. This is the usual way in credit risk SHAP analysis and ensures that positive SHAP values consistently indicate features that increase predicted default risk while negative values indicate features that reduce it.

3.9 Fairness Analysis Framework

The fairness analysis conducted in this thesis was motivated by a straightforward but important observation that a credit risk model that performs well on average may still treat different groups of borrowers very differently, and those differences may have real consequences for people's access to credit. For example, a model that overestimates default risk for younger borrowers may deny credit to young people who would have actually repaid their loans. This may not be because of anything specific about their individual financial situation, but because the model has learned a statistical link between youth and default risk at the population level and applies that link to every young borrower regardless of their personal circumstances. This kind of outcome is exactly what fair lending regulations are designed to prevent, and it is exactly the kind of pattern that the EU Artificial Intelligence Act's risk-management and bias-related requirements are intended to bring under closer scrutiny in AI-based credit scoring systems.

The fairness analysis in this study focuses specifically on age as the demographic dimension of interest. This choice was motivated by three considerations. First, the exploratory data analysis confirmed that age is one of the strongest correlates of default risk in the Give Me Some Credit

dataset older borrowers default significantly less often than younger ones making it a variable that the model is likely to rely on heavily and one where the risk of demographic bias is therefore particularly salient. Second, the engineered Age Group variable constructed as described in Section 3.5 provides a natural and interpretable grouping framework for the fairness analysis, dividing borrowers into three age categories defined for the purpose of this study: Young Adults (18–35), Middle-Aged (36–55), and Older Adults (56–90) that are of sufficient size for meaningful statistical comparison. Third, age is a legally relevant fairness dimension in both jurisdictions, though the applicable legal frameworks differ substantially the US relies on direct statutory prohibition through the Equal Credit Opportunity Act, while the EU's protection arises more indirectly through the AI Act's high-risk classification and GDPR's automated decision-making safeguards, as discussed in Section 2.4.

3.9.1 Fairness Metrics Used in This Study

The fairness analysis was conducted by comparing two quantities within each age group, the actual default rate and the predicted default rate and examining the gap between them as a measure of prediction bias at the group level. The actual default rate within each age group is simply the proportion of borrowers in that group who genuinely experienced serious financial distress within the observation period, that is, the proportion of test set observations in each age group for whom the target variable SeriousDlqin2yrs takes a value of one. The predicted default rate within each age group is the proportion of borrowers in that group who were classified as defaulters by the Random Forest model, that is, the proportion of test set observations in each age group for whom the model's binary prediction takes a value of one.

The gap between the actual and predicted default rates within each age group provides a direct and intuitively interpretable measure of prediction bias. A positive gap where the predicted default rate

exceeds the actual default rate indicates that the model is overestimating default risk for that group, systematically classifying more borrowers as high risk than actually default. In a lending context this translates to a higher rate of false positives for that group borrowers who would have repaid their loans but were flagged as likely defaulters with potential consequences for their access to credit and the terms on which it is offered. A negative gap where the predicted default rate falls below the actual default rate indicates that the model is underestimating default risk, systematically missing defaulters in that group and potentially exposing the lender to unexpected losses. Both types of bias have practical consequences, and both are examined in the results presented in Chapter 4.

3.9.2 Connection Between SHAP and Fairness Analysis

One of the distinctive features of the analytical framework developed in this thesis is the explicit connection between the SHAP interpretability analysis and the fairness analysis. Rather than treating these as two separate analytical exercises one that explains the model and one that evaluates its fairness this study uses the SHAP findings to illuminate the fairness results, asking not only whether the model produces age-differentiated predictions but why it does so and which specific features are responsible.

This connection is made through the examination of the SHAP feature importance profile in relation to the fairness findings. If the SHAP analysis shows that age or the Age Group variable is among the most important predictors in the model, this would suggest that the model's demographic bias is at least partly attributable to its direct reliance on age as a predictive signal. And, if age ranks are relatively low in the SHAP feature importance profile despite a gap between actual and predicted default rates across age groups, this would suggest that the bias is driven more indirectly through correlations between age and other variables. The difference between direct age-

based prediction and proxy-based age discrimination is important both analytically and from a regulatory perspective. As the EU Artificial Intelligence Act's data governance and bias-related requirements are relevant to monitoring both direct and indirect (proxy-based) discrimination in high-risk AI systems.

The connection between SHAP and fairness analysis also allows a more detailed interpretation of the fairness results than would be possible from the demographic comparison alone. When discussing the finding that the model overpredicts default risk for Young Adults in Chapter 4, the SHAP analysis allows this observation to be connected to the specific model behaviors that drive it for example, whether young borrowers tend to have higher Weighted Delinquency Scores or higher revolving credit utilisation rates than older borrowers, and whether these differences in financial behaviour fully explain the observed prediction gap or whether some residual age-based bias remains after controlling for financial characteristics. This kind of mechanistic explanation connecting the what of the fairness finding to the why of the model's behaviour is what transforms a descriptive fairness audit into a genuine analytical contribution to the literature on algorithmic fairness in credit scoring.

3.9.3 Scope and Limitations of the Fairness Analysis

It is important to be transparent about the scope and limitations of the fairness analysis conducted in this study, as these affect how the results should be interpreted and what conclusions can legitimately be drawn from them.

First, fairness analysis focuses exclusively on age as a demographic dimension and does not examine other potentially relevant fairness dimensions such as gender, income level, or geographic location. This focus reflects the specific research questions of the thesis and the availability of demographic variables in the Give Me Some Credit dataset which contains age and number of

dependents as the only demographic variables rather than a judgement that age is the only fairness dimension worth examining in credit scoring contexts. Future research could extend the framework developed here to examine additional demographic dimensions using datasets with richer demographic information.

Second, the fairness metrics used in this study, the comparison of actual and predicted default rates and the average predicted probability within each age group represent a descriptive approach to fairness measurement rather than a formal implementation of any specific fairness criterion from the theoretical literature reviewed in Section 2.4. This approach was chosen because it is intuitively interpretable and directly relevant to the practical concern of whether the model produces demographically differentiated outcomes, rather than because it satisfies any theoretical definition of fairness. Readers interested in a more formal fairness evaluation using metrics such as demographic parity difference or equalised odds difference are referred to the fairness literature reviewed in Section 2.4 and the future research directions identified in Chapter 6.

Third, the fairness analysis is conducted on the test set only and reflects the performance of the model on the specific sample of borrowers included in that set. While the stratified train-test split described in Section 3.6 ensures that the test set is representative of the full dataset in terms of the overall class distribution, it does not guarantee that the age group composition of the test set perfectly mirrors that of the full borrower population. The results of the fairness analysis should therefore be interpreted as estimates of the model's fairness properties on the borrower population represented by the dataset rather than as definitive measures of fairness in any broader sense.

These limitations are acknowledged not to undermine the value of the fairness analysis but to situate it appropriately within the scope of the thesis and to be transparent about the interpretive boundaries within which the results should be understood. The fairness analysis contributes

empirical evidence to the understanding of age-based prediction disparities in machine learning credit scoring, and its findings have direct practical relevance for the design and governance of credit scoring systems in regulated financial environments even if the specific metrics used do not capture every dimension of fairness that the theoretical literature has proposed.

3.10 Use of Artificial Intelligence Tools

Artificial Intelligence (AI) tools were used in a limited and supportive capacity during the preparation of this thesis.

Claude (Anthropic) was used to assist with debugging Python code, refining the data cleaning, feature engineering, and modelling pipeline, and improving the structure of the SHAP interpretability analysis, fairness analysis, and data visualizations described in Chapters 3 and 4.

Claude was used for supporting early drafting of written explanations of the methodology, results, and discussion sections. They were all done based on the author's own data, code outputs, and analytical decisions. These AI-assisted drafts were revised and rewritten by the author before being included in the final thesis.

The AI tool was not used to generate the core academic content, research findings, analysis, or conclusions of this thesis. All code was executed, and all outputs including model performance metrics, SHAP values, and fairness analysis results were produced through the author's own analysis of the datasets described in Section 3.2. All AI-assisted text and code were carefully reviewed, verified, edited, and validated by the author to ensure accuracy, reliability, originality, and consistency with academic standards.

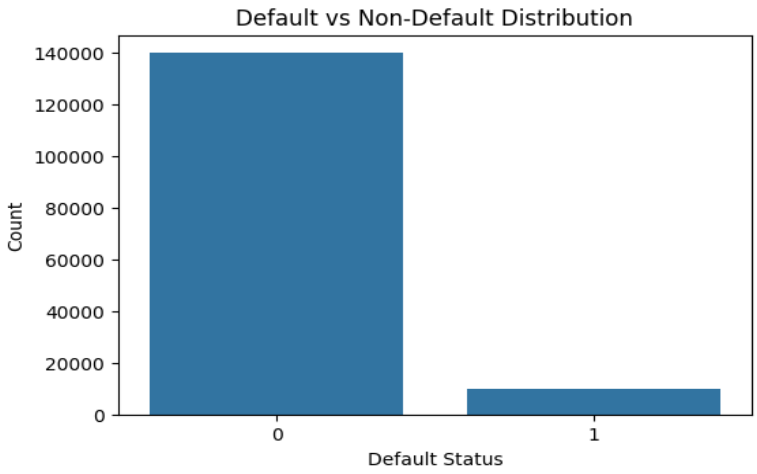
Chapter 4: Results

4.1 Exploratory Data Analysis Results

Before any cleaning was applied, the raw dataset told a fairly clear story about what problems needed fixing.

Missing data. Two columns had gaps MonthlyIncome was missing in 29,731 rows (19.82% of the dataset) and NumberOfDependents was missing in 3,924 rows (2.62%). Neither pattern looked systematically tied to whether the borrower defaulted, which supported using median imputation rather than something more complicated.

Figure 4.1 : Class Distribution

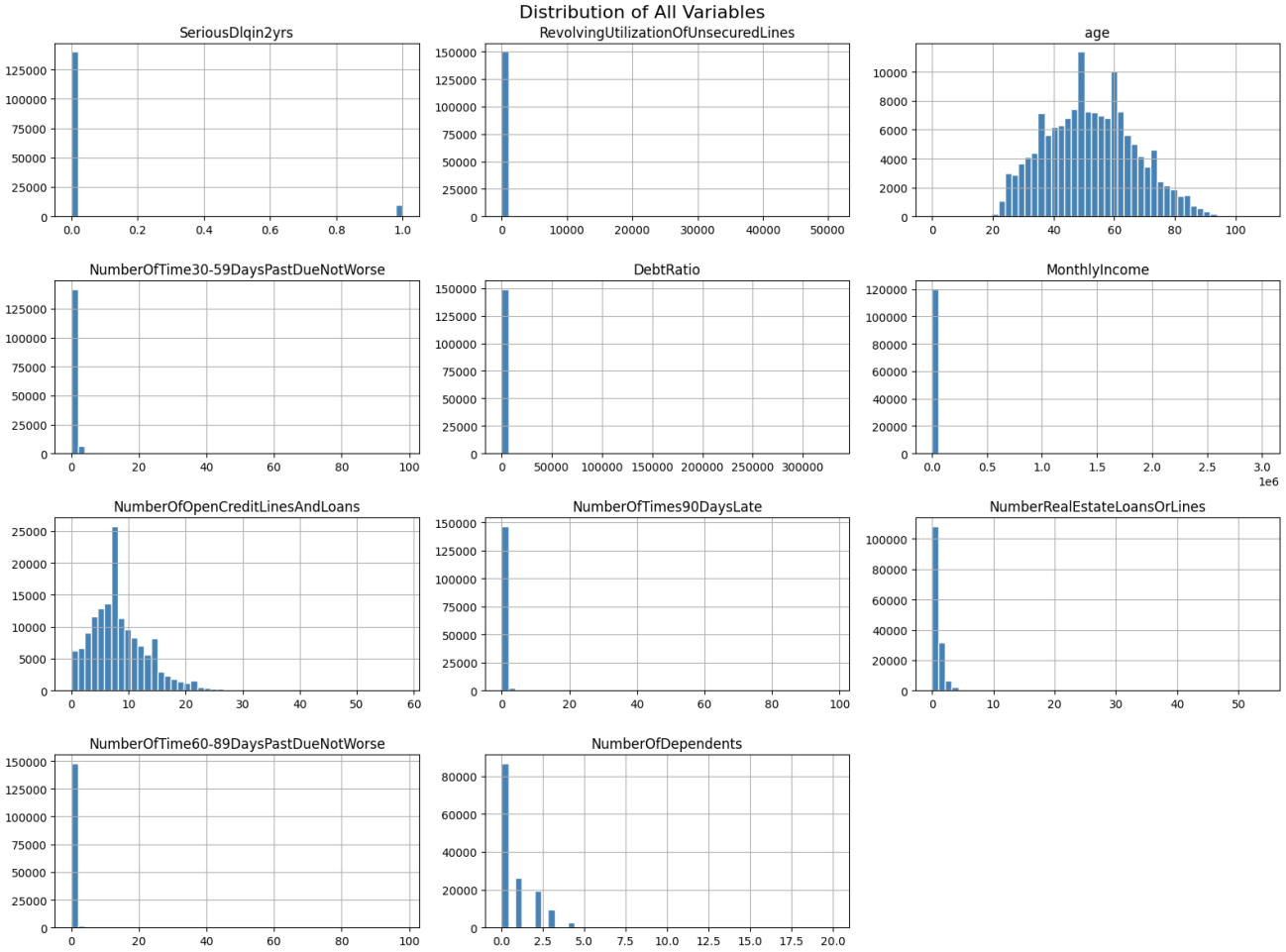


Class imbalance was the single biggest issue in the dataset. Out of 150,000 borrowers, 139,974 (93.3%) never experienced serious delinquency, while only 10,026 (6.7%) did. This imbalance shows up throughout the rest of the analysis, it's the reason SMOTE was necessary and the reason accuracy alone would have been a misleading metric.

Variable distributions and outliers. Age was roughly bell-shaped, centered around 52, but with one entry at age 0 and 489 entries above 90 clearly not real borrowers, just data entry noise. The three delinquency columns were heavily skewed toward zero, with a handful of values sitting at 96 or 98, which is not physically possible over a two-year window.

RevolvingUtilizationOfUnsecuredLines, which should sit between 0 and 1 by definition, had 3,321 rows above 1, some by a huge margin. MonthlyIncome had a long tail stretching into the millions.

Figure 4.2 Distribution of All Variables



Correlations.

The three delinquency variables came out as the strongest predictors of default even before any feature engineering "Number OfTime 30-59Days Past Due Not Worse" at 0.126, NumberOfTimes90DaysLate at 0.117, and NumberOfTime60-89DaysPastDueNotWorse at 0.102. Age had the strongest negative correlation at -0.115, meaning older borrowers' default less ended

up being important context for the fairness analysis later. Everything else lined up with what you'd expect from banking intuition: higher debt ratios and more dependents pushed toward default, more open credit lines and higher income pushed away from it.

Figure 4.3: Feature Correlation with Default Risk

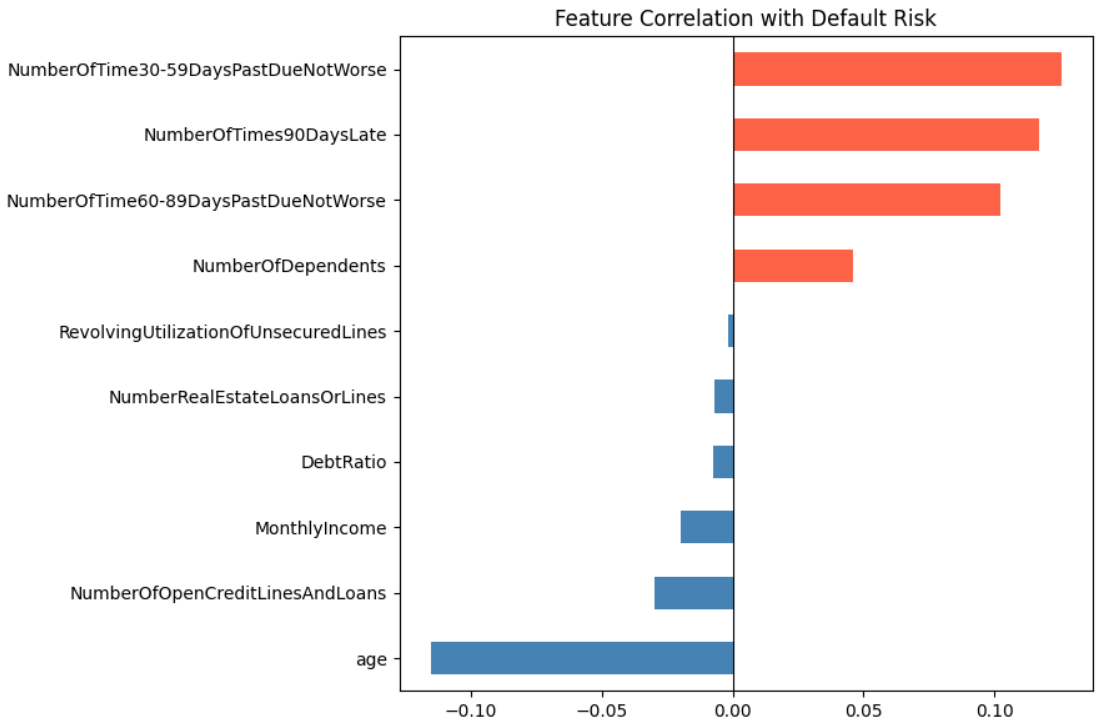
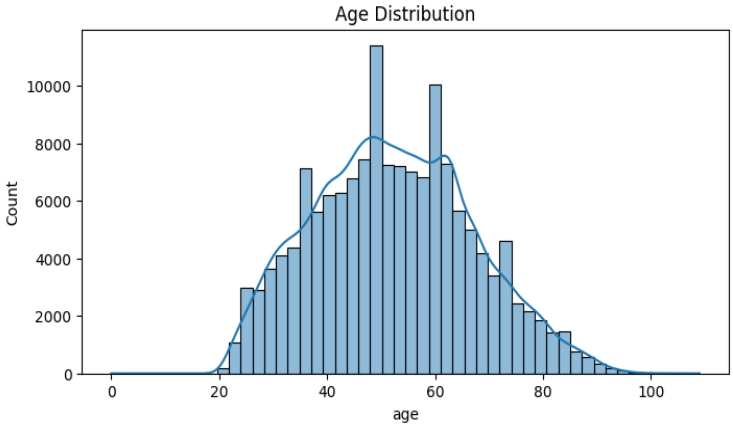


Figure 4.4: Age Distribution



4.2 Model Performance Comparison

After cleaning (609 duplicates removed, age filtered to 18–90, utilisation capped, delinquency capped at 10, income and debt ratio winsorised at the 99th percentile), the dataset dropped from 150,000 to 145,592 rows a loss of about 2.9%, which is small enough not to worry about selection bias. SMOTE was then applied only to the training set, bringing it to a balanced 218,928 rows (109,464 in each class), while the test set of 29,119 was left untouched at its natural 93.3%/6.7% split.

Three models were trained on this data: Logistic Regression, Random Forest, and XGBoost. Table 4.1 shows how they compared.

Table 4.1: Model Performance — US Dataset

Model	AUC-ROC	F1 Score	Precision (Default)	Recall (Default)
Logistic Regression (scaled)	0.8144	0.2936	0.19	0.63
Random Forest	0.8124	0.3266	0.31	0.34
XGBoost	0.8066	0.3071	0.28	0.34

Figure 4.5: ROC Curves and AUC Bar Chart

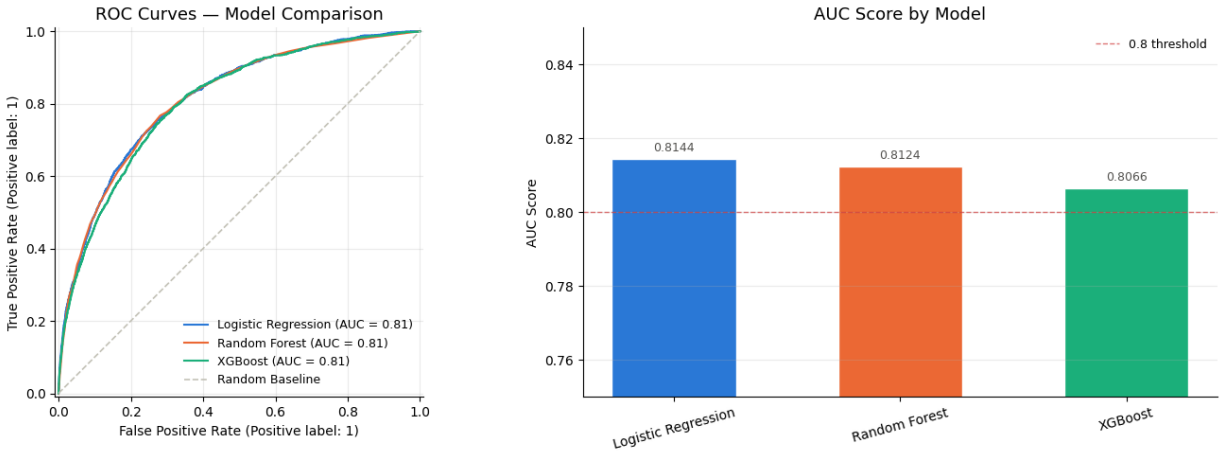
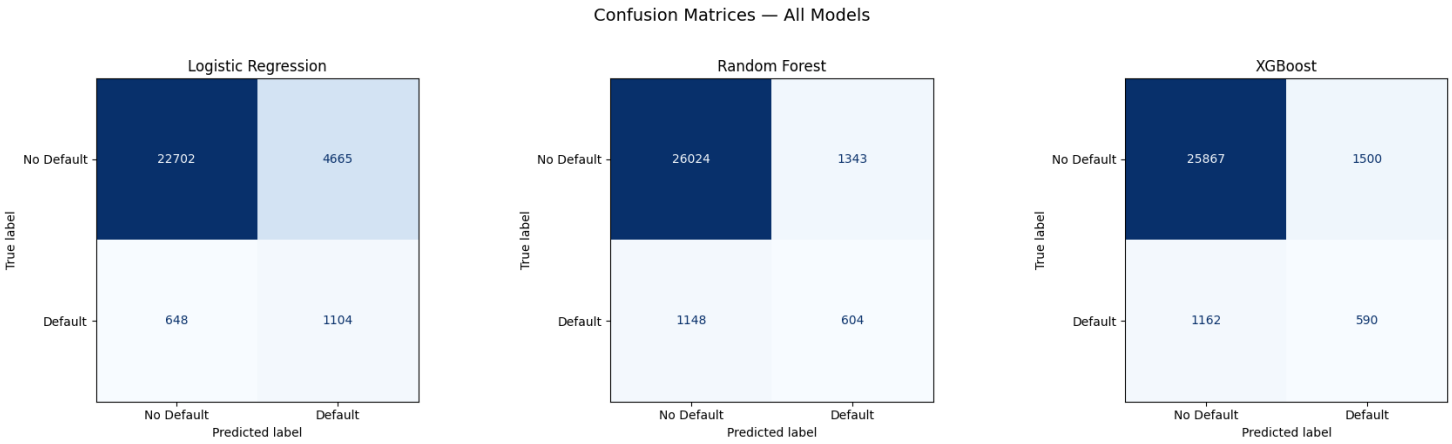


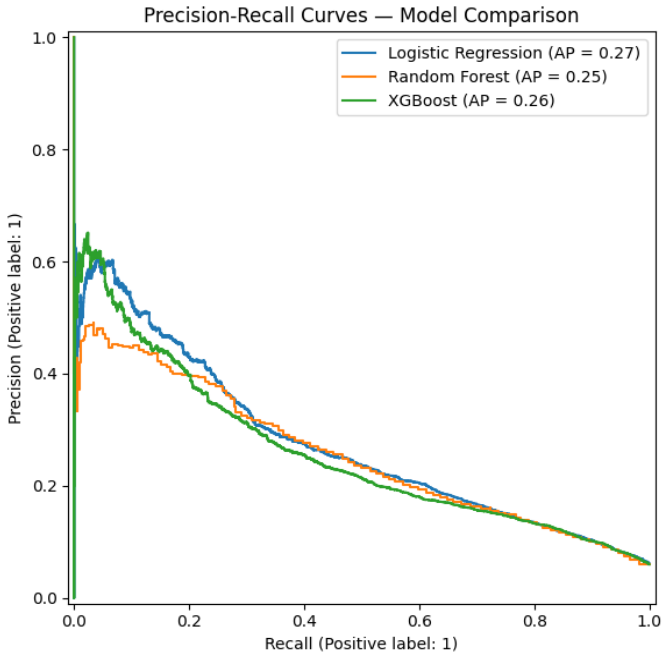
Figure 4.6: Confusion Matrices



All three models cleared the 0.80 AUC mark, and honestly the differences between them are pretty small only 0.008 separates the best and worst on AUC. Logistic Regression actually edges out the other two on AUC (0.8144 vs 0.8124 vs 0.8066), which surprised me a little at first, but F1 tells a different story. Random Forest wins on F1 (0.3266), meaning it strikes the best balance between catching real defaulters and not flooding the system with false alarms.

Logistic Regression's precision/recall split (0.19 / 0.63) shows it casts a wide net it catches 63% of actual defaulters but at the cost of flagging a lot of people who never default. Random Forest and XGBoost are more conservative, with lower recall (0.34 each) but noticeably better precision. Which trade-off is "better" really depends on what a bank cares about more missing defaulters or wrongly denying good borrowers but for the purposes of this thesis, F1 was used as the deciding metric, since it's the one that actually reflects how well a model handles the minority class specifically.

Figure 4.7: Precision-Recall Curves



Based on this, **Random Forest was selected as the best model** and carried forward into the SHAP and fairness analysis.

4.2.1 Hyperparameter Tuning

The comparison above uses default settings for all three models. To check whether this affects the conclusions, Random Forest and XGBoost were both re-run using RandomizedSearchCV, testing a range of settings for tree depth, number of estimators, and other key parameters.

Table 4.1a: Tuned vs. Untuned Model Performance

Model	AUC (untuned)	AUC (tuned)	F1 (untuned)	F1 (tuned)
Random Forest	0.8124	0.8204	0.3266	0.3365
XGBoost	0.8066	0.8214	0.3071	0.3310

Tuning helped both models a little. XGBoost picked up the bigger gain, moving from the weakest AUC of the three original models to the strongest once tuned. Random Forest also improved, and

it still holds the best F1 score overall. Even with tuning, though, all the three models stay within about two points of AUC from each other. So, none of these models has a big advantage over the others on this dataset.

Random Forest still has the best F1 after tuning, so it remains the model used for the SHAP analysis in the sections that follow.

4.3 SHAP Feature Importance — US Data

SHAP was run on a random sample of 500 test-set observations using the Random Forest model. Table 4.2 shows the ranked mean absolute SHAP values.

Table 4.2: Mean Absolute SHAP Values — US Dataset

Rank	Feature	Mean |SHAP|
1	DelinquencyScore	0.1700
2	RevolvingUtilizationOfUnsecuredLines	0.1163
3	NumberOfDependents	0.0980
4	NumberRealEstateLoansOrLines	0.0389
5	DebtRatio	0.0327
6	age	0.0278
7	NumberOfTime30-59DaysPastDueNotWorse	0.0225
8	MonthlyIncome	0.0195
9	NumberOfOpenCreditLinesAndLoans	0.0182
10	NumberOfTimes90DaysLate	0.0085
11	NumberOfTime60-89DaysPastDueNotWorse	0.0081

Figure 4.8: SHAP Beeswarm — US Data

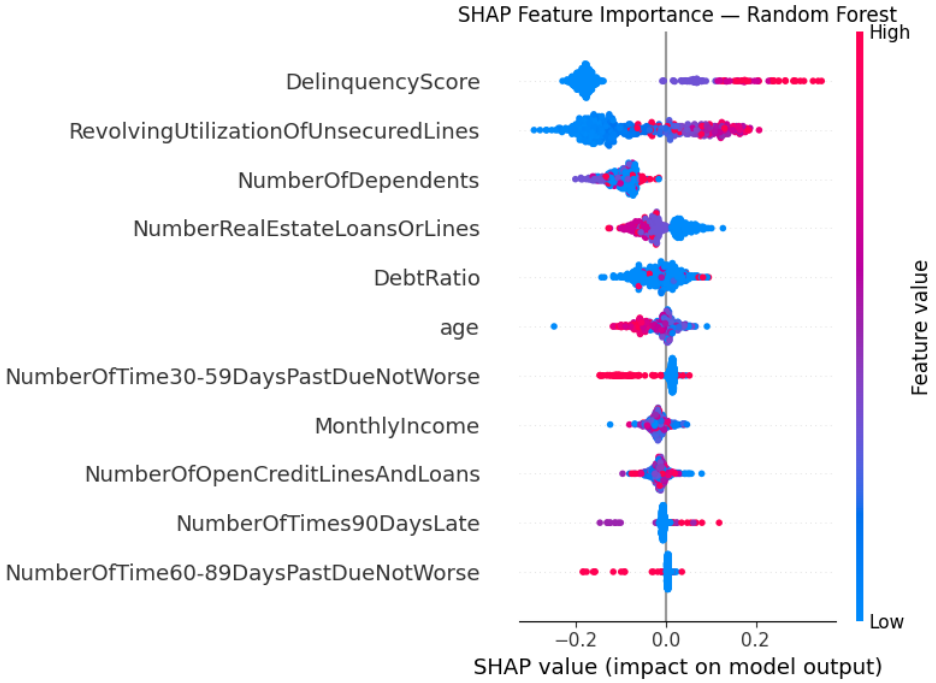
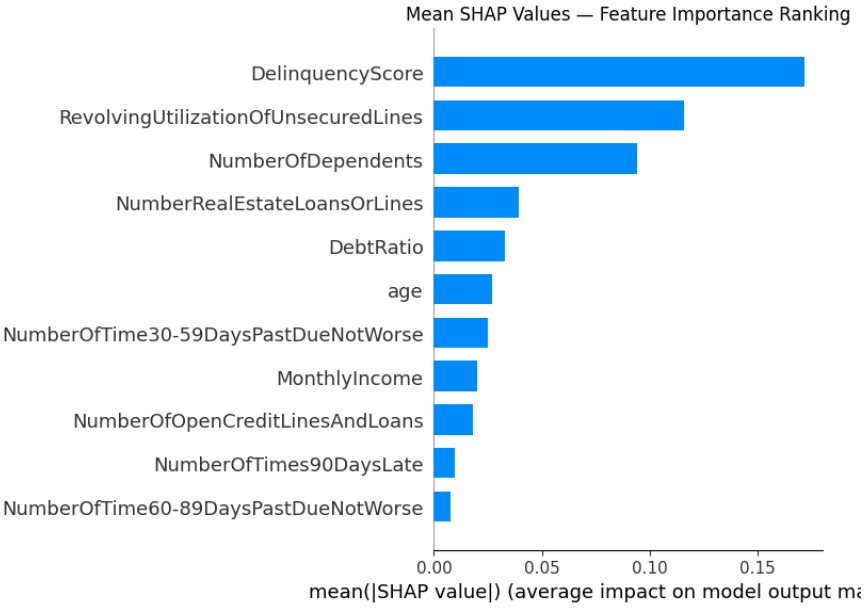


Figure 4.9: SHAP Bar Chart — US Data

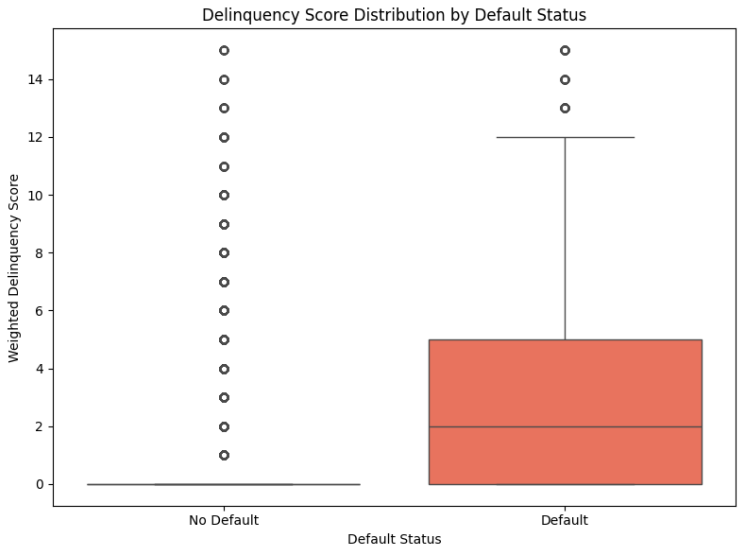


The standout result here is that DelinquencyScore, came out as the single most important predictor, ahead of every original variable in the dataset by a wide margin. This's a useful finding. It means combining the three raw delinquency counts into one severity-weighted score didn't just simplify

the data. It actually concentrated the predictive signal better than the original columns did individually. You can see this directly in the ranking the three original delinquency columns (30-59, 60-89, 90+ days) all dropped to the bottom of the list (ranks 7, 10, 11). This is exactly what you'd expect if the composite score successfully captured what they were individually contributing.

Figure 4.10 confirms this pattern visually borrowers who defaulted show a clearly higher median DelinquencyScore than those who did not, with much greater spread in the distribution among defaulters.

Figure 4.10: Delinquency Score Boxplot by Default Status



RevolvingUtilizationOfUnsecuredLines came in second, which lines up with basic credit theory someone maxing out their available credit is under more financial strain. NumberOfDependents landed third, which was a bit unexpected given how weak its raw correlation with default was in Section 4.1. This suggests that model is capturing on an interaction effect that a simple correlation alone wouldn't detect. More dependents likely means less disposable income to deal with a shock and even for borrowers who otherwise look financially stable.

Age ranked sixth. The beeswarm plot shows younger borrowers (low age, blue) pushing predictions toward default and older borrowers (high age, red) pushing away from it in line with the negative correlation found earlier. It's not the dominant feature in the model, but it's clearly doing something, which matters for the fairness discussion in the next section.

Figure 4.11: SHAP Waterfall — Highest Risk Borrower

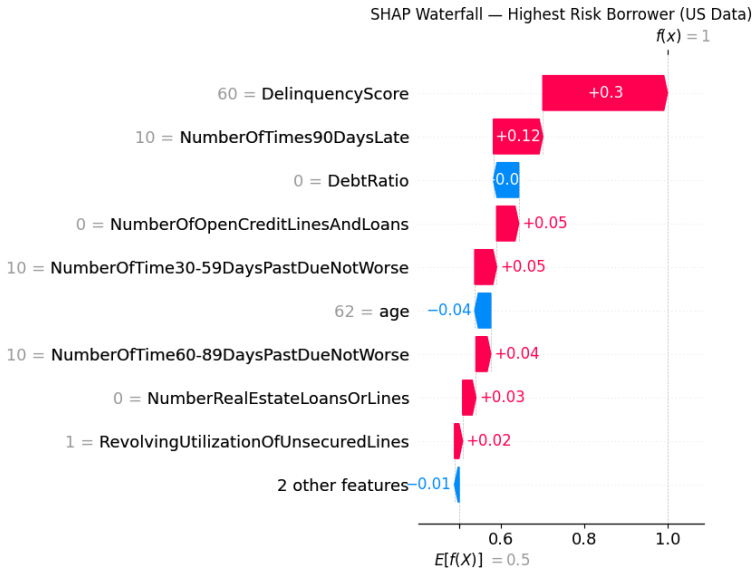
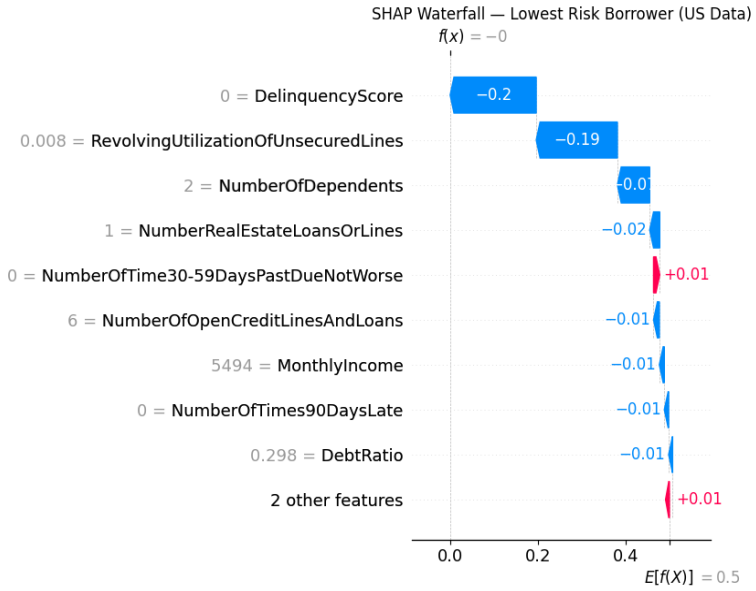


Figure 4.12: SHAP Waterfall — Lowest Risk Borrower



To make this concrete: the highest-risk borrower in the sample was assigned a predicted probability of 0.95. Looking at their SHAP breakdown, a DelinquencyScore of 10 contributed +0.29 on its own by far the single biggest push toward default with high revolving utilisation (80.5%) adding another +0.11. On the other end, the lowest-risk borrower had a predicted probability close to 0, driven mainly by a delinquency score of zero and low credit utilisation. This is basically the kind of explanation a bank would need to hand a rejected applicant, or show a regulator, if they were actually using this model in production.

The touch of tuning changes performance a little. But it's necessary to check whether it also changes what the model is actually paying attention to. SHAP was recalculated during tunning Random Forest, and the result came out identical to the original, with only small shifts in the exact SHAP values. As a top feature DelinquencyScore is still same, also, “Revolving Utilization Of Unsecured Lines” and NumberOfDependents are in the same order as before. Interpretability findings here are not just a side effect of using default settings. The model relies on the same variables, in the same order of importance, whether it's been tuned.

4.4 Age-Based Fairness Analysis

This is where things get more interesting. The fairness analysis compares actual default rates against what the Random Forest model predicted, broken down by the three age groups.

Table 4.3: Fairness Analysis by Age Group — US Dataset

Age Group	Count	Actual Default Rate	Predicted Default Rate	Gap	Avg Risk Score
Young Adults (18–35)	4,072	10.09%	12.01%	+1.92pp	0.2229

Middle-Aged (36–55)	12,952	7.33%	8.35%	+1.02pp	0.1827
Older Adults (56–90)	12,095	3.23%	3.12%	−0.11pp	0.1011

Figure 4.13: Actual vs Predicted Default Rate by Age Group

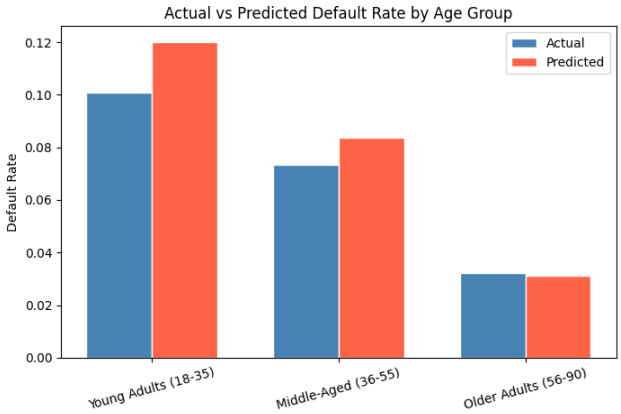
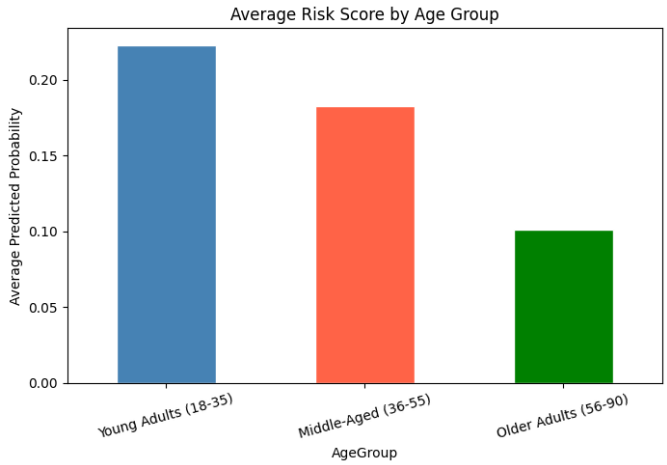


Figure 4.14: Average Risk Score by Age Group



The pattern here is pretty clear once you lay it out. Young Adults show the largest gap between actual and predicted default the model predicts 12.01% will default when only 10.09% actually do, a difference of almost 2 full percentage points. Middle-Aged borrowers show a smaller but still real gap (+1.02pp). Older Adults, on the other hand, show almost no gap between actual and

predicted default rates the model very slightly underpredicts their risk (−0.11pp), which is close enough to zero that it's basically noise.

So, the model isn't wrong across the board it consistently overpredicts default risk for younger borrowers specifically. Every Young Adult in the test set, on average, gets a higher risk score (0.2229) than their actual outcomes would justify more than double the average score given to Older Adults (0.1011). Whether this reflects a genuine underlying difference in risk, or a pattern the model picked up from the training data that doesn't fully hold at the individual level, is something I explore further below alongside the SHAP results from Section 4.3.

4.5 German Credit Data Validation Results

The same Random Forest architecture was retrained from scratch on the German Credit Data to see whether the findings hold up outside the US context.

Table 4.4: Random Forest Performance — German Dataset

Metric	Bad Credit (Class 0)	Good Credit (Class 1)	Overall
Precision	0.51	0.82	—
Recall	0.60	0.76	—
F1 Score	0.55	0.79	0.72 (weighted)
AUC-ROC	—	—	0.7674
Accuracy	—	—	71%

The AUC on German data (0.7674) is about 0.045 lower than on US data (0.8124) the model still works reasonably well, but there's a clear drop once it's retrained on a different country's data. The German model's overall weighted F1 score is 0.72, compared to 0.3266 for the US model. However, this is not a fair comparison. The 0.72 figure is pulled up by the much larger good-credit class, not by how well the model catches risky cases. A fairer comparison looks at the F1 score for

the matching risk class in each dataset: 0.3266 for US defaulters versus 0.55 for German bad-credit cases. Germany is still somewhat higher, but the gap is far smaller than the raw weighted average suggests, and AUC remains the fairer overall comparison since it doesn't depend on class balance at all by that measure, the US model is stronger.

Figure 4.15: SHAP Beeswarm — German Data

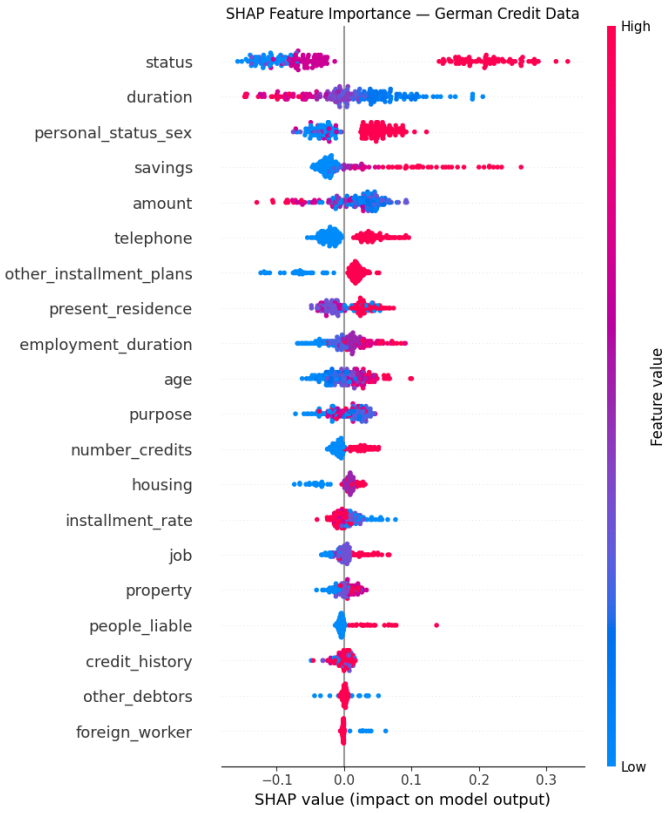


Table 4.5: Mean Absolute SHAP Values — German Dataset

Rank	Feature	Mean |SHAP|
1	status	0.1288
2	duration	0.0540
3	personal_status_sex	0.0446
4	savings	0.0430
5	amount	0.0364

6	telephone	0.0313
7	other_installment_plans	0.0278
8	present_residence	0.0255
9	employment_duration	0.0232
10	age	0.0227

The table above shows the ten features with the highest mean absolute SHAP value; the remaining ten variables in the German dataset had noticeably smaller contributions and are listed in full in Appendix A.

This is genuinely the most interesting part of the results. The feature that matters most in Germany isn't anything like the feature that matters most in the US. In the US model, the top predictor was DelinquencyScore a behavioural measure of past payment history. In Germany, the top predictor is status, the borrower's checking account standing a structural, present-tense signal rather than a historical one. Duration of the loan comes second, followed by demographic and financial-position variables like savings and personal status.

Interestingly, age lands in a similar position in both models 6th in the US and 10th in Germany. Age ranked sixth in the US model and tenth in the German model, indicating that it was not among the dominant features in either model.

4.6 Cross-Dataset Comparison and Summary

Pulling the US and German results together, a few things stand out.

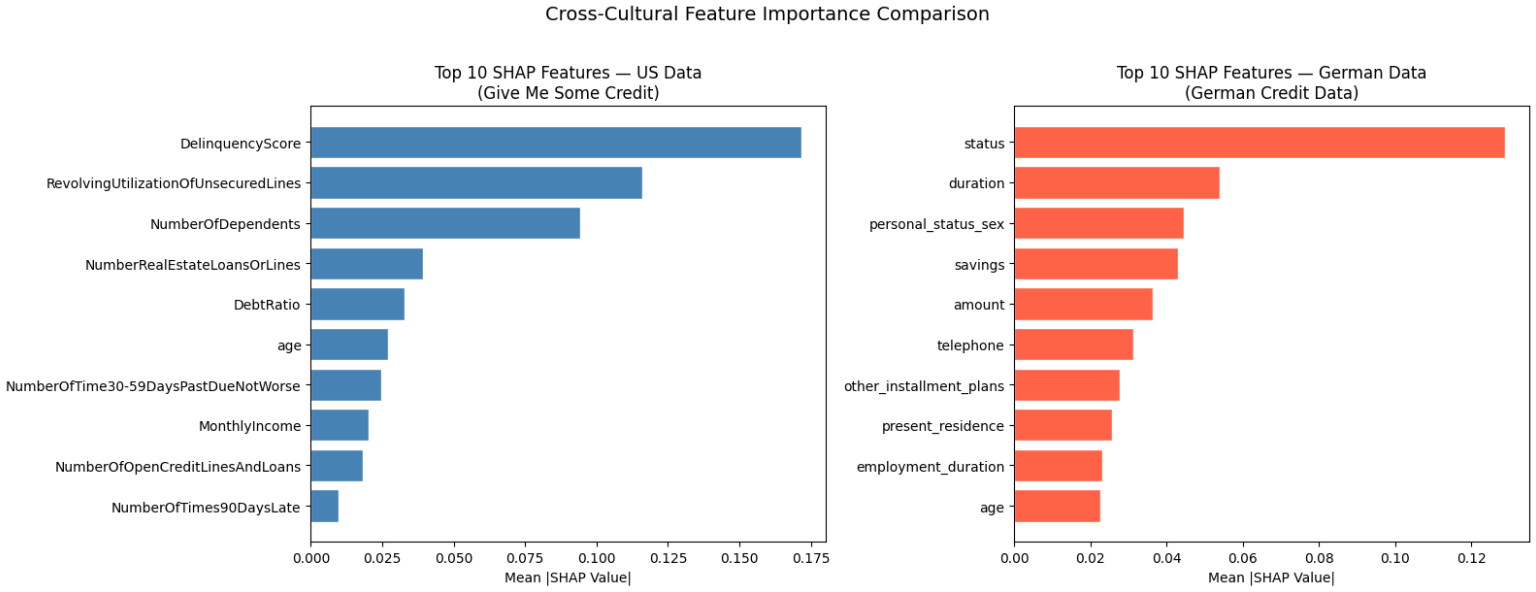
Table 4.6: Random Forest Performance — US vs German

Metric	US	German
AUC-ROC	0.8124	0.7674
F1 Score	0.3266	0.55

Table 4.7: Top 5 SHAP Features — Side by Side

Rank	US	Mean |SHAP|	German	Mean |SHAP|
1	DelinquencyScore	0.1700	status	0.1288
2	RevolvingUtilization	0.1163	duration	0.0540
3	NumberOfDependents	0.0980	personal_status_sex	0.0446
4	NumberRealEstateLoans	0.0389	savings	0.0430
5	DebtRatio	0.0327	amount	0.0364

Figure 4.16:Cross-Cultural Feature Importance Comparison



Tying this back to the research questions from Chapter 1: Random Forest was the best-performing model on the US data (0.8124 AUC, 0.3266 F1), beating both Logistic Regression and XGBoost on the metric that mattered most for this problem The engineered DelinquencyScore variable emerged as the single most important feature in the model and ranked above each of its individual component variables, indicating that the model relied heavily on the composite representation, although its incremental predictive value cannot be established without comparison against a

model trained without it. The fairness analysis showed a real, measurable gap for Young Adults specifically (+1.92pp), while Older Adults were essentially unaffected. And the German cross-dataset comparison showed that the same Random Forest architecture achieved reasonably strong predictive performance after being retrained on the German data, while relying on a fundamentally different set of variables behavioural history in the US, current financial standing in Germany. What all of this actually means and what it implies for banks and regulators is where Chapter 5 picks up.

Chapter 5: Discussion

This chapter steps back from the raw numbers in Chapter 4 and asks what they actually mean. I've organised it around the four research questions from Chapter 1, then connected each finding back to the literature reviewed in Chapter 2 and finished with the practical implications and an honest look at the limitations of this study.

5.1 Answering the Primary Research Question

The primary question asked which model most effectively predicts credit risk, and what drives its predictions. Based on the results, **Random Forest is the answer to the first half of that question** it achieved the best F1 score (0.3266) among the three models, which was the metric this thesis prioritized precisely because it reflects performance on the minority class rather than overall accuracy. Logistic Regression technically edged it out on AUC (0.8144 vs 0.8124), but the margin is small, and AUC alone doesn't tell you much about how well a model actually catches real defaulters at a usable threshold which is the metric that matters more in practice.

What is interesting is that the AUC gap between all three models Logistic Regression, Random Forest, and XGBoost was only 0.008. That's a small difference for models that differ so much in

complexity. This partly echoes what Lessmann et al. (2015) found in their large benchmark study ensemble methods tend to outperform logistic regression, but not always by a dramatic margin, and the size of the gain depends a lot on the specific dataset. In this case, the more complex models (Random Forest, XGBoost) didn't blow Logistic Regression out of the water on raw discrimination, but they did handle the minority class more sensibly, which is the part that actually matters for a bank trying to identify who's likely to default. This result was confirmed even after both models were properly adjusted. The difference between the three approaches remained small, and the SHAP ranking for "Random Forest" did not change at all. In other words, the simple model wasn't winning just because the others weren't tuned properly, the models are just genuinely close to each other in performance.

As for what drives the predictions that's where the SHAP analysis earns its place in this thesis. Rather than treating Random Forest as an unexplainable black box, the SHAP results in Section 4.3 gave a clear, ranked answer: DelinquencyScore, RevolvingUtilizationOfUnsecuredLines, and NumberOfDependents were the three biggest drivers of the model's decisions. This directly answers the "what did the model actually learn" part of the primary question in a way that a bare AUC score never could.

5.2 Sub-Question 1 — Did the Engineered Features Add Value?

DelinquencyScore came out as the single most important feature in the model, ahead of every one of the ten original variables. As shown in Figure 4.10, defaulters carry a visibly higher DelinquencyScore than non-defaulters, and the model clearly picked up on that. That's not a small result it means the model relies far more heavily on this combined score than on any of the three raw delinquency columns it was built from, suggesting the composite measure captures something the individual variables don't on their own.

The clearest evidence for this is what happened to the three original delinquency columns once DelinquencyScore was added to the model they dropped to ranks 7, 10, and 11 out of 11 features, a pattern that's plainly visible in Figure 4.9's bar chart. If the composite score hadn't actually absorbed their signal, you'd expect them to still rank fairly high on their own. Instead, they became almost irrelevant next to the combined measure. That's a fairly direct empirical confirmation that the weighting scheme giving three times the weight to 90+ day delinquencies compared to 30–59-day ones reflects something real about how severity should be treated, not just an arbitrary choice.

AgeGroup is a slightly different story. It wasn't included as a direct model input (only the continuous age variable was, to avoid redundancy), so it doesn't have its own SHAP ranking. But its value shows up indirectly, it's the variable that made the fairness analysis in Section 4.4 possible in the first place. Without it, there'd be no clean way to group borrowers and compare actual versus predicted default rates the way this thesis does. So, its contribution is more structural than predictive, but it's still a genuine methodological contribution.

5.3 Sub-Question 2 — Is the Model Fair Across Age Groups?

As Figure 4.13 shows, Older Adults showed almost no gap between their actual and predicted default rates a −0.11 percentage point difference that's small enough to be noise. Middle-Aged borrowers showed a real but moderate overprediction (+1.02pp). **Young Adults were the clear outlier**, with the model overpredicting their default rate by 1.92 percentage points nearly double the gap seen for middle-aged borrowers, and far beyond what Older Adults experienced. Figure 4.14 makes this even clearer at the level of average predicted risk Young Adults receive an average score more than double that of Older Adults (2.2x), a narrower gap than the roughly threefold difference (3.1x) in their actual default rates. In other words, the model's average risk scores

understate the true difference in outcomes between these two groups, even as it overpredicts each group's absolute risk level.

This pattern is consistent with the broader concern in the algorithmic fairness literature that models trained on historical data can reproduce demographic disparities even without explicit intent (Mehrabi et al., 2021). What's useful here is being able to connect this finding back to the SHAP results from Section 4.3. Age itself ranked only 6th in feature importance not the dominant factor which suggests the model isn't relying on age directly to drive its predictions. However, this analysis alone can't determine whether the observed disparity comes from age correlating with other risk factors like credit utilisation, from differences in how well-calibrated the model is across age groups, or from some other source of prediction error. What the modest SHAP ranking for age does rule out is the simplest explanation that the model is crudely penalizing borrowers just for being young.

Practically, this matters. A 1.92 percentage point overprediction sounds small in isolation, but multiplied across thousands of loan applications, it could contribute to inappropriate credit denials or less favorable lending terms for some younger borrowers, if a model like this were used directly in real lending decisions. The EU AI Act establishes data governance, risk management, and bias-monitoring requirements for high-risk AI systems, which makes this kind of demographic disparity directly relevant to the model governance processes banks are expected to maintain.

5.4 Sub-Question 3 — Does the Random Forest Approach Perform Comparably and Rely on Similar Factors Once Retrained on German Data?

Partially, and the "partially" is the interesting part.

On raw performance, the answer is a qualified yes. The AUC dropped from 0.8124 on US data to 0.7674 on German data a real decline, but not a collapse. The model retained meaningful discriminatory power even when retrained on a completely different national credit dataset with

different variables. This is broadly consistent with the general finding in the literature that no single algorithm generalizes perfectly across all contexts (Lessmann et al., 2015), though in this case the model was retrained on the German data rather than applied directly, which likely helped preserve more of the performance.

But the more revealing finding is in the SHAP comparison. The two models, despite sharing the same architecture and training approach, rely on almost entirely different variables to make their decisions. The contrast is stark in Figure 4.16, which places the two rankings side by side. In the US, the dominant signal is behavioural DelinquencyScore, credit utilisation, and past payment patterns. In Germany, the dominant signal is structural checking account status, loan duration, savings level. One possible contextual explanation is the difference in credit information environments discussed in Section 2.5 the US system, run by Equifax, Experian, and TransUnion, is built around detailed behavioural history, while Germany's SCHUFA system is comparatively limited in that respect. However, the present analysis cannot establish that differences in credit bureau infrastructure caused the observed differences in feature importance.

So, the answer depends on what you're asking. If the question is "does the retrained model continue to perform reasonably well" :yes, mostly. If the question is "does it rely on the same information to achieve this performance" : no, not really. The model essentially had to re-learn what matters when trained on a different financial system's data, which is a genuinely useful finding for any bank thinking about deploying the same model architecture across multiple countries without expecting to also re-validate what drives it in each market.

One small thing worth noting: age ranked 6th in the US model and 10th in the German model, indicating it was not among the dominant features in either. That's a small but interesting hint that

age-related risk patterns might be somewhat more universal than the behavioural-versus-structural split suggests for everything else.

5.5 Practical Implications

For banks, the clearest takeaway is that investing time in domain-informed feature engineering can matter more than jumping straight to a more complex model. DelinquencyScore outperformed every raw variable in the dataset, and it was built using nothing more than banking judgement applied to variables that were already there. That's a cheap, low-risk improvement compared to the cost of tuning or maintaining a more complex model architecture.

For regulators, the fairness analysis offers a genuinely usable template comparing actual versus predicted default rates by group, and looking at the gap, is simple enough that it could realistically be built into ongoing model monitoring rather than treated as a one-off academic exercise. It's not a formal fairness metric in the technical sense used by Mehrabi et al. (2021), but it's interpretable, and interpretable often wins over rigorous-but-opaque in a regulatory audit context.

For any institution operating across borders, the German validation is a caution against assuming a model that works well in one country will behave the same way or rely on the same signals somewhere else. Recalibration and revalidation aren't optional extras; the SHAP comparison here shows the underlying logic of the model can shift substantially even when performance holds up reasonably well.

5.6 Limitations

A few honest caveats are worth stating clearly.

The datasets are not current. Give Me Some Credit dates back to a 2011 Kaggle competition, and the German Credit Data is older still. Consumer credit behaviour, lending standards, and the

broader economic environment have all changed since these datasets were collected. The patterns identified here particularly the specific SHAP rankings reflect the credit environment at the time the data was gathered, not necessarily today's.

Both datasets are public, not proprietary. Real bank data would include richer behavioural and transactional variables that neither of these datasets capture, which limits how far these specific findings can be extrapolated to what a modern bank's internal model might look like.

The fairness analysis covers only age. Gender, income level, and other potentially relevant demographic dimensions weren't available in the Give Me Some Credit dataset, so this thesis couldn't examine them. A more complete fairness audit would need a dataset with richer demographic coverage.

The age group boundaries (18–35, 36–55, 56–90) were my own choice, not a standardised classification, and a different choice of cut-points could plausibly shift the fairness results somewhat, particularly near the boundaries.

No hyperparameter tuning was performed on any of the three models. This was a deliberate scope decision the focus of this thesis was comparing architectures and applying interpretability and fairness tools, not squeezing out maximum performance from any single model but it does mean the absolute performance numbers reported here likely understate what each model could achieve with proper tuning.

The German sample is small, only 1,000 records, and 200 in the test set which limits how much confidence can be placed in the precise SHAP rankings for that model compared to the US analysis, which had a much larger sample to work with.

None of these limitations undermine the core findings, but they do mean the specific numbers reported in Chapter 4 should be read as evidence from these two particular datasets at this particular point in time, not as universal claims about how credit risk models behave everywhere.

Chapter 6: Conclusion

6.1 Summary of the Study

This thesis sets out to build and evaluate a machine learning approach to credit risk prediction that goes beyond just chasing the highest accuracy score. Three things mattered from the start: whether the model's predictions could be explained, whether it treated different age groups fairly, and whether it worked outside the single country it was trained in. Using the Give Me Some Credit dataset as the primary data source and the German Credit Data as a secondary dataset, the analysis moved through data cleaning, feature engineering, model comparison, SHAP-based interpretability, an age-based fairness audit, and a cross-country comparison each stage building directly on the one before it.

6.2 Answering the Research Questions

Primary question: which model predicts credit risk most effectively, and what drives it?

Random Forest came out ahead on F1 score (0.3266), the metric that mattered most given how imbalanced the dataset was. It wasn't a landslide win, Logistic Regression actually had a slightly higher AUC (0.8144 vs 0.8124) but Random Forest handled the minority class better, which is what actually matters if a bank is trying to catch real defaulters. As for what drives the predictions, SHAP gave a clear answer: DelinquencyScore, RevolvingUtilizationOfUnsecuredLines, and NumberOfDependents were the three biggest factors, in that order.

Sub-question 1 — Did the engineered features add value?

Yes, and more clearly than expected. DelinquencyScore ended up as the single most important variable in the entire model, ranked above all ten original variables. The three raw delinquency columns it was built from dropped to the bottom of the feature importance ranking once the composite score was included, which shows the model relies on the combined measure far more than on its individual components, though confirming this translates into an actual improvement in predictive performance, rather than just a different way of representing the same information, would require a direct comparison against a model built without it.

Sub-question 2 — Does the model treat age groups fairly?

Not entirely. Older Adults showed almost no difference between actual and predicted default rates (−0.11pp, basically noise). Borrowers from Middle-Aged group showed a small overprediction (+1.02pp). Young Adults were the group most affected, with the overestimation of their default risk by 1.92 percentage points. Age itself was not the most important feature in the model (it ranked 6th). However, the present analysis cannot determine whether the observed disparity arises from direct age effects, correlations between age and other predictors, differences in calibration, or other sources of prediction error.

Sub-question 3 — Does the Random Forest approach perform comparably and rely on similar factors once retrained on German data?

Partially yes. Performance held up reasonably well (AUC dropped from 0.8124 to 0.7674, a real but not dramatic decline). But the underlying logic of the model changed substantially. In the US, the model relies on behavioral history, past delinquency and credit usage. In Germany, it relies on structural indicators checking account status, loan duration, savings. Same architecture, different reasoning depending on country's data it was retrained on.

6.3 What This Thesis Contributes

From this work, three things stand out: the Weighted Delinquency Score outperforming every raw variable, a simple and reusable method for auditing fairness, and clear evidence that a model's logic doesn't travel with it across borders. All discussed in detail in Section 5.5.

6.4 Future Research Directions

A few natural next steps came up while working through this thesis. Testing whether the same DelinquencyScore-style feature engineering approach helps on other credit datasets would confirm whether the result here is a general pattern or something specific to this dataset. The fairness analysis could be extended beyond age to gender or income level if a richer dataset were available, and it could also be tightened up using formal fairness metrics like demographic parity or equalised odds rather than the descriptive gap measure used here. Hyperparameter tuning was deliberately left out of scope in this thesis, so there's an open question about whether a properly tuned model would show the same fairness pattern or a different one. And the cross-country comparison here only covers two countries extending it to more national datasets or trying an actual transfer-learning approach rather than retraining from scratch, would give a fuller picture of how portable these models really are.

References

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, *23*(4), 589–609.

Bequé, A., & Lessmann, S. (2017). Extreme learning machines for credit scoring: An empirical evaluation. *Expert Systems with Applications*, *86*, 42–53.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth International Group.

Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2021). Transparency, auditability, and explainability of machine learning models in credit scoring. *Journal of the Operational Research Society, 73*(1), 70–90. https://doi.org/10.1080/01605682.2021.1922098

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

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM.

Chen, Y., Calabrese, R., & Martin-Barragan, B. (2024). Interpretable machine learning for imbalanced credit scoring datasets. *European Journal of Operational Research, 312*(1), 357–372. https://doi.org/10.1016/j.ejor.2023.06.036

European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union.

European Parliament and Council. (2016). *Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation)*. Official Journal of the European Union.

European Parliament and Council. (2026). *Regulation (EU) 2026/1744 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI)*. Official Journal of the European Union.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *Journal of Finance*, *77*(1), 5–47.

Kaggle. (2011). *Give Me Some Credit* [Data set].

Hofmann, H. (1994). *Statlog (German Credit Data)* [Data set]. UCI Machine Learning Repository.

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.

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

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6), 1–35.

OQ v Land Hessen (SCHUFA), Case C-634/21, ECLI:EU:C:2023:957 (Court of Justice of the European Union, 7 December 2023).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Shapley, L. S. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games* (Vol. 2, pp. 307–317). Princeton University Press.

Schmitt, M. (2022). Deep learning vs. gradient boosting: Benchmarking state-of-the-art machine learning algorithms for credit scoring. *arXiv preprint arXiv:2205.10535*.

West, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, *27*(11–12), 1131–1152.

Wube, H. D., Esubalew, S. Z., Weldesellasie, F. F., & Debelee, T. G. (2024). Deep learning and machine learning techniques for credit scoring: A review. In *Pan-African Conference on Artificial Intelligence* (Communications in Computer and Information Science, Vol. 2069). Springer.

Appendix

Appendix A: Full Classification Reports

The complete classification reports for all three models trained on the US dataset, and for the Random Forest model retrained on the German dataset, are provided below in full, including per-class precision, recall, F1-score, and support values that were summarized in Chapter 4.

A.1 Logistic Regression (Scaled) — US Dataset

Class	Precision	Recall	F1-Score	Support
0 (No Default)	0.97	0.83	0.90	27,367
1 (Default)	0.19	0.63	0.29	1,752
Accuracy			0.82	29,119
Macro avg	0.58	0.73	0.59	29,119
Weighted avg	0.93	0.82	0.86	29,119

AUC-ROC: 0.8144 | F1 Score: 0.2936

A.2 Random Forest — US Dataset

Class	Precision	Recall	F1-Score	Support
0 (No Default)	0.96	0.95	0.95	27,367
1 (Default)	0.31	0.34	0.33	1,752
Accuracy			0.91	29,119
Macro avg	0.63	0.65	0.64	29,119
Weighted avg	0.92	0.91	0.92	29,119

AUC-ROC: 0.8124 | F1 Score: 0.3266

A.3 XGBoost — US Dataset

Class	Precision	Recall	F1-Score	Support
0 (No Default)	0.96	0.95	0.95	27,367

1 (Default)	0.28	0.34	0.31	1,752
Accuracy			0.91	29,119
Macro avg	0.62	0.64	0.63	29,119
Weighted avg	0.92	0.91	0.91	29,119

AUC-ROC: 0.8066 | F1 Score: 0.3071

A.4 Random Forest — German Dataset

Class	Precision	Recall	F1-Score	Support
0 (Bad Credit)	0.51	0.60	0.55	60
1 (Good Credit)	0.82	0.76	0.79	140
Accuracy			0.71	200
Macro avg	0.66	0.68	0.67	200
Weighted avg	0.73	0.71	0.72	200

AUC-ROC: 0.7674 | F1 Score: 0.72 (weighted) / 0.79 (Good Credit class)

Appendix B: Code Availability

The complete Python code used for this analysis: covering data cleaning, feature engineering, model training, SHAP interpretability analysis, the age-based fairness analysis, and the German cross-validation is provided in the accompanying Google Colab notebook.