EDC PARIS BUSINESS SCHOOL

MSc Data Science and Business Analysis

Academic Year 2025–2026

MASTER'S THESIS

Explainable AI for Financial Inclusion: A Machine Learning and Interpretable Credit Scoring Framework with Applications to Underbanked Populations

Submitted by: Nanduri Kushath Sri Krishna Rao

Thesis Supervisor: BA Adama

Date of Submission: 22/05/2026

Abstract

Traditional credit scoring models, constructed upon bureau-reported credit histories and logistic regression, structurally exclude approximately 1.4 billion adults worldwide who lack formal financial records. This exclusion perpetuates a cyclical barrier to economic participation: individuals without credit histories cannot access credit, and without credit access, they cannot build histories. The expansion of financial technology (FinTech) has generated unprecedented volumes of alternative behavioural and transactional data including e-commerce activity, utility payments, and mobile usage patterns, that could serve as proxies for creditworthiness. Simultaneously, advances in machine learning (ML) have demonstrated that ensemble algorithms such as XGBoost and LightGBM significantly outperform logistic regression in predicting credit default risk. However, the opacity of these ML models creates critical barriers to regulatory compliance, ethical deployment, and managerial trust. This tension is particularly acute under the European Union Artificial Intelligence Act, which classifies credit scoring as a high-risk AI application requiring transparency and the provision of meaningful explanations to affected individuals.

This thesis develops an Explainable Inclusive Credit Scoring (EICS) Framework that integrates three pillars, high-performance ML prediction, SHAP-based explainable AI (XAI), and algorithmic fairness evaluation all into a unified decision-support tool for FinTech institutions. Using real-world lending data from the Lending Club dataset (2.26 million loan records, 2007–2018) and the Home Credit Default Risk dataset (307,000 applications), the study compares Logistic Regression, Random Forest, XGBoost, and LightGBM in predicting credit default. Feature importance is decomposed through SHAP analysis to provide transparent, regulation-compliant explanations. Fairness is evaluated across demographic subgroups using established metrics including Demographic Parity and Equalised Odds. The findings contribute to academic knowledge and professional practice by proposing a deployable framework that reconciles predictive accuracy with regulatory transparency and financial inclusion.

Keywords: *credit scoring, machine learning, explainable AI, SHAP, financial inclusion, algorithmic fairness, FinTech, EU AI Act, underbanked populations, XGBoost, LightGBM*

Résumé

Les modèles traditionnels de notation de crédit, fondés sur les historiques déclarés aux agences et sur la régression logistique, excluent structurellement environ 1,4 milliard d'adultes dans le monde ne disposant pas de dossiers financiers formels. Cette exclusion perpétue un obstacle cyclique à la participation économique : sans historique de crédit, les individus ne peuvent accéder au crédit, et sans accès au crédit, ils ne peuvent constituer d'historique. L'essor de la technologie financière (FinTech) a généré des volumes inédits de données comportementales et transactionnelles alternatives, notamment l'activité de commerce électronique, les paiements de services publics et les schémas d'utilisation mobile, pouvant servir d'indicateurs de solvabilité. Parallèlement, les avancées en apprentissage automatique (ML) ont démontré que les algorithmes d'ensemble tels que XGBoost et LightGBM surpassent significativement la régression logistique dans la prédiction du risque de défaut. Toutefois, l'opacité de ces modèles crée des obstacles critiques à la conformité réglementaire, au déploiement éthique et à la confiance managériale. Cette tension est particulièrement vive au regard du Règlement européen sur l'intelligence artificielle, qui classe la notation de crédit parmi les applications d'IA à haut risque exigeant transparence et explications significatives aux personnes concernées.

Ce mémoire développe un cadre de notation de crédit explicable et inclusif (EICS) intégrant trois piliers, prédiction ML à haute performance, IA explicable fondée sur SHAP et évaluation de l'équité algorithmique, en un outil d'aide à la décision unifié destiné aux institutions FinTech. À partir des bases Lending Club (2,26 millions de dossiers, 2007–2018) et Home Credit Default Risk (307 000 demandes), l'étude compare la régression logistique, la forêt aléatoire, XGBoost et LightGBM pour la prédiction du défaut. L'importance des variables est décomposée par analyse SHAP afin de fournir des explications transparentes et conformes. L'équité est évaluée entre sous-groupes démographiques via la parité démographique et l'égalité des chances. Les résultats contribuent aux connaissances académiques et à la pratique en proposant un cadre déployable conciliant précision prédictive, transparence réglementaire et inclusion financière.

Mots-clés : notation de crédit, apprentissage automatique, IA explicable, SHAP, inclusion financière, équité algorithmique, FinTech, Règlement IA de l'UE, populations sous-bancarisées, XGBoost, LightGBM

Acknowledgements

I would like to express my sincere gratitude to my thesis supervisor, BA Adama, for the guidance, expertise, and constructive feedback provided throughout this research process. The intellectual rigour and scholarly direction offered at each stage of this work have been invaluable.

I also wish to thank the faculty and staff of EDC Paris Business School for creating an academic environment that encourages independent thinking, methodological ambition, and the pursuit of research that bridges academic inquiry with real-world impact.

Finally, I extend my appreciation to my family, friends, and peers whose encouragement and support sustained me throughout this demanding but rewarding journey.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence

AUC-ROC Area Under the Receiver Operating Characteristic Curve

CFPB Consumer Financial Protection Bureau

ECOA Equal Credit Opportunity Act (United States)

EICS Explainable Inclusive Credit Scoring (Framework)

EU European Union

FICO Fair Isaac Corporation

FinTech Financial Technology

GDPR General Data Protection Regulation

GOSS Gradient-based One-Side Sampling

LightGBM Light Gradient Boosting Machine

LIME Local Interpretable Model-Agnostic Explanations

MCC Matthews Correlation Coefficient

ML Machine Learning

PII Personally Identifiable Information

SDG Sustainable Development Goal

SHAP SHapley Additive exPlanations

SMOTE Synthetic Minority Oversampling Technique

XAI Explainable Artificial Intelligence

XGBoost Extreme Gradient Boosting

Chapter 1: Introduction

1.1 Background and Context

Financial inclusion, defined by the World Bank as the ability of individuals and businesses to access useful and affordable financial products and services that meet their needs, delivered in a responsible and sustainable way, remains one of the most pressing socioeconomic challenges of the twenty-first century. The most recent edition of the World Bank's Global Findex Database reveals that approximately 1.4 billion adults worldwide remain unbanked, lacking access to even a basic transaction account at a financial institution or through a mobile money provider (Demirgüç-Kunt et al., 2022). This figure, while representing a notable improvement from the 2.5 billion unbanked adults recorded in 2011, nonetheless signals that roughly one-quarter of the global adult population remains outside the formal financial system. The consequences of this exclusion extend far beyond the inability to open a bank account. Without access to formal credit, individuals cannot finance education, invest in small businesses, smooth consumption during economic shocks, or build the asset base necessary to escape poverty. Beck et al. (2007), drawing on cross-country panel data, demonstrated that financial development disproportionately benefits the poorest quintile of the income distribution, implying that financial exclusion is not merely a symptom of poverty but an active mechanism that perpetuates it.

At the structural core of this exclusion lies a fundamental information problem. The credit scoring systems that underpin modern lending decisions, pioneered by the Fair Isaac Corporation (FICO) in the United States in the 1950s and subsequently adopted, in various forms, across global financial markets which are built upon a specific type of data: bureau-reported credit histories. These histories record an individual's prior interactions with formal credit markets: loan repayment patterns, credit card utilisation rates, mortgage payment regularity, the number and types of open credit accounts, and the presence of negative events such as bankruptcies or collection actions. Using these inputs, traditional credit scoring models, most commonly logistic regression, the dominant method since the work of Hand and Henley (1997) and the comprehensive treatment by Thomas et al. (2002) do estimate the probability that a given

borrower will default within a specified time horizon. The output is a numerical score that lenders use to make binary decisions: approve or deny, and at what interest rate.

The critical limitation of this system becomes apparent when one considers who it excludes. For the 1.4 billion adults who have never interacted with formal credit markets, are variously described in the literature as “thin-file” borrowers (those with limited credit history), “no-file” borrowers (those with no credit history at all), or “credit-invisible” individuals (those who do not appear in any credit bureau’s records) the traditional scoring model cannot function. It requires as input precisely the data that these individuals, by definition, do not possess. The result is a paradox that scholars of financial inclusion have long identified: individuals cannot access credit because they lack a credit history, and they cannot build a credit history because they lack access to credit (Agarwal et al., 2020). In the United States alone, the Consumer Financial Protection Bureau (CFPB, 2015) estimated that approximately 26 million adults are “credit invisible” they have no credit record at any of the three major credit bureaus and an additional 19 million have credit records that are considered “unscorable” by current models. In developing economies, where formal credit infrastructure is less mature, the proportion of the population excluded from traditional scoring is substantially higher.

The rise of financial technology (FinTech) over the past decade has introduced a potentially transformative disruption to this paradigm. FinTech lenders who are operating predominantly through digital platforms and often outside the regulatory frameworks that govern traditional banks, have begun to experiment with alternative data sources to evaluate creditworthiness in the absence of bureau data. These alternative data sources include mobile phone usage patterns (call frequency, airtime top-ups, the types of applications installed on a device), e-commerce transaction histories (purchase regularity, average transaction value, return rates), utility payment records (electricity, water, and telecommunications bill payment consistency), and even social network characteristics (Berg et al., 2020; Gambacorta et al., 2019). The underlying premise is that these behavioural traces are generated through ordinary daily activities which contain latent information about an individual’s reliability, financial discipline, and capacity to repay a loan, even when no formal credit history exists.

Simultaneously, advances in machine learning (ML) have demonstrated that algorithms capable of modelling non-linear relationships and complex feature interactions such as Random Forest, Gradient Boosting Machines (including XGBoost and LightGBM), and neural networks can significantly outperform the logistic regression models that have been the industry standard for decades. The landmark benchmark study by Lessmann et al. (2015), which systematically compared 41 classification algorithms across eight real-world credit scoring datasets, demonstrated that ensemble methods achieve statistically significant improvements in discriminatory power (AUC-ROC) over logistic regression, with improvements ranging from 2 to 7 percentage points depending on the dataset and the specific algorithm. When these ML methods are combined with alternative data, the potential to serve previously unscorable populations becomes concrete, a proposition supported by empirical evidence from Berg et al. (2020), Agarwal et al. (2020), and Gambacorta et al. (2019), each of which is discussed in detail in Chapter 2.

However, the deployment of ML-based credit scoring models introduces a critical and unresolved tension between predictive performance and transparency. While logistic regression, despite its statistical limitations, offers a transparent decision logic each coefficient directly indicates the direction and magnitude of a feature's contribution to the predicted probability, ensemble ML models operate as what the literature terms "black boxes." Their predictions emerge from the aggregation of hundreds or thousands of decision trees, each splitting on different features at different thresholds, producing an output that cannot be reduced to a simple, human-interpretable formula. For a loan officer who must justify a lending decision to a customer, for a regulator who must audit algorithmic fairness, or for a rejected applicant who wishes to understand why they were denied credit, knowing that an algorithm assigned a particular risk score is insufficient. Understanding why the algorithm arrived at that score, which specific factors contributed, in what direction, and by how much is essential for trust, accountability, and legal compliance.

This transparency imperative has become particularly acute in the European regulatory context. The European Union's Artificial Intelligence Act (Regulation (EU) 2024/1689), which entered into force in 2024 after years of legislative deliberation, explicitly classifies AI systems used for credit scoring and the assessment of creditworthiness as "high-risk" under Annex III. High-risk

AI systems are subject to stringent requirements including the implementation of risk management systems, the maintenance of data governance standards, the provision of transparency and information to deployers and users, the establishment of human oversight mechanisms, and the demonstration of accuracy, robustness, and cybersecurity. Article 13 of the AI Act specifically mandates that high-risk AI systems "shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately." Article 86 further establishes the right of affected persons to obtain "clear and meaningful explanations" of the role of the AI system in the decision-making procedure and the main elements of the decision taken. The General Data Protection Regulation (GDPR), which has been in force since 2018, reinforces this through Article 22 and Recital 71, which address the right to meaningful information about the logic involved in automated decision-making that significantly affects individuals.

The field of Explainable Artificial Intelligence (XAI) has emerged as the primary response to this demand for transparency. Techniques such as SHAP (SHapley Additive exPlanations), developed by Lundberg and Lee (2017) on the foundation of Shapley values from cooperative game theory (Shapley, 1953), and LIME (Local Interpretable Model-Agnostic Explanations), introduced by Ribeiro et al. (2016), enable the decomposition of individual model predictions into interpretable feature contributions. These methods make it possible, in principle, to explain precisely why a specific applicant was approved or denied, transforming a black-box output into a transparent, auditable decision narrative. Yet, a further dimension complicates the deployment of ML in credit scoring: algorithmic fairness. Credit scoring models trained on historical data may inherit and amplify the biases embedded in that data, biases that reflect decades of discriminatory lending practices, structural inequalities in economic opportunity, and systematic disparities in data availability across demographic groups (Barocas & Selbst, 2016; Mehrabi et al., 2021). Ensuring that ML-based credit models are not only accurate and explainable but also fair across demographic lines is therefore a prerequisite for their responsible and sustainable deployment.

1.2 Research Problem and Justification

Despite significant progress in each of these individual domains, machine learning for credit scoring, explainable AI, and algorithmic fairness, the academic literature reveals a persistent and consequential gap at their intersection. This gap is not merely a matter of academic incompleteness; it has direct operational, regulatory, and societal implications that this thesis seeks to address.

The fragmentation manifests in three ways. First, the majority of benchmark studies that compare ML algorithms for credit scoring focus exclusively on predictive performance metrics which are AUC-ROC, accuracy, F1-score, without incorporating explainability or fairness into their evaluation frameworks. The influential study by Lessmann et al. (2015), for example, compares 41 classification algorithms across eight datasets but does not assess whether the top-performing models can produce interpretable explanations or whether their predictions are equitable across demographic groups. Similarly, Bao et al. (2019), who integrate unsupervised and supervised ML for credit risk, optimise purely for predictive accuracy. This performance-centric paradigm implicitly treats credit scoring as a purely technical optimisation problem, ignoring the reality that lending decisions affect human lives and are subject to legal and ethical scrutiny.

Second, the XAI literature as applied to finance remains largely conceptual or limited in empirical scope. The comprehensive survey by Arrieta et al. (2020) provides an excellent taxonomy of XAI methods and their potential applications across domains, but does not empirically validate these methods on large-scale real-world credit data. Where empirical applications exist, they tend to demonstrate explainability techniques on small or synthetic datasets without integrating fairness evaluation. The gap is not in the availability of XAI methods, SHAP and LIME are well-developed and computationally tractable but in their integration into end-to-end credit scoring frameworks that simultaneously address prediction, transparency, and equity.

Third, the algorithmic fairness literature, while theoretically sophisticated, often addresses bias detection and mitigation in isolation from the practical requirements of model deployment, regulatory compliance, and stakeholder communication. Chouldechova (2017) and Hardt et al.

(2016) establish important impossibility results and fairness definitions, but their work operates at a level of abstraction that does not directly translate into operational guidance for a FinTech institution seeking to comply with the EU AI Act while maintaining competitive predictive performance.

No comprehensive framework currently exists that integrates all three dimensions of high-performance ML prediction, SHAP-based regulatory-grade explainability, and fairness-aware evaluation, into a unified, deployable artifact validated on real-world lending data at scale. This gap has direct consequences: for financial institutions seeking to comply with the EU AI Act, which demands both accuracy and transparency; for regulators tasked with auditing algorithmic lending decisions for discriminatory outcomes; and for the 1.4 billion individuals whose access to credit depends on the development of assessment mechanisms that are simultaneously inclusive, transparent, and fair.

The relevance of this study is therefore threefold. Academically, it contributes to the credit risk, XAI, and algorithmic fairness literatures by bridging them through an integrated empirical investigation on real-world data comprising over 2.5 million records. Professionally, it produces a deployable Explainable Inclusive Credit Scoring (EICS) Framework that FinTech institutions, banks, and microfinance organisations can adapt to implement regulation-compliant, transparent, and inclusive lending practices. Societally, it addresses financial inclusion, recognised by the United Nations as integral to Sustainable Development Goal 1 (No Poverty) and Sustainable Development Goal 10 (Reduced Inequalities) by demonstrating how data science, deployed responsibly, can extend credit access to populations that traditional models structurally exclude.

1.3 Research Questions and Objectives

This thesis is guided by the following overarching research question:

To what extent can explainable machine learning models, applied to alternative and traditional credit data, improve credit risk assessment accuracy while maintaining regulatory transparency and promoting financial inclusion for underbanked populations?

This main research question is deliberately broad, encompassing the three pillars prediction, explainability, and fairness that the thesis seeks to integrate. To operationalise this overarching question into testable and measurable components, four sub-questions are formulated:

Sub-Question 1 (Prediction): How do ensemble machine learning models, specifically Random Forest, XGBoost, and LightGBM, compared to the traditional logistic regression baseline in predicting credit default risk on real-world lending data, as measured by AUC-ROC, F1-score, precision, recall, and Matthews Correlation Coefficient?

Sub-Question 2 (Explainability): What are the most influential features driving credit risk predictions in the best-performing model, and how can SHAP-based explainability decompose these predictions into transparent, human-interpretable decision narratives that satisfy the EU AI Act's requirements for meaningful explanations?

Sub-Question 3 (Fairness): To what extent do the ML-based credit scoring models exhibit disparities in prediction outcomes across demographic subgroups, and what post-processing mitigation strategies, particularly threshold calibration can be applied to reduce algorithmic bias while quantifying the resulting impact on predictive accuracy?

Sub-Question 4 (Integration): How can the three pillars of prediction, explainability, and fairness be synthesised into an integrated Explainable Inclusive Credit Scoring (EICS) Framework that translates predictive insights into actionable, regulation-compliant lending decisions for FinTech institutions?

Correspondingly, the research pursues five objectives:

1. To critically review the evolution of credit scoring from classical statistical methods to contemporary ML-based approaches, situating the study within the theoretical frameworks of information asymmetry, financial inclusion, and AI regulation.

2. To design and implement a rigorous quantitative methodology that compares Logistic Regression, Random Forest, XGBoost, and LightGBM on real-world lending datasets comprising over 2.5 million records, with appropriate treatment of class imbalance, feature engineering, and hyperparameter optimization.

3. To apply SHAP explainability analysis at both global and local levels, decomposing model predictions into feature-level contributions that can serve as the basis for regulation-compliant credit decision narratives.

4. To evaluate model fairness across available demographic subgroups using established metrics (Demographic Parity Ratio, Equalised Odds Difference, Disparate Impact Ratio) and to apply and assess post-processing bias mitigation strategies.

5. To develop an original Explainable Inclusive Credit Scoring (EICS) Framework that integrates prediction, explanation, and fairness evaluation into a unified, deployable decision-support architecture for FinTech lending.

Additionally, the thesis tests four hypotheses that correspond to the sub-questions and provide falsifiable empirical claims:

H1: Ensemble machine learning models (XGBoost, LightGBM, Random Forest) achieve statistically significantly higher predictive performance (AUC-ROC, F1-score) than logistic regression for credit default prediction on real-world lending data.

H2: SHAP-based feature attribution analysis reveals that alternative behavioural features (e.g., loan purpose patterns, income-to-debt dynamics, employment stability proxies) contribute significant predictive power beyond traditional bureau variables, as measured by their mean absolute SHAP values.

H3: Fairness-aware model evaluation reveals measurable disparities in prediction outcomes across demographic subgroups, which can be partially mitigated through post-processing threshold calibration without statistically significant loss in overall predictive accuracy.

H4: An integrated explainable credit scoring framework improves decision transparency measured by explanation fidelity and consistency metrics relative to a non-explainable baseline, without statistically significant reduction in AUC-ROC.

1.4 Structure of the Thesis

This thesis is organised into six chapters, each serving a distinct function within the overall research architecture. The present chapter has established the context, identified the problem, stated the research questions and objectives, and formulated the hypotheses that guide the empirical investigation.

Chapter 2 presents a critical review of the academic literature organised into four thematic domains: (i) credit risk theory, traditional scoring methods, and the mechanisms of financial exclusion; (ii) machine learning applications in credit scoring, including ensemble methods and the role of alternative data; (iii) explainable artificial intelligence, with particular attention to SHAP and the regulatory requirements of the EU AI Act; and (iv) algorithmic fairness in credit decisions, including fairness definitions, bias sources, and mitigation strategies. The chapter culminates in the identification of the research gap and the construction of the conceptual framework, the three-pillar EICS model that underpins the empirical work.

Chapter 3 details the research methodology, including the positivist research philosophy, quantitative research design, data sources and their justification, the complete data preprocessing pipeline (missing value treatment, feature engineering, class imbalance handling, encoding and scaling), the model training and evaluation protocol (stratified k-fold cross-validation, hyperparameter tuning, statistical significance testing), the SHAP explainability and fairness evaluation procedures, ethical considerations, and tool selection.

Chapter 4 presents the findings and results of the analysis in a systematic, objective manner, using tables, figures, and statistical tests. It reports descriptive statistics, model performance comparisons, SHAP explainability results (global and local), fairness evaluation outcomes, and the robustness check on the supplementary Home Credit dataset.

Chapter 5 provides the discussion, interpreting the findings in light of the literature reviewed in Chapter 2, presenting the complete EICS Framework architecture, and addressing theoretical contributions, managerial and policy implications, limitations, and avenues for future research.

Chapter 6 concludes with a synthesis of key insights, actionable recommendations for practitioners and policymakers, and a closing reflection on the contribution of this research to the intersection of data science, financial inclusion, and responsible AI.

Chapter 2: Literature Review

This chapter presents a critical synthesis of the academic literature that underpins the research. It is organised into five thematic sections that progressively construct the theoretical and empirical foundation upon which the Explainable Inclusive Credit Scoring (EICS) Framework is built. Section 2.1 examines the theoretical foundations of credit risk assessment, tracing the evolution from classical quantitative models to logistic regression and situating financial exclusion within information asymmetry theory. Section 2.2 reviews the empirical evidence on machine learning in credit scoring, analysing the performance of ensemble methods, the role of alternative data, and the methodological challenges that must be addressed. Section 2.3 explores the field of Explainable AI, focusing on the SHAP framework and the regulatory environment created by the EU AI Act. Section 2.4 examines algorithmic fairness in credit decisions, including competing definitions of fairness, sources of bias, and mitigation strategies. Section 2.5 synthesises the preceding analysis to identify the research gap and present the conceptual framework that guides the empirical investigation.

The review draws upon a minimum of 30 peer-reviewed academic sources, supplemented by institutional publications from the World Bank, the Bank for International Settlements, and the European Union. Sources are primarily from the period 2015–2025, with foundational works from earlier decades included where their theoretical contributions remain central to the field.

2.1 Credit Risk Assessment: Theoretical Foundations

2.1.1 The Concept of Credit Risk and Its Economic Significance

Credit risk, the probability that a borrower will fail to meet their contractual obligations to repay a debt is a foundational concept in financial theory and a primary concern of lending institutions worldwide. The effective management of credit risk is not merely a matter of institutional profitability; it determines the efficiency of capital allocation within an economy, influences the structure of interest rates, and shapes the availability of credit to individuals, small businesses, and corporations. When credit risk is poorly assessed, the consequences can be severe: at the

micro level, individual lenders suffer losses; at the macro level, systemic misallocation of credit can contribute to financial crises, as demonstrated by the global financial crisis of 2007–2009, which was precipitated in significant part by inadequate credit risk assessment in the subprime mortgage market (Altman & Saunders, 1997).

The academic study of credit risk assessment can be traced to two foundational contributions that established distinct but complementary analytical traditions. Altman (1968) introduced the Z-Score model for predicting corporate bankruptcy, employing multiple discriminant analysis (MDA) to combine five financial ratios working capital to total assets, retained earnings to total assets, earnings before interest and taxes to total assets, market value of equity to book value of total debt, and sales to total assets into a single composite score. The model achieved a classification accuracy of approximately 95% one year prior to bankruptcy in its original sample and has been extensively validated across subsequent decades and geographies. Altman's contribution established the principle that quantitative statistical models could systematically outperform subjective expert judgement in credit assessment, a finding that would prove foundational for the entire field of automated credit scoring.

Merton (1974) complemented this with a structural approach, modelling default as a function of the relationship between a firm's asset value and its debt obligations. Drawing on the Black-Scholes option pricing framework, Merton treated a firm's equity as a call option on its assets, with the strike price equal to the face value of its debt. Default occurs when the value of the firm's assets falls below the debt threshold at maturity. While the Merton model was developed for corporate credit risk rather than consumer lending, its conceptual contribution is significant: it formalised the idea that default is a probabilistic event that can be modelled as a function of observable financial variables, subject to stochastic uncertainty. Together, the Altman and Merton traditions established credit risk assessment as a quantitative discipline grounded in statistical inference, a discipline that would evolve dramatically over the following decades.

2.1.2 Traditional Credit Scoring: The Logistic Regression Paradigm

While the Altman and Merton models addressed corporate credit risk, the parallel development of consumer credit scoring followed a distinct trajectory. Beginning in the 1960s with the adoption

of linear discriminant analysis and progressing through various statistical techniques, consumer credit scoring converged by the 1990s on logistic regression as the dominant methodology (Hand & Henley, 1997). The logistic regression model estimates the probability of a binary outcome default (1) or non-default (0) as a function of a linear combination of input features, mapped through the logistic (sigmoid) function to constrain the output to the [0, 1] probability interval. Thomas et al. (2002), in their comprehensive treatment of credit scoring methods, documented the widespread adoption of logistic regression across the retail banking, consumer finance, and credit card industries, noting its advantages in terms of interpretability, computational simplicity, and regulatory acceptability.

The appeal of logistic regression for credit scoring rests on several specific properties. First, each model coefficient can be directly interpreted as the change in the log-odds of default associated with a one-unit change in the corresponding input feature, holding all other features constant. This interpretability is not merely an academic convenience; it is a regulatory requirement in many jurisdictions. Under the United States' Equal Credit Opportunity Act (ECOA) and its implementing regulation (Regulation B), lenders who deny credit must provide the applicant with the specific reasons for the adverse decision. A logistic regression model's coefficients directly support this requirement by identifying the features that most contributed to the denial. Second, logistic regression produces calibrated probability estimates, meaning that its output can be directly interpreted as a probability rather than a score requiring external calibration. Third, the model is computationally efficient and can be deployed at scale with minimal infrastructure, a significant advantage in the pre-cloud computing era when many of these systems were first built.

However, logistic regression carries inherent limitations that become increasingly consequential as lending environments grow more complex and data-rich. The most fundamental limitation is the assumption of linearity: logistic regression assumes that the relationship between each input feature and the log-odds of default is linear. This assumption is empirically violated in numerous real-world credit scoring contexts. For example, the relationship between a borrower's debt-to-income ratio and default probability is typically non-linear, risk may increase slowly at low ratios, accelerate sharply as the ratio approaches a critical threshold, and plateau at very high values. Similarly, interaction effects between features such as the combined impact of high

revolving credit utilisation and recent credit inquiries which are not captured by the additive structure of logistic regression unless explicitly engineered by the analyst. Lessmann et al. (2015) provide systematic empirical evidence that this linearity assumption imposes a performance ceiling that more flexible algorithms can surpass.

The second critical limitation of logistic regression is its dependence on the availability of bureau-reported credit data. The model's inputs, prior credit accounts, payment histories, delinquencies, utilisation rates, length of credit history are drawn from credit bureau records. For thin-file and no-file borrowers who lack these records, the model simply cannot be computed. This is not a failure of the algorithm per se; it is a structural feature of the scoring paradigm that renders traditional credit scoring fundamentally incapable of serving populations without prior credit market participation. The significance of this limitation becomes apparent when one considers its scale: as documented by the World Bank (Demirgüç-Kunt et al., 2022), the populations excluded by this data requirement number in the hundreds of millions globally.

2.1.3 Information Asymmetry Theory and the Mechanism of Financial Exclusion

The problem of financial exclusion through traditional credit scoring can be understood more precisely through the lens of information asymmetry theory, a body of economic theory that examines how unequal distribution of information between parties to a transaction affects market outcomes. Akerlof's (1970) seminal analysis of "The Market for Lemons" demonstrated that in markets where sellers possess information about product quality that buyers lack, a process of adverse selection can unfold: high-quality sellers withdraw from the market because they cannot command prices commensurate with their quality, leaving only low-quality sellers, which in turn further depresses buyers' willingness to pay. In extreme cases, the market may collapse entirely.

Applied to credit markets, Akerlof's framework illuminates why lenders cannot simply offer credit to all applicants at appropriately risk-adjusted prices. When lenders cannot observe the true creditworthiness of applicants because the information infrastructure (credit bureaus, documented credit histories) that would make this observable does not cover certain populations, they face a version of the lemons problem. If they offer credit at an average interest rate, they attract a disproportionate share of high-risk borrowers (who benefit from the below-risk pricing) while

low-risk borrowers (who would be penalised by above-risk pricing) may self-select out. The rational response, as Akerlof demonstrated, is either to refuse to lend entirely or to charge prohibitively high rates, both of which effectively exclude the underbanked population from formal credit access.

Stiglitz and Weiss (1981) extended this analysis in a contribution that is directly relevant to the present study. They demonstrated formally that credit rationing, the systematic denial of credit to certain categories of borrowers, even when those borrowers would be willing to accept higher interest rates, is a rational equilibrium response by lenders facing imperfect information. Raising interest rates, Stiglitz and Weiss showed, does not simply reduce demand for credit; it also changes the composition of the borrower pool through two mechanisms. The adverse selection effect means that higher rates disproportionately deter low-risk borrowers, who have better outside options. The moral hazard effect means that higher rates incentivise borrowers to pursue riskier projects with higher potential returns. Both effects worsen the quality of the lender's loan portfolio, potentially reducing expected profits even as the nominal interest rate increases. The implication is stark: in the absence of information that would allow lenders to distinguish between high-risk and low-risk borrowers, entire populations are excluded from credit markets not because they are inherently uncreditworthy, but because lenders cannot verify their creditworthiness with the available data.

This theoretical insight provides the intellectual foundation for the present study. If the core problem is information asymmetry, if lenders exclude underbanked populations because they lack information about their creditworthiness, not because these populations are inherently risky, then any method that reduces information asymmetry has the potential to unlock credit access. Alternative data sources and advanced ML methods represent precisely such information-generating mechanisms. By extracting predictive signals from behavioural data that thin-file borrowers do generate (mobile phone usage, utility payments, e-commerce transactions), ML models can reduce the information gap that Akerlof and Stiglitz identified as the root cause of financial exclusion.

2.1.4 The Scale and Geography of Financial Exclusion

The empirical reality of financial exclusion, as documented by the World Bank's Global Findex Database, gives concrete magnitude to the theoretical problem. Demirgüç-Kunt et al. (2022) report that the global unbanked population decreased from approximately 2.5 billion in 2011 to 1.4 billion in 2021, a significant improvement driven largely by mobile money expansion in Sub-Saharan Africa and government-led financial inclusion initiatives in South Asia. Nevertheless, the remaining 1.4 billion unbanked adults are concentrated in regions and demographic groups that traditional credit scoring is structurally unable to serve: Sub-Saharan Africa (where only 55% of adults have an account), South Asia, and the Middle East and North Africa. Women are disproportionately represented among the unbanked: the gender gap in account ownership, while narrowing, persists across most developing economies. Rural residents, lower-income individuals, and young adults are similarly overrepresented.

Even in developed economies, a significant "credit-invisible" population exists. The Consumer Financial Protection Bureau (CFPB, 2015) estimated that approximately 26 million adults in the United States lack a credit file at any of the three major credit bureaus (Equifax, Experian, TransUnion), and an additional 19 million have files that are considered "unscorable" by conventional models due to insufficient or stale information. These 45 million individuals, roughly 19% of the adult US population, are effectively invisible to the traditional credit scoring infrastructure. They are disproportionately Black and Hispanic, live in lower-income neighbourhoods, and are younger than the average credit-visible population, suggesting that credit invisibility intersects with and reinforces pre-existing socioeconomic disparities.

Financial inclusion has been recognised at the highest levels of international policy as a critical enabler of sustainable development. It is embedded in the United Nations' 2030 Agenda for Sustainable Development, where it appears as an enabler across multiple Sustainable Development Goals: SDG 1 (No Poverty), SDG 2 (Zero Hunger, through agricultural credit), SDG 5 (Gender Equality, through women's financial empowerment), SDG 8 (Decent Work and Economic Growth), and SDG 10 (Reduced Inequalities). The World Bank has made "Universal Financial Access" a strategic priority, and the G20's Global Partnership for Financial Inclusion

coordinates policy efforts across major economies. These institutional commitments underscore the significance of the problem that this thesis addresses: the development of credit assessment mechanisms that can serve the excluded without compromising risk management.

2.2 Machine Learning in Credit Scoring: Empirical Evidence

2.2.1 The Paradigm Shift from Statistical Models to Machine Learning

The limitations of logistic regression documented in Section 2.1, combined with the exponential growth in available data and computational power, have driven a paradigm shift in credit scoring research toward machine learning methods over the past two decades. Machine learning algorithms differ from traditional statistical models in a fundamental respect: rather than requiring the analyst to specify the functional form of the relationship between inputs and outputs (as logistic regression requires linearity in the log-odds), ML algorithms learn this relationship directly from the data, allowing them to capture non-linear patterns, threshold effects, and complex multi-way interactions without explicit specification (Hastie et al., 2009). This flexibility is both the source of ML's superior predictive performance and the origin of its interpretability challenge, a tension that is central to this thesis.

The landmark benchmark study by Lessmann et al. (2015) represents the most comprehensive comparative evaluation of classification algorithms for credit scoring in the literature. The authors evaluated 41 algorithms, including logistic regression, linear and quadratic discriminant analysis, k-nearest neighbours, support vector machines, decision trees, Random Forest, Gradient Boosting, and neural networks, across eight real-world credit scoring datasets from multiple countries and lending contexts. Their findings were decisive: ensemble methods, particularly Random Forest and Gradient Boosting variants, consistently outperformed logistic regression in terms of AUC-ROC, with statistically significant improvements ranging from approximately 2 to 7 percentage points. The authors further demonstrated that the performance differences were not artifacts of particular datasets or evaluation protocols but were robust across multiple evaluation settings. This study established what has become a widely accepted empirical conclusion in the

field: for predictive accuracy in credit scoring, ensemble ML methods are superior to logistic regression.

It is important, however, to contextualise the magnitude of these improvements. In credit scoring, even small increases in AUC-ROC translate into substantial economic value. A 2-percentage-point improvement in AUC from 0.78 to 0.80, for example, can mean thousands fewer misclassified borrowers across a large loan portfolio, with corresponding reductions in both false approvals (loans that will default) and false rejections (creditworthy applicants who are denied). Hand (2009), while cautioning against over-reliance on AUC as a sole evaluation metric, acknowledged that in the context of credit scoring, where the costs of misclassification are asymmetric and context-dependent, improvements in discriminatory power are economically meaningful.

2.2.2 Ensemble Methods: Random Forest, XGBoost, and LightGBM

Among the ML algorithms evaluated in the credit scoring literature, three ensemble methods have emerged as consistently top-performing and practically deployable: Random Forest, XGBoost, and LightGBM. Each represents a distinct approach to combining multiple weak learners into a strong ensemble, and understanding their mechanisms is essential for the methodological choices made in this thesis.

Random Forest, introduced by Breiman (2001), constructs an ensemble of decision trees, each trained on a bootstrap sample of the training data and considering a random subset of features at each split. The final prediction is obtained by aggregating the predictions of all trees through majority voting (for classification) or averaging (for regression). Two properties make Random Forest particularly well-suited to credit scoring. First, the randomisation in both data sampling and feature selection makes the ensemble robust to overfitting, each individual tree may overfit to its particular data subset, but the diversity among trees ensures that idiosyncratic errors cancel out in the aggregate. Second, Random Forest naturally handles high-dimensional feature spaces without requiring dimensionality reduction, which is valuable when the analyst wishes to explore a wide range of candidate features without committing to a specific feature selection procedure in advance. In the credit scoring literature, Random Forest has demonstrated consistently strong

performance across multiple datasets and evaluation frameworks (Lessmann et al., 2015; Bao et al., 2019), typically achieving AUC-ROC values in the range of 0.75–0.85 depending on the dataset characteristics.

XGBoost (Extreme Gradient Boosting), developed by Chen and Guestrin (2016), represents an advancement of the gradient boosting paradigm. Unlike Random Forest, which trains trees independently and aggregates their outputs, gradient boosting trains trees sequentially: each new tree is fitted to the residual errors of the ensemble constructed so far, gradually reducing the overall prediction error. XGBoost introduces several key innovations over earlier gradient boosting implementations: L1 and L2 regularisation terms in the objective function to penalise model complexity and prevent overfitting; efficient handling of sparse data through a sparsity-aware algorithm; and parallel tree construction through a novel column-block data structure that dramatically reduces training time. Since its introduction, XGBoost has achieved state-of-the-art results across a wide range of structured data problems, including numerous credit scoring benchmarks and competitive machine learning challenges. In the context of credit scoring, XGBoost's regularisation capabilities are particularly valuable: credit datasets often contain noisy features, and regularisation helps the model focus on genuine predictive signals rather than overfitting to noise.

LightGBM (Light Gradient Boosting Machine), developed by Ke et al. (2017) at Microsoft Research, builds upon the gradient boosting framework while introducing two algorithmic innovations that specifically address the computational challenges of large-scale datasets. The first innovation, Gradient-based One-Side Sampling (GOSS), reduces the number of data instances used for computing information gain at each split by retaining all instances with large gradients (which contribute most to the information gain) while randomly sampling from instances with small gradients. The second innovation, Exclusive Feature Bundling (EFB), reduces the number of features by bundling mutually exclusive features, features that rarely take non-zero values simultaneously into single features, effectively reducing the dimensionality without losing information. These innovations collectively reduce training time and memory consumption by an order of magnitude compared to XGBoost on large datasets, while maintaining or slightly improving predictive accuracy. For the present study, which involves the

Lending Club dataset with over 2.26 million records, LightGBM's computational efficiency is a significant practical advantage.

Comparative studies generally find that XGBoost and LightGBM achieve similar predictive performance on credit scoring tasks, with LightGBM holding an advantage in training speed and memory efficiency and XGBoost sometimes achieving marginally better accuracy when computational resources are not constrained (Ke et al., 2017). Both consistently outperform logistic regression and Random Forest, though the margin over Random Forest is typically smaller than the margin over logistic regression. This hierarchy with gradient boosting methods at the top, Random Forest close behind, and logistic regression as the interpretable baseline motivates the model selection strategy adopted in this thesis, which includes all four algorithms to provide a comprehensive comparative evaluation.

2.2.3 The Role of Alternative Data in Expanding Credit Access

A particularly consequential development in ML-based credit scoring has been the incorporation of alternative data, information not traditionally included in credit bureau reports to assess the creditworthiness of populations for whom bureau data is sparse or absent. The theoretical rationale is grounded in the information asymmetry framework established in Section 2.1.3: if financial exclusion results from lenders' inability to observe creditworthiness due to missing bureau data, then any data source that contains predictive signals about repayment capacity can, in principle, serve as a substitute or complement.

Berg et al. (2020), in a study published in The Review of Financial Studies, one of the top five finance journals globally, provided rigorous evidence on this proposition. Using data from a German e-commerce company, the authors examined whether "digital footprints" information generated through individuals' online activities, including the type of device used to access the platform, the email provider, the time of day of the application, and whether the applicant used uppercase or lowercase letters in their name entry, contain predictive power for credit default. Their findings were striking: a credit scoring model based solely on digital footprints achieved an AUC-ROC of approximately 0.69, comparable to the AUC of 0.71 achieved by the traditional credit bureau score. Crucially, the predictive power of digital footprints was strongest precisely

for the borrowers for whom the bureau score was weakest, young borrowers, borrowers with short credit histories, and borrowers with few prior credit relationships. This finding suggests that alternative data and traditional data are complementary rather than substitutive: they capture different dimensions of creditworthiness, and their combination yields better predictions than either source alone.

Agarwal et al. (2020) extended this line of research to a developing-economy context, examining data from a large FinTech lender in India that uses mobile phone usage data to score applicants who lack traditional credit bureau records. The alternative data features included the number and types of applications installed on the applicant's smartphone, the diversity of the contact list, call and SMS frequency patterns, and device characteristics. Using Random Forest and XGBoost, Agarwal et al. demonstrated that these alternative data features could substitute for traditional credit scores in predicting default, enabling the lender to extend credit to individuals who would have been automatically rejected under a bureau-dependent model. Importantly, the default rates among these newly included borrowers were not significantly higher than among bureau-scored borrowers, suggesting that the alternative scoring model was identifying genuinely creditworthy individuals who had been excluded by information constraints rather than by risk.

Gambacorta et al. (2019), working with data from a major Chinese FinTech firm in a Bank for International Settlements working paper, provided a further dimension to this evidence. They compared three types of credit scoring models: a traditional model using only bank-sourced data (credit history, assets, liabilities), a FinTech model using only alternative data (e-commerce transaction patterns, social media activity, payment of non-financial obligations), and a combined model using both data sources. The FinTech model significantly outperformed the traditional model in high-risk environments, precisely the environments in which underbanked populations are concentrated, and the combined model outperformed both individual models across all risk segments. This finding has important implications for the present thesis: it suggests that the optimal credit scoring framework is not one that replaces traditional data with alternative data, but one that integrates both sources to maximise both accuracy and inclusion.

The datasets used in the present thesis, the Lending Club dataset and the Home Credit Default Risk dataset, reflect this evolving landscape. While neither dataset contains "true" alternative data such as mobile phone records or social media profiles (which are rarely available in public datasets for privacy reasons), both contain rich feature spaces that include traditional bureau-type variables (FICO score range, delinquency records), financial behaviour variables (debt-to-income ratio, revolving credit utilisation), and contextual variables (loan purpose, employment length, home ownership status) that serve as proxies for the types of alternative signals that FinTech lenders utilise. The feature engineering strategy detailed in the Methodology chapter is designed to maximise the extraction of behavioural predictive signals from these features, enabling the analysis to demonstrate how ML methods can exploit richer information sets to improve credit assessment.

2.2.4 Methodological Challenges: Class Imbalance, Overfitting, and Model Evaluation

The application of ML to credit scoring, despite its demonstrated performance advantages, introduces several methodological challenges that must be rigorously addressed to ensure reliable and generalisable results. These challenges are not unique to credit scoring, but their specific manifestations in the lending context have important implications for the design of this study.

Class imbalance is perhaps the most pervasive challenge. In typical credit scoring datasets, the proportion of defaulting borrowers is substantially smaller than the proportion of non-defaulting borrowers, often in the range of 5–15% of the total sample, depending on the lending context and the definition of default. This imbalance biases standard classifiers toward the majority class: a naïve model that classifies every applicant as "non-default" achieves 85–95% accuracy while failing to identify any of the defaulters that are the primary objects of interest. The Synthetic Minority Oversampling Technique (SMOTE), introduced by Chawla et al. (2002), has become one of the most widely used approaches to addressing this problem. SMOTE generates synthetic minority-class examples by interpolating between existing minority instances in feature space, thereby increasing the representation of the minority class without simply duplicating existing observations. However, SMOTE is not without limitations: it can generate synthetic examples in

ambiguous regions of the feature space where the boundaries between classes are unclear, potentially introducing noise. Alternative approaches include random undersampling of the majority class, which sacrifices data but avoids synthetic generation, and cost-sensitive learning through class weighting, which modifies the algorithm's loss function to assign higher costs to misclassifications of the minority class. The present study employs SMOTE as the primary resampling strategy, with comparisons to class weighting to assess robustness.

Overfitting, the tendency of complex models to learn patterns specific to the training data that do not generalise to unseen data, is a further concern, particularly with the high-dimensional feature spaces and large model capacities characteristic of ensemble methods. Stratified k-fold cross-validation, in which the dataset is partitioned into k folds with preserved class proportions and the model is trained on $k-1$ folds and evaluated on the remaining fold k times, is the standard approach to obtaining unbiased performance estimates and detecting overfitting (variance across folds signals instability). Regularisation of L1 and L2 penalties in XGBoost, maximum depth constraints in Random Forest, and early stopping in gradient boosting methods, provides a complementary defence by penalising model complexity during training. The evaluation protocol for this study uses stratified 5-fold cross-validation with an 80/20 train-test split, consistent with best practices in the credit scoring literature (Lessmann et al., 2015).

The selection of appropriate evaluation metrics is also critical and has been the subject of scholarly debate. AUC-ROC is widely considered the most informative single metric for credit scoring, as it evaluates the model's ability to discriminate between defaulters and non-defaulters across all possible classification thresholds rather than at a single, arbitrary cutoff. Hand (2009), however, raised important concerns about the AUC's implicit assumption that the ratio of the costs of false positives and false negatives is uniformly distributed, arguing that in practice, this ratio is context-dependent and rarely uniform. To address this concern, the present study supplements AUC-ROC with multiple additional metrics: F1-score (the harmonic mean of precision and recall, which balances the trade-off between identifying defaulters and avoiding false alarms), precision, recall, and Matthews Correlation Coefficient (MCC) the latter being particularly robust for imbalanced datasets. Statistical significance of performance differences between models is assessed using McNemar's test, a non-parametric test for paired binary

classification outcomes that is more appropriate than parametric alternatives when comparing classifiers on the same test set.

2.3 Explainable AI in Financial Decision-Making

2.3.1 The Black-Box Problem in High-Stakes Financial Decisions

The adoption of ensemble ML models in credit scoring creates a fundamental tension between predictive performance and interpretability that is widely recognised in both the academic literature and the regulatory community. Logistic regression, as discussed in Section 2.1.2, offers a decision logic that is transparent by construction: each coefficient maps directly to a feature's influence on the predicted probability. Ensemble methods derive their superior performance precisely from their ability to model complex, non-linear, multi-way interactions among features, interactions that cannot be captured by a linear model and that, by the same token, cannot be summarised by a simple list of coefficients (Arrieta et al., 2020). A Random Forest ensemble of 500 trees, each with depths of 10–20 levels, encodes millions of decision pathways; an XGBoost model with 1,000 sequential boosting rounds compounds this complexity further. The result is a model whose aggregate prediction is highly accurate but whose internal logic is opaque to human inspection.

In financial contexts, this opacity has consequences that go beyond academic concern. First, there are legal consequences. As discussed in Section 1.1, the EU AI Act requires that high-risk AI systems, including credit scoring systems, provide deployers with sufficient information to interpret outputs and provide affected persons with meaningful explanations. The GDPR's Article 22, combined with Recital 71, establishes a right to meaningful information about the logic of automated decisions. In the United States, the ECOA's adverse action notice requirements demand that denied applicants receive specific, understandable reasons for denial. A black-box model that produces a risk score without an accompanying explanation cannot satisfy these requirements.

Second, there are operational consequences. Internal model risk management teams within financial institutions need to understand how a model arrives at its predictions in order to validate

the model, detect degradation, and intervene when the model produces anomalous outputs. A model whose logic is inaccessible cannot be effectively governed. Third, there are consequences for adoption and trust. Loan officers and credit managers who do not understand why a model recommends a particular decision are less likely to follow its recommendations, reducing the practical value of the model's predictive accuracy.

Rudin (2019), in a widely cited commentary published in Nature Machine Intelligence, argued that for high-stakes decisions, explicitly including credit scoring, criminal justice, and healthcare, the research community should prioritise inherently interpretable models (such as sparse logistic regression, rule-based systems, or scorecards) over post-hoc explanations of black-box models. Rudin's argument rests on the premise that post-hoc explanations are inherently unfaithful: they approximate the behaviour of the complex model with a simpler surrogate, and any discrepancy between the surrogate and the original model introduces a risk of misleading the user. This position, while intellectually rigorous, faces a practical counterargument: the performance gap between interpretable models and black-box models is, in many credit scoring applications, economically significant. The present thesis adopts a pragmatic middle ground: it deploys the best-performing ensemble models and applies SHAP, the XAI method with the strongest theoretical guarantees, to provide explanations that are as faithful as possible to the underlying model, while acknowledging and measuring the inherent approximation involved.

2.3.2 Taxonomy of Explainability Methods

The field of Explainable AI has developed a rich taxonomy of methods to address the interpretability challenge. Arrieta et al. (2020), in their comprehensive survey, which has been cited over 10,000 times since its publication, organise XAI methods along several dimensions that are relevant to the present study. The first dimension is model-specific versus model-agnostic. Model-specific methods are designed for particular algorithm families: for example, feature importance derived from the information gain of splits in a Random Forest is a model-specific explainability measure. Model-agnostic methods, by contrast, can be applied to any classifier regardless of its internal architecture, treating the model as a black box and probing its behaviour through systematic input perturbation or output analysis.

The second dimension is global versus local. Global explainability methods characterise the overall behaviour of the model across the entire dataset-for example, identifying which features are, on average, most influential in driving predictions. Local explainability methods explain individual predictions-for example, explaining why a specific applicant received a particular risk score. For credit scoring, both levels are important: global explanations help model developers and risk managers understand the model's general decision logic, while local explanations are necessary to satisfy the regulatory requirement for individual-level decision justification.

The third dimension is ante-hoc versus post-hoc. Ante-hoc methods build interpretability into the model architecture itself (e.g., decision trees, rule-based systems, generalised additive models). Post-hoc methods explain the behaviour of an already-trained model without modifying its structure. For the present thesis, which seeks to combine high predictive accuracy (requiring complex ensemble models) with transparent explanations, post-hoc model-agnostic methods that provide both global and local explanations are the most appropriate choice.

2.3.3 SHAP: Theory, Properties, and Application to Credit Scoring

SHAP (SHapley Additive exPlanations), introduced by Lundberg and Lee (2017) in a paper presented at the Conference on Neural Information Processing Systems (NeurIPS) one of the premier venues in machine learning research, is grounded in the cooperative game theory concept of Shapley values, originally developed by Shapley (1953) and awarded the Nobel Memorial Prize in Economics in 2012 (jointly with Alvin Roth). In the context of cooperative game theory, a Shapley value quantifies the contribution of each player to the total payoff of a coalition game, distributing the total payoff among players in a way that satisfies four desirable axioms: efficiency (the contributions sum to the total payoff), symmetry (players who contribute identically receive identical attributions), null player (a player who contributes nothing receives zero attribution), and additivity (the value function of a combined game equals the sum of value functions of its component games).

Lundberg and Lee's innovation was to reformulate ML model interpretation as a cooperative game in which the "players" are input features and the "payoff" is the model's prediction for a specific instance. The SHAP value of each feature quantifies its contribution to the difference

between the model's prediction for that instance and the average prediction across the dataset (the "baseline" or "expected value"). Three properties make SHAP particularly suitable for credit scoring explanations. First, local accuracy: the SHAP values for a given prediction sum exactly to the difference between the model's prediction and the baseline, meaning that the explanation fully accounts for the model's output without residual. Second, missingness: features that are not present in the model's input receive SHAP values of exactly zero. Third, consistency: if a feature's contribution to the model's output increases in a modified model, its SHAP value does not decrease. These properties, which are uniquely satisfied by Shapley values among all possible additive attribution methods (Lundberg & Lee, 2017), provide a theoretically rigorous foundation for credit decision explanations, an important advantage when these explanations may be subject to regulatory scrutiny.

SHAP provides a suite of visualisation tools that address different stakeholder needs. Summary bar plots rank features by their mean absolute SHAP values across all instances, providing a global importance ranking suitable for model developers and risk managers. Beeswarm plots display the SHAP value of every instance for every feature, revealing not just which features matter but how their values map to positive or negative contributions, capturing non-linear relationships and threshold effects that are invisible in linear coefficient tables. Dependence plots show the relationship between a specific feature's values and its SHAP contributions, optionally coloured by an interacting feature to reveal interaction effects. Waterfall plots and force plots decompose individual predictions, showing exactly which features pushed a specific applicant's risk assessment above or below the baseline. For credit scoring, this local-level capability is particularly valuable: it enables the generation of human-readable decision narratives. For example, a waterfall plot might show that a specific applicant was classified as high-risk primarily because their revolving credit utilisation exceeded 85% (SHAP contribution: +0.15), they had three delinquencies in the past 24 months (+0.09), and their debt-to-income ratio was above 40% (+0.07), partially offset by their employment tenure of 10+ years (−0.06) and absence of public records (−0.04). Such a narrative is directly responsive to the EU AI Act's requirement for "clear and meaningful explanations" of AI-assisted credit decisions.

LIME (Local Interpretable Model-Agnostic Explanations), proposed by Ribeiro et al. (2016), takes a different approach: it constructs a local surrogate model, typically a weighted linear regression that approximates the complex model's behaviour in the neighbourhood of a specific prediction. While LIME is computationally simpler than SHAP and does not require the exponential computation of all possible feature coalitions, it lacks the theoretical guarantees that SHAP inherits from Shapley value theory. Specifically, LIME explanations can be inconsistent across similar instances (the same feature may appear important for one instance and unimportant for a nearly identical instance), and the choice of neighbourhood definition and sampling procedure introduces subjective parameters that affect the explanation. Lundberg and Lee (2017) demonstrated that SHAP subsumes LIME as a special case and provides stronger guarantees, a finding that has been corroborated by subsequent comparative studies (Molnar, 2020). For these reasons, the present thesis adopts SHAP as the primary explainability method.

2.3.4 The Regulatory Imperative: EU AI Act and GDPR

The regulatory context for XAI in credit scoring has been fundamentally transformed by the European Union's Artificial Intelligence Act (Regulation (EU) 2024/1689). Following years of legislative development from the European Commission's original proposal in April 2021, through extensive negotiations in the European Parliament and Council, to political agreement in December 2023 and formal adoption in 2024 the AI Act establishes a comprehensive, risk-based regulatory framework for AI systems deployed within the EU. The Act classifies AI systems into four risk categories: unacceptable risk (banned), high risk (subject to mandatory requirements), limited risk (subject to transparency obligations), and minimal risk (no specific obligations).

Credit scoring and creditworthiness assessment are explicitly listed as high-risk AI applications under Annex III, Section 5(b) of the Act. This classification subjects credit scoring AI systems to a comprehensive set of mandatory requirements. Article 9 requires the establishment and maintenance of a risk management system that identifies and analyses foreseeable risks. Article 10 imposes data governance requirements, including that training data be relevant, sufficiently representative, and free of errors to the extent possible. Article 13 the transparency provision mandates that high-risk AI systems "shall be designed and developed in such a way as to ensure

that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately." Article 14 requires human oversight mechanisms. Article 15 requires accuracy, robustness, and cybersecurity. Article 86 establishes an individual right to explanation, stipulating that affected persons have the right to obtain "clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken."

The GDPR, which has been in force since May 2018, complements the AI Act through its provisions on automated decision-making. Article 22 establishes that data subjects have the right not to be subject to decisions based solely on automated processing, including profiling, that produce legal or similarly significant effects. Recital 71 specifies that the data controller should implement suitable safeguards, which should include specific information to the data subject and the right to obtain an explanation of the decision reached after such assessment. While the precise scope of this "right to explanation" remains debated in the legal literature, there is broad consensus that it creates, at minimum, a requirement for lenders using AI systems to be able to explain the basis of their credit decisions in meaningful terms.

These regulatory provisions create a concrete and immediate demand for the kind of explainability that SHAP provides. A credit scoring model that can decompose each decision into interpretable feature contributions demonstrating that a rejection was driven by specific, identifiable, and legitimate risk factors is substantially better positioned to comply with the EU AI Act's transparency requirements than an opaque model accompanied only by aggregate performance statistics. This regulatory alignment provides both the practical motivation and the timeliness for the present study.

2.4 Algorithmic Fairness in Credit Decisions

2.4.1 Defining Fairness: Competing Frameworks and Impossibility Results

Algorithmic fairness has become a critical concern in ML-based credit scoring, driven by accumulating evidence that automated decision systems can perpetuate or amplify historical biases embedded in their training data. However, defining what constitutes "fairness" in an

algorithmic context is a substantially more complex task than it may initially appear. The literature identifies multiple, formally defined fairness criteria, several of which are mutually incompatible, a finding with profound implications for both research and practice (Mehrabi et al., 2021).

Three fairness criteria are particularly prominent in the credit scoring context. Demographic parity (also termed statistical parity or group fairness) requires that the proportion of positive outcomes (e.g., loan approvals) be equal across demographic groups, regardless of the base rates of creditworthiness within those groups. Formally, if $\hat{Y}$ denotes the model's prediction and A denotes group membership, demographic parity requires $P(\hat{Y} = 1 \mid A = a) = P(\hat{Y} = 1 \mid A = b)$ for all groups a and b. The appeal of this criterion is its simplicity and its direct correspondence to the legal concept of disparate impact: a lending practice that produces substantially different approval rates across protected groups may be challenged as discriminatory under US fair lending law and EU anti-discrimination directives.

Equalised odds, proposed by Hardt et al. (2016), imposes a more nuanced requirement: the true positive rate (probability of correctly approving a creditworthy applicant) and the false positive rate (probability of incorrectly approving a non-creditworthy applicant) must be equal across groups. Formally, equalised odds requires $P(\hat{Y} = 1 \mid Y = y, A = a) = P(\hat{Y} = 1 \mid Y = y, A = b)$ for $y \in \{0, 1\}$. This criterion allows for different approval rates across groups, reflecting different base rates of creditworthiness, but demands that the model's accuracy be equitable: it should be equally good at identifying creditworthy applicants and equally likely to make errors across all groups.

Predictive parity (also termed calibration) requires that the positive predictive value, the proportion of approved applicants who actually repay, be equal across groups. This criterion focuses on the reliability of the model's positive predictions: a model satisfies predictive parity if, among applicants who receive the same risk score, the actual default rate is the same regardless of group membership.

Chouldechova (2017) proved a foundational impossibility result that has shaped all subsequent research on algorithmic fairness: when the base rates of the target outcome (e.g., default rates)

differ across groups, as they typically do in credit scoring, reflecting structural socioeconomic inequalities, it is mathematically impossible to simultaneously satisfy demographic parity, equalised odds, and predictive parity. This impossibility theorem implies that practitioners must make explicit, deliberate choices about which dimension of fairness to prioritise. These choices are not purely technical; they involve value judgements about what constitutes equitable treatment in the context of lending, and they may have different implications under different legal frameworks. The present thesis evaluates multiple fairness criteria simultaneously, transparently reporting the trade-offs rather than optimising for a single criterion.

2.4.2 Sources of Bias in Credit Scoring Models

Understanding the mechanisms through which bias enters ML-based credit scoring models is essential for designing effective mitigation strategies. Barocas and Selbst (2016), in a seminal analysis published in the California Law Review, identified multiple pathways through which automated decision systems can produce discriminatory outcomes, even when the model designer has no discriminatory intent and even when protected characteristics (race, gender, ethnicity) are not explicitly included as input features.

Historical bias is perhaps the most insidious source. If the training data reflects past discriminatory lending practices, for example, if minority applicants were historically denied credit at higher rates due to prejudice rather than risk assessment, then a model trained on this data will learn to replicate these patterns, treating historical discrimination as a predictive signal. The model does not "know" that the denials were discriminatory; it simply observes that applicants with certain characteristics were denied and incorporates this pattern into its predictions. Representation bias arises when certain groups are underrepresented in the training data, leading to less accurate predictions for those groups. If a lending platform historically served primarily urban, employed, middle-class borrowers, its model will be better calibrated for that population than for rural, self-employed, or lower-income applicants, precisely the populations that financial inclusion efforts aim to reach.

Measurement bias occurs when the features used as model inputs are themselves correlated with protected characteristics, even when those characteristics are not directly included. Geographic

variables (zip code, state) are well-documented proxies for race and ethnicity in the United States due to persistent residential segregation. Employment type and industry can proxy for gender due to occupational segregation. Even seemingly neutral features like the channel through which a loan application was submitted (online versus in-branch) can correlate with age and technology access. This phenomenon of "proxy discrimination" means that simply removing protected characteristics from the model's input features, the approach known as "fairness through unawareness" is insufficient to prevent discriminatory outcomes.

Empirical evidence confirms that these theoretical concerns are not hypothetical. Bartlett et al. (2022), published in the Journal of Financial Economics, found evidence of discrimination in both face-to-face and algorithmic lending in the US mortgage market, with minority borrowers paying interest rates approximately 7.9 basis points higher than comparable non-minority borrowers after controlling for creditworthiness. Fuster et al. (2022), published in The Journal of Finance, demonstrated that while ML-based credit scoring models reduce average prediction errors relative to traditional models (confirming the performance advantages documented in Section 2.2), they can simultaneously increase the disparity in error rates across racial groups. Specifically, ML models improved accuracy most for the majority group while providing smaller improvements, or in some cases, no improvement for minority groups. This finding reveals a fairness-accuracy trade-off that requires explicit management: improving overall model performance does not automatically improve fairness, and in some cases, may worsen it.

2.4.3 Bias Mitigation Strategies

The literature proposes three categories of bias mitigation strategies, classified by the stage in the ML pipeline at which they intervene (Mehrabi et al., 2021). Pre-processing techniques modify the training data before model training to remove or reduce bias. These include resampling to equalise group representation, feature transformation to decorrelate features from protected attributes, and relabelling strategies that correct historical labelling biases. Pre-processing methods have the advantage of being model-agnostic, they modify the data, not the algorithm but they can distort the data distribution in ways that reduce overall predictive accuracy, and they require access to protected attribute information in the training data.

In-processing techniques modify the learning algorithm itself to incorporate fairness constraints during training. Examples include adding a fairness regularisation term to the loss function (penalising the model for producing different error rates across groups), adversarial debiasing (training a secondary "adversary" model to detect whether predictions are correlated with protected attributes, and using this feedback to remove such correlations from the primary model), and constrained optimisation approaches that directly optimise for fairness metrics alongside accuracy. In-processing methods can be theoretically elegant but require modifications to standard training algorithms that may be complex to implement and may not be compatible with all model architectures.

Post-processing techniques adjust the model's output after training, without modifying the model itself. The most common approach is threshold calibration: rather than applying a single decision threshold to all applicants, group-specific thresholds are set to equalise a chosen fairness metric across demographic groups. For example, if the model's false positive rate is higher for Group A than Group B at a threshold of 0.5, the threshold for Group A could be raised slightly (reducing its false positive rate) while the threshold for Group B could be lowered (increasing its true positive rate), until the rates equalise. Post-processing methods are the most straightforward to implement, do not require retraining the model, and are compatible with any model architecture. Their primary limitation is that they do not address the root causes of bias in the model's learned representations; they merely adjust the mapping from predictions to decisions.

For practical deployment in FinTech credit scoring where model retraining may be costly and frequent, where regulatory compliance timelines are tight, and where the model architecture may not be easily modifiable, post-processing approaches, particularly threshold calibration, offer the most pragmatic balance between fairness improvement and implementation feasibility. The present thesis adopts this approach as its primary bias mitigation strategy, while documenting the fairness-accuracy trade-off that threshold adjustment entails.

2.5 Conceptual Framework and Research Gap

2.5.1 Synthesis: What the Literature Establishes

The preceding four sections have reviewed interconnected bodies of literature that, taken together, delineate both the opportunity and the challenge that this thesis addresses. From credit risk theory (Section 2.1), the review established that information asymmetry, the inability of lenders to observe the creditworthiness of thin-file and no-file borrowers, is the fundamental mechanism driving financial exclusion, and that traditional logistic regression scoring is structurally dependent on bureau data that these populations lack. From the ML in credit scoring literature (Section 2.2), the review demonstrated that ensemble methods (Random Forest, XGBoost, LightGBM) significantly and consistently outperform logistic regression in predictive accuracy, and that alternative data sources can meaningfully expand credit access to populations previously deemed unscorable. From the XAI literature (Section 2.3), the review showed that SHAP provides a theoretically rigorous, practically implementable method for decomposing ML predictions into interpretable feature contributions, a capability now demanded by the EU AI Act for all high-risk AI systems, including credit scoring. From the algorithmic fairness literature (Section 2.4), the review highlighted that ML models can inherit and amplify historical biases, that multiple competing definitions of fairness exist with inherent mathematical trade-offs, and that post-processing mitigation strategies offer a pragmatic approach to reducing disparities.

2.5.2 Identification of the Research Gap

Despite these individual advances, the literature reveals a critical gap at the intersection of these three domains. The fragmentation is systematic and consequential:

Credit scoring studies that compare ML algorithms (Lessmann et al., 2015; Bao et al., 2019) optimize for predictive accuracy without incorporating explainability or fairness. Their contribution is to establish that ML outperforms logistic regression, but they do not address whether the superior models can be made transparent or whether their predictions are equitable two questions that are essential for regulatory compliance and ethical deployment.

XAI studies applied to finance (Arrieta et al., 2020; Lundberg & Lee, 2017) demonstrate that explanation methods exist and can be applied to financial models, but they typically operate on small or illustrative datasets, do not integrate fairness evaluation, and do not propose end-to-end frameworks that a financial institution could deploy operationally.

Fairness studies in credit scoring (Chouldechova, 2017; Hardt et al., 2016; Fuster et al., 2022) establish important theoretical results and empirical findings about bias, but they address detection and mitigation in isolation from the practical requirements of model explainability, regulatory compliance, and stakeholder communication.

No existing study or framework integrates all three pillars, high-performance ML prediction, SHAP-based regulatory-grade explainability, and fairness-aware evaluation into a unified, empirically validated decision-support framework for credit scoring, validated on real-world lending data at scale. This is the gap that the present thesis addresses.

2.5.3 The Proposed Conceptual Framework: The EICS Model

To address this gap, the thesis proposes an Explainable Inclusive Credit Scoring (EICS) Framework structured around three interconnected and sequentially dependent pillars:

Pillar 1 – Prediction. High-performance ML models (Logistic Regression as the interpretable baseline; Random Forest, XGBoost, and LightGBM as ensemble candidates) are trained on real-world lending data and rigorously compared using multiple evaluation metrics (AUC-ROC, F1-score, precision, recall, MCC) with statistical significance testing (McNemar's test). Class imbalance is addressed through SMOTE and class weighting. The output of this pillar is the identification of the best-performing model and a quantified understanding of how much predictive improvement ML delivers over the logistic regression baseline.

Pillar 2 – Explanation. The best-performing model is subjected to comprehensive SHAP analysis at both global and local levels. Global SHAP analysis (summary bar plots, beeswarm plots) reveals the features that are, on average, most influential in driving predictions across the entire portfolio. Local SHAP analysis (waterfall plots, force plots) decomposes individual predictions into feature-level contributions, enabling the generation of human-readable decision

narratives. Feature interaction analysis (SHAP dependence plots) captures non-linear and interactive effects. The output of this pillar is a transparent mapping from model predictions to interpretable explanations that satisfy the EU AI Act's requirements for meaningful information about AI-assisted decisions.

Pillar 3 – Fairness. The model's predictions are evaluated for demographic fairness using established metrics: Demographic Parity Ratio, Equalised Odds Difference, and Disparate Impact Ratio. Where disparities are detected, post-processing mitigation strategies specifically, threshold calibration per demographic subgroup are applied. The impact of mitigation on both fairness metrics and predictive accuracy is measured, providing an empirical quantification of the fairness-accuracy trade-off. The output of this pillar is a fairness-audited model with documented trade-off analysis that supports responsible deployment decisions.

The three pillars are designed to be sequentially dependent but iteratively refinable. The Prediction pillar identifies the best model; the Explanation pillar makes its logic transparent; the Fairness pillar audits its equity. Together, they constitute a complete decision-support workflow that addresses the requirements of predictive accuracy (demanded by business stakeholders), transparency (demanded by regulators and affected individuals), and equity (demanded by ethical principles and anti-discrimination law). The EICS Framework is the thesis's primary original contribution and is designed to be adaptable to different lending contexts, regulatory environments, and data availability scenarios.

The empirical validation of this framework on real-world lending data comprising over 2.5 million records with a robustness check on a second, independent dataset distinguishes this thesis from conceptual proposals that lack empirical grounding. The following chapter details the methodology through which this validation is conducted.

Chapter 3: Methodology

This chapter presents the research methodology through which the Explainable Inclusive Credit Scoring (EICS) Framework is empirically implemented and validated. The chapter is organised into nine sections that collectively describe and justify every methodological decision, from the philosophical foundations of the research design to the specific computational tools employed. Section 3.1 establishes the research philosophy and approach. Section 3.2 describes the research design. Section 3.3 details the data sources and their justification. Section 3.4 presents the complete data preprocessing pipeline. Section 3.5 describes the model selection, training, and evaluation protocol. Section 3.6 outlines the SHAP-based explainability analysis. Section 3.7 details the fairness evaluation methodology. Section 3.8 addresses ethical considerations. Section 3.9 justifies the selection of software tools. Each methodological choice is explicitly linked to the research questions formulated in Chapter 1 and justified with reference to the academic literature reviewed in Chapter 2.

3.1 Research Philosophy and Approach

The philosophical foundation of a research study determines the assumptions about knowledge, reality, and the relationship between the researcher and the phenomenon under investigation. These assumptions shape the methodological choices that follow. The present study is grounded in a positivist epistemology, which holds that objective knowledge about the world can be obtained through systematic observation, measurement, and the testing of falsifiable hypotheses (Saunders et al., 2019). Positivism is the dominant philosophical paradigm in quantitative financial research, where the objects of inquiry—credit default events, predictive model performance, feature importance rankings, fairness metrics—are empirically observable and measurable phenomena that exist independently of the researcher's interpretation.

The ontological position is realist: credit default is an objective event (a borrower either repays or does not), model performance metrics are deterministic functions of predictions and true labels, and fairness metrics are mathematical computations over defined subgroups. There is no interpretive ambiguity in whether a loan was charged off or fully paid; the reality is recorded in

the data. This realist ontology supports the use of quantitative methods that measure, compare, and test these objective quantities.

The research approach is deductive. The study begins with theoretical propositions derived from the literature (the hypotheses H1–H4 formulated in Section 1.3), designs a systematic empirical investigation to test these propositions, collects and analyses data, and draws conclusions about whether the evidence supports or refutes the hypotheses. This deductive logic is appropriate because the research questions are specific, the variables are quantifiable, and the analytical techniques produce numerical results that can be compared against predefined benchmarks and tested for statistical significance (Creswell & Creswell, 2018).

A qualitative or interpretivist approach was considered but rejected. While interviews with FinTech practitioners or regulators about their experiences with AI-based credit scoring could yield valuable complementary insights, they would not address the core research questions, which are fundamentally quantitative: comparing model performance metrics, decomposing feature contributions through SHAP values, and measuring fairness disparities. A mixed-methods approach, while theoretically appealing, was assessed as impractical within the scope and timeline of a Master's thesis. The EDC handbook notes that mixed methods are "not necessarily recommended at this level" (EDC Paris Business School, 2025, p. 8), and the depth of the quantitative analysis required by this study is sufficient to constitute a rigorous standalone methodology.

3.2 Research Design

The research design is a quantitative, comparative-analytical, secondary data study. Each element of this design is justified below.

The design is quantitative because all four sub-questions require numerical answers: SQ1 requires performance metrics (AUC-ROC, F1-score) for model comparison; SQ2 requires SHAP values quantifying feature contributions; SQ3 requires fairness metrics measuring demographic disparities; and SQ4 requires the synthesis of these quantitative outputs into a structured

framework. There is no interpretive or exploratory component that would benefit from a qualitative design.

The design is comparative-analytical because the core analytical strategy involves the systematic comparison of four machine learning models (Logistic Regression, Random Forest, XGBoost, LightGBM) against standardised performance, explainability, and fairness criteria. This comparative structure is essential for answering SQ1 and is a well-established design in credit scoring research (Lessmann et al., 2015; Bao et al., 2019).

The design relies on secondary data, publicly available loan-level datasets containing real borrower characteristics and actual default outcomes. Secondary data is justified on multiple grounds. First, the datasets available (Lending Club: 2.26 million records; Home Credit: 307,000 records) are substantially larger and more representative than any primary dataset collectible within a Master's thesis, far exceeding the EDC minimum of 150 records for quantitative research. Second, these datasets have been extensively used in peer-reviewed research (Berg et al., 2020; Lessmann et al., 2015), lending credibility to results derived from them. Third, primary data collection in the form of real lending outcomes would require partnership with a financial institution, infeasible at this stage. The EDC handbook explicitly lists "secondary data studies" among common business research designs and states that data collection methods can include "analysing secondary data (e.g., financial reports, databases, archival materials)" (EDC Paris Business School, 2025, p. 8).

The design is cross-sectional: it analyses a snapshot of historical loan data rather than tracking outcomes longitudinally. While longitudinal analysis could capture temporal dynamics such as concept drift, the complexity of implementing longitudinal ML with explainability and fairness evaluation exceeds the scope of this thesis. The cross-sectional design allows focused demonstration of the EICS Framework, with temporal extension identified as an avenue for future research.

3.3 Data Sources and Description

3.3.1 Primary Dataset: Lending Club Loan Data (2007–2018)

The primary dataset is the Lending Club Accepted Loans dataset, publicly available on Kaggle and originally released by Lending Club, one of the largest peer-to-peer lending platforms in the United States. Lending Club operated from 2007 to 2020 as an online marketplace connecting borrowers seeking personal loans with investors willing to fund those loans, publicly disclosing loan-level data as part of its transparency and regulatory compliance practices.

The dataset comprises approximately 2.26 million individual loan records spanning 2007 to Q4 2018. Each record represents a single funded loan and contains over 150 features describing the borrower's financial profile, the loan's characteristics, and the repayment outcome. Key feature categories include the following.

Borrower financial profile: annual income, employment length, home ownership status (own, mortgage, rent), debt-to-income ratio, number of open credit lines, total credit revolving balance, revolving utilisation rate, and number of derogatory public records.

Credit bureau information: FICO score range (upper and lower bounds at origination), number of inquiries in the last six months, number of accounts with delinquency, earliest credit line date, and months since last delinquency.

Loan characteristics: loan amount, funded amount, interest rate, loan term (36 or 60 months), loan grade and sub-grade (assigned by Lending Club's internal scoring), loan purpose (debt consolidation, credit card refinancing, home improvement, major purchase, medical, etc.), and income verification status.

Repayment outcome (target variable): loan status, recording the final disposition of the loan. Loans classified as "Fully Paid" are coded as non-default (0); loans classified as "Charged Off" are coded as default (1). Loans with intermediate statuses (Current, In Grace Period, Late 16–30 Days, Late 31–120 Days) are excluded to ensure a clean binary outcome, consistent with standard practice in the credit scoring literature.

The choice of Lending Club as the primary data source is justified on four grounds. First, it contains real borrower data with actual default outcomes, not simulated or synthetic records providing ecological validity essential for a study aimed at practical relevance. Second, its scale (2.26 million records) far exceeds the EDC minimum and provides sufficient statistical power for reliable model comparison, SHAP analysis, and fairness evaluation across subgroups. Third, its feature richness (150+ variables) enables both traditional bureau-style scoring and the engineering of behavioural proxy features approximating alternative data signals used by FinTech lenders. Fourth, the dataset has been extensively used in peer-reviewed credit scoring literature, facilitating comparison with existing findings.

3.3.2 Supplementary Dataset: Home Credit Default Risk

The supplementary dataset is the Home Credit Default Risk dataset, sourced from the Home Credit Group, an international consumer finance provider operating across Central and Eastern Europe, South and Southeast Asia, and other emerging markets. Home Credit's stated mission is to provide loans to individuals with little or no credit history, a population directly corresponding to the "underbanked" applicational focus of this thesis.

The main application table contains approximately 307,000 loan application records with 122 features. Unlike Lending Club, the Home Credit dataset includes alternative data dimensions: previous application history at the same institution, credit card balance and payment behaviour over time, instalment payment patterns, and point-of-sale loan records. These features more closely approximate the alternative data streams that FinTech lenders use to score underbanked populations, making this dataset particularly relevant for testing whether the EICS Framework generalises beyond the US peer-to-peer lending context.

The Home Credit dataset serves specifically as a robustness check. The primary analysis is conducted on Lending Club; the key findings (model performance rankings, top SHAP features, fairness outcomes) are then replicated on Home Credit to assess generalisability. Consistent findings across both datasets strengthen the EICS Framework's external validity; divergent findings provide insight into context-dependence, discussed in Chapter 5.

3.3.3 Anticipation of Data Quality Issues

Both datasets present data quality challenges that must be anticipated and addressed before modelling. The Lending Club dataset, given its scale and organic nature, exhibits significant missing values in several features. Preliminary exploration reveals that approximately 20–30 features have missingness rates exceeding 50%, primarily in fields related to joint applications, secondary credit bureau records, and late-stage repayment indicators. These extensively missing features are dropped, as imputation at >50% missingness introduces more noise than signal.

For features with moderate missingness (5–50%), the pattern of missingness is assessed. If data is missing at random (MAR) or missing completely at random (MCAR), imputation is justified. If data is missing not at random (MNAR) for example, "months since last delinquency" is missing for borrowers who have never had a delinquency, this is documented as a limitation and handled through indicator variables that flag missingness as a separate signal, preserving information that naïve imputation would destroy (Enders, 2010).

The Home Credit dataset presents a different challenge: its multi-table relational structure. Supplementary tables (previous applications, credit card balance, instalment payments, POS loan records) contain time-series data requiring aggregation to the application level before merging. Aggregation functions (mean, median, max, min, count, sum) generate summary features, following the methodology used by top-performing Kaggle solutions and validated in subsequent academic analyses. All aggregation decisions are documented for reproducibility.

3.4 Data Preprocessing Pipeline

The data preprocessing pipeline transforms raw datasets into analysis-ready formats through a systematic, fully documented sequence of steps. Each step is justified with its rationale and specific decisions.

3.4.1 Data Cleaning and Missing Value Treatment

Features with more than 50% missing values are removed. Features that are identifiers (loan ID, member ID) or that contain post-origination information (collection recovery amounts, last

payment date) are removed to prevent data leakage, the inadvertent inclusion of future information in training data that would artificially inflate model performance.

For remaining features with missing values, the imputation strategy is as follows. Numerical features are imputed with the median, which is more robust to outliers than mean imputation. Categorical features are imputed with the mode (most frequent category). Crucially, imputation parameters are computed on the training set only and then applied to the test set, preventing information leakage from test data into imputation values. For features where missingness itself is informative (e.g., "months since last delinquency" is missing for borrowers who have never been delinquent), a binary indicator variable is created to flag missing status, allowing the model to learn from the missingness pattern.

3.4.2 Target Variable Definition

In the Lending Club dataset, loans with status "Fully Paid" are assigned target value 0 (non-default); loans with status "Charged Off" are assigned target value 1 (default). Lending Club defines a charge-off as a loan for which the borrower has failed to make payments for 150 days or more and recovery is deemed unlikely. This definition is consistent with industry practice and academic literature (Lessmann et al., 2015). Loans with intermediate statuses (Current, In Grace Period, Late) are excluded to avoid label ambiguity: their final outcome is unknown and including them would degrade both model reliability and SHAP interpretability.

In the Home Credit dataset, the target variable "TARGET" is pre-coded as binary (1 = default, 0 = non-default), requiring no further transformation.

3.4.3 Feature Engineering

Feature engineering creates new features from existing data to improve model performance and interpretability. The strategy is designed to extract maximum predictive signal while creating features that are interpretable in the context of credit risk assessment.

Financial behaviour ratios: Loan-to-income ratio (loan amount divided by annual income), revolving balance to total credit limit ratio, and instalment-to-income ratio (monthly payment

divided by monthly income). These ratios capture the borrower's financial leverage relative to their capacity, among the strongest default predictors in the literature (Thomas et al., 2002).

Credit history indicators: Years since earliest credit line (credit maturity proxy), ratio of delinquent accounts to total accounts (historical repayment reliability), inquiry-to-account ratio (credit-seeking intensity), and a binary flag for any delinquency in the past two years.

Stability proxies: Employment length recoded numerically (0–10+ years) and grouped into categories (<1, 1–3, 3–5, 5–10, 10+ years) to capture the non-linear relationship between employment tenure and default risk. Home ownership status is retained as a categorical stability indicator. These features approximate the alternative data signals (employment stability, residential permanence) used by FinTech lenders within the constraints of the available dataset.

Loan context features: Loan purpose grouped into broader categories (debt consolidation, credit card refinancing, home improvement, other) to reduce cardinality while preserving meaningful distinctions. Loan grade and sub-grade encoded numerically to preserve ordinal risk relationships.

3.4.4 Class Imbalance Treatment

Credit scoring datasets are characteristically imbalanced: the Lending Club dataset exhibits an approximate default rate of 20–25% among analysed loans (after removing intermediate-status records), representing a moderate but significant class imbalance. Failure to address this imbalance would bias classifiers toward predicting non-default for all applicants, producing misleadingly high accuracy while failing to identify defaulters.

The primary resampling strategy is SMOTE (Synthetic Minority Oversampling Technique; Chawla et al., 2002). SMOTE generates synthetic minority-class examples through interpolation between existing minority instances in feature space: for each minority instance, SMOTE identifies its k = 5 nearest neighbours, randomly selects one or more, and creates synthetic instances at random points along connecting line segments. This increases minority representation without simply duplicating existing observations, which would risk overfitting to specific instances.

SMOTE is applied exclusively to the training set. The test set remains unmodified to ensure evaluation metrics reflect performance on the natural (imbalanced) distribution, what the model encounters in deployment. This train-only application is critical for preventing optimistic bias in performance estimates (Chawla et al., 2002).

As a robustness check, the analysis is also conducted using class weighting, where the loss function assigns higher misclassification costs to minority-class instances using weights inversely proportional to class frequencies: weight = n_samples / (n_classes × n_samples_class). Comparing SMOTE and class-weighted results tests whether findings are sensitive to the specific resampling strategy.

3.4.5 Feature Encoding and Scaling

Categorical features are encoded using one-hot encoding for nominal variables (home ownership, loan purpose) and ordinal encoding for ordinal variables (loan grade). One-hot encoding is preferred for nominal features because it does not impose artificial ordinal relationships that tree-based models might exploit misleadingly.

Numerical features are standardised (zero mean, unit variance). Standardisation is necessary for Logistic Regression, which is sensitive to feature scale, and applied uniformly across all models for consistency. Standardisation parameters are computed on the training set only and applied to the test set to prevent information leakage.

3.5 Model Selection, Training, and Evaluation

3.5.1 Model Selection and Justification

Four classification models are selected, each serving a specific role within the analytical framework:

Logistic Regression (Baseline): The industry standard for credit scoring (Hand & Henley, 1997; Thomas et al., 2002). Serves as the interpretable baseline against which ensemble models are measured. If ensemble models do not significantly outperform logistic regression, the case for

deploying more complex models, with their associated explainability and governance challenges is weakened. Implemented with L2 (ridge) regularisation, with regularisation strength C optimised through cross-validation.

Random Forest: A bagging-based ensemble constructing independent decision trees on bootstrapped samples (Breiman, 2001). Selected because it consistently ranks among top performers in credit scoring benchmarks (Lessmann et al., 2015), provides built-in feature importance for comparison with SHAP, and is robust to overfitting through random feature subsetting. Key hyperparameters: n_estimators, max_depth, min_samples_leaf, max_features.

XGBoost: A gradient boosting ensemble training trees sequentially on residual errors (Chen & Guestrin, 2016). Selected because it achieves state-of-the-art performance on structured data, its L1/L2 regularisation prevents overfitting on noisy credit data, and it supports TreeSHAP for exact SHAP value computation. Key hyperparameters: n_estimators, learning_rate, max_depth, subsample, colsample_bytree, reg_alpha, reg_lambda.

LightGBM: A gradient boosting ensemble with GOSS and EFB optimisations for large-scale efficiency (Ke et al., 2017). Selected because its performance is comparable to XGBoost at substantially lower training time critical for the 2.26-million-record dataset and it supports TreeSHAP. Including both XGBoost and LightGBM enables comparison of the two dominant gradient boosting implementations. Key hyperparameters: num_leaves, learning_rate, n_estimators, feature_fraction, bagging_fraction.

Alternative models were considered and excluded. Deep neural networks were excluded because they typically do not outperform gradient boosting on structured/tabular data (Grinsztajn et al., 2022), require substantially more computational resources, and SHAP computation for deep networks is computationally expensive and less exact than TreeSHAP. Support vector machines were excluded due to poor scalability to millions of records. k-Nearest Neighbours was excluded for scalability concerns and absence of meaningful feature importance mechanisms.

3.5.2 Train-Test Split and Cross-Validation Protocol

The dataset is partitioned into training (80%) and hold-out test (20%) sets using stratified random sampling, preserving the default/non-default proportion in both partitions. The test set is isolated before any preprocessing (imputation, encoding, scaling, SMOTE) and used exclusively for final evaluation, providing an unbiased performance estimate on unseen data.

Within the training set, hyperparameter optimisation uses stratified 5-fold cross-validation. The training set is partitioned into five folds with preserved class proportions; for each hyperparameter configuration, the model is trained on four folds and evaluated on the fifth, rotating five times. The average AUC-ROC across folds selects the optimal configuration. Large variance across folds signals model instability and potential overfitting.

Hyperparameter search uses Bayesian optimisation via the Optuna framework, which builds a probabilistic model of the objective function from previous evaluations and concentrates search in promising regions, achieving better results than exhaustive grid search within the same computational budget. For each model, 100 optimisation trials are conducted.

3.5.3 Evaluation Metrics

Model performance is assessed using multiple complementary metrics, each capturing a different aspect of classification quality. No single metric is sufficient for credit scoring evaluation; the use of multiple metrics provides a comprehensive picture and prevents over-optimisation on any single criterion.

AUC-ROC (primary metric): The Area Under the Receiver Operating Characteristic Curve measures the model's ability to discriminate between defaulters and non-defaulters across all possible classification thresholds. It is threshold-independent, which is valuable because the optimal decision threshold depends on the lender's specific cost structure. AUC-ROC is the most widely used metric in credit scoring research (Lessmann et al., 2015) and serves as the primary basis for model comparison.

F1-Score: The harmonic mean of precision and recall, balancing the trade-off between correctly identifying defaulters (recall) and minimising false alarms among predicted defaulters

(precision). F1-score is particularly informative for imbalanced datasets where accuracy alone is misleading.

Precision and Recall (separately): Precision measures the proportion of predicted defaulters who actually defaulted; recall measures the proportion of actual defaulters correctly identified. Reporting both separately reveals whether a model achieves its F1-score through balanced performance or by trading off one component against the other a distinction with direct business implications (high recall minimises missed defaults; high precision minimises rejected creditworthy applicants).

Matthews Correlation Coefficient (MCC): A balanced measure that accounts for all four cells of the confusion matrix (true positives, true negatives, false positives, false negatives), producing a value between −1 and +1. MCC is considered the most informative single metric for binary classification on imbalanced datasets because it gives high scores only when the model performs well on both classes (Chicco & Jurman, 2020).

Confusion Matrix: The full 2×2 confusion matrix is reported for each model at the optimal threshold, providing the raw counts of true positives, true negatives, false positives, and false negatives that underlie all derived metrics.

3.5.4 Statistical Significance Testing

To determine whether observed performance differences between models are statistically significant rather than attributable to random variation, McNemar's test is applied to pairwise model comparisons on the hold-out test set. McNemar's test is a non-parametric test for paired binary classification outcomes that examines whether two models disagree on the same instances at significantly different rates. Specifically, it tests the null hypothesis that the two models have the same error rate by analysing the discordant cells of a 2×2 contingency table: instances where Model A is correct and Model B is wrong, and vice versa. The test statistic follows a chi-squared distribution with one degree of freedom under the null hypothesis. A significance level of $\alpha = 0.05$ is adopted throughout.

McNemar's test is preferred over alternatives (such as paired t-tests across cross-validation folds or bootstrap confidence intervals) because it directly compares the models' predictions on the same held-out instances, avoids assumptions of normality, and has been specifically recommended for comparing classifiers in the machine learning literature (Dietterich, 1998).

3.6 SHAP-Based Explainability Analysis

The explainability analysis constitutes Pillar 2 of the EICS Framework and directly addresses Sub-Question 2. Once the best-performing model is identified through the comparative evaluation in Section 3.5, it is subjected to comprehensive SHAP analysis at both global and local levels. The choice of SHAP over alternative XAI methods (notably LIME) is justified in Chapter 2, Section 2.3.3, on the grounds of SHAP's unique theoretical properties: local accuracy, missingness, and consistency—properties that are uniquely satisfied by Shapley values among all additive attribution methods (Lundberg & Lee, 2017).

3.6.1 TreeSHAP Implementation

For tree-based ensemble models (Random Forest, XGBoost, LightGBM), SHAP values are computed using TreeSHAP, an algorithm developed by Lundberg et al. (2020) that exploits the tree structure to compute exact Shapley values in polynomial time—$O(TLD^2)$ where T is the number of trees, L is the maximum number of leaves per tree, and D is the maximum depth. This is dramatically faster than the general kernel-based SHAP algorithm, which has exponential complexity in the number of features. For the Logistic Regression baseline, the linear SHAP method is used, which computes exact Shapley values in closed form for linear models.

SHAP values are computed on the hold-out test set (not the training set) to ensure that the explanations reflect the model's behaviour on unseen data, the same data on which performance metrics are evaluated. This alignment between the evaluation data and the explanation data ensures that the explanations are representative of the model's deployed behaviour rather than its behaviour on the data it was trained to fit.

3.6.2 Global Explainability

Global explainability characterises the model's overall decision logic across the entire test set. Two primary visualisations are generated. The SHAP summary bar plot ranks features by their mean absolute SHAP value across all test instances, identifying the features that are, on average, most influential in driving predictions. This provides a global importance ranking that can be compared with the model's native feature importance (e.g., information gain in XGBoost) and with domain knowledge about credit risk drivers. The SHAP beeswarm plot displays the SHAP value of every test instance for each feature, colour-coded by feature value (red = high, blue = low). This visualisation reveals not just which features matter but how their values map to positive or negative contributions capturing non-linear relationships (e.g., high revolving utilisation consistently pushes predictions toward default) and heterogeneous effects (e.g., moderate income has little effect but very low income strongly increases risk).

Additionally, SHAP dependence plots are generated for the top five features, showing the relationship between each feature's values and its SHAP contributions, coloured by the feature with the strongest interaction effect. These plots reveal interaction dynamics that are invisible in global importance rankings and that are particularly relevant for credit scoring, where feature interactions (e.g., the combined effect of high debt-to-income ratio and short employment length) may be more predictive than individual features in isolation.

3.6.3 Local Explainability and Decision Narrative Generation

Local explainability decomposes individual predictions into feature-level contributions. SHAP waterfall plots are generated for a set of representative individual cases selected to illustrate diverse decision scenarios: a clearly high-risk application (predicted default probability > 0.8), a clearly low-risk application (predicted default probability < 0.2), a borderline case (predicted probability near the decision threshold), and critically a case where the model's prediction diverges from what the borrower's observable characteristics might suggest (e.g., a high-income borrower classified as high-risk due to high revolving utilisation and recent delinquencies). This

last case is particularly important because it demonstrates SHAP's ability to identify non-obvious risk drivers, which is exactly the kind of explanation that the EU AI Act's Article 86 envisions.

For each selected case, the SHAP waterfall plot is translated into a structured human-readable decision narrative following a standardised template: "This applicant was classified as [high-risk/ low-risk] primarily because [Feature A: specific value, contribution direction and magnitude], [Feature B: specific value, contribution direction and magnitude], and [Feature C], partially offset by [Feature D] and [Feature E]." This narrative template represents a methodological contribution of the thesis: it demonstrates concretely how SHAP outputs can be operationalised into the "clear and meaningful explanations" required by the EU AI Act, bridging the gap between technical XAI output and regulatory compliance language.

3.7 Fairness Evaluation Methodology

The fairness evaluation constitutes Pillar 3 of the EICS Framework and directly addresses Sub-Question 3. It assesses whether the model's predictions produce systematically different outcomes across demographic subgroups and, where disparities are detected, applies mitigation strategies.

3.7.1 Subgroup Definition and Proxy Variables

A critical methodological challenge in the fairness evaluation is that the Lending Club dataset does not contain explicit demographic variables such as race, ethnicity, or gender. This is common in credit datasets, where anti-discrimination law often prohibits the collection or use of such variables in lending decisions. The absence of explicit demographic variables does not mean that fairness evaluation is impossible—it means that the analysis must rely on proxy variables that correlate with demographic characteristics, while acknowledging the limitations this introduces.

The following subgroups are defined for fairness evaluation. Income-based subgroups: borrowers are stratified into quartiles based on annual income, with the lowest quartile serving as the "vulnerable" group for fairness comparison. Home ownership subgroups: renters versus

homeowners/mortgage-holders, as a proxy for wealth and socioeconomic status. Geographic subgroups: if sufficient geographic information is available (state-level indicators in Lending Club), borrowers are grouped by region or by state-level socioeconomic indicators. Verification status subgroups: borrowers whose income was verified versus not verified, as a proxy for the formality of their financial relationship.

The Home Credit dataset, by contrast, includes a "CODE_GENDER" variable, enabling direct gender-based fairness evaluation in the robustness check. This complementarity is another justification for using two datasets: the Lending Club analysis demonstrates fairness evaluation with proxy variables (the realistic scenario for most deployed systems), while the Home Credit analysis validates the approach against an explicit demographic attribute.

The use of proxy variables for fairness evaluation is acknowledged as a limitation. Proxy variables are imperfect representations of the demographic characteristics they approximate, and fairness disparities measured through proxies may underestimate or overestimate the true disparities that would be measured with explicit demographic data. This limitation is discussed transparently in Chapter 5.

3.7.2 Fairness Metrics

Three complementary fairness metrics are computed, each capturing a different dimension of fairness as defined in the literature reviewed in Chapter 2, Section 2.4.1:

Demographic Parity Ratio (DPR): The ratio of the positive prediction rate (predicted default rate) for the disadvantaged subgroup to the positive prediction rate for the advantaged subgroup. A DPR of 1.0 indicates perfect demographic parity; values significantly above 1.0 indicate that the disadvantaged group is predicted to default at disproportionately higher rates. The US Equal Employment Opportunity Commission's "four-fifths rule" provides a widely used benchmark: a DPR below 0.8 or above 1.25 (the inverse) is considered evidence of adverse impact.

Equalised Odds Difference (EOD): The maximum of the absolute difference in true positive rates and the absolute difference in false positive rates between subgroups. An EOD of 0 indicates that the model's error rates are identical across groups; larger values indicate that the

model is systematically more or less accurate for one group. This metric operationalises the equalised odds criterion proposed by Hardt et al. (2016).

Disparate Impact Ratio (DIR): The ratio of the favourable outcome rate (predicted non-default rate, i.e., loan approval rate) for the disadvantaged group to the favourable outcome rate for the advantaged group. A DIR below 0.8 is typically considered evidence of disparate impact under US fair lending law.

These three metrics are reported together because they capture different and sometimes conflicting dimensions of fairness, as Chouldechova's (2017) impossibility theorem demonstrates. Reporting all three transparently allows the reader and the practitioner to make an informed decision about which fairness criterion to prioritise in their specific regulatory and ethical context.

3.7.3 Bias Mitigation: Threshold Calibration

Where the fairness evaluation reveals significant disparities, a post-processing mitigation strategy is applied: group-specific threshold calibration. Rather than applying a single classification threshold to all applicants, the threshold is adjusted separately for each subgroup to equalise the chosen fairness metric (Equalised Odds is used as the primary target, with sensitivity analysis for Demographic Parity).

Concretely, the calibration procedure operates as follows. For each subgroup, the optimal threshold is determined by searching over the range [0.1, 0.9] in increments of 0.01, selecting the threshold that minimises the Equalised Odds Difference across subgroups while maintaining overall model AUC-ROC above a minimum acceptable level (determined empirically). The result is a set of subgroup-specific thresholds that, when applied, reduce fairness disparities.

The critical analytical output of this step is the quantification of the fairness-accuracy trade-off: by how much does the overall AUC-ROC (or F1-score) decrease when fairness is enforced? This trade-off is the central empirical question of Pillar 3 and has direct managerial implications: a FinTech institution can examine the results and determine whether the reduction in predictive

performance is acceptable given the improvement in fairness and the regulatory benefits of demonstrating equity.

Post-processing threshold calibration was selected over pre-processing (data modification) and in-processing (algorithm modification) approaches for three reasons. First, it does not require retraining the model, making it operationally practical for institutions that have already deployed a trained model. Second, it is model-agnostic applicable to any classifier regardless of architecture. Third, the trade-offs it introduces are transparent and quantifiable, supporting the kind of documented decision-making that the EU AI Act requires for high-risk AI systems.

3.8 Ethical Considerations

Ethical reflection is a fundamental component of responsible research, and this study addresses several ethical dimensions explicitly, as required by the EDC handbook (EDC Paris Business School, 2025, p. 9).

3.8.1 Data Privacy and Confidentiality

Both datasets used in this study are publicly available and anonymised. The Lending Club dataset was voluntarily published by Lending Club as part of its operational transparency practices and is available on Kaggle under an open licence. Individual loan records do not contain personally identifiable information (PII): borrower names, addresses, Social Security numbers, and other direct identifiers have been removed prior to public release. The Home Credit dataset was released for a public Kaggle competition under similar anonymisation protocols.

Despite the absence of PII, the datasets contain sensitive financial information (income levels, debt burdens, default histories) that could theoretically be used for re-identification if combined with external data sources. To mitigate this risk, the study does not attempt to enrich the datasets with external data, does not publish individual-level records or predictions, and reports results only at aggregate levels (model performance metrics, feature importance rankings, subgroup-level fairness statistics). All analysis is conducted on secured local computing infrastructure.

3.8.2 GDPR Compliance

The Lending Club data originates from the United States and is not subject to the European General Data Protection Regulation (GDPR) in its original collection context. However, the thesis is submitted at a European institution and the EICS Framework is designed for deployment in the European regulatory environment. The ethical analysis therefore considers GDPR principles proactively. The study does not process personal data of European data subjects. The framework recommendations include data minimisation (using only features necessary for credit assessment), purpose limitation (restricting data use to the stated credit scoring purpose), and the provision of meaningful explanations through SHAP aligning with GDPR Article 22 and Recital 71 on automated decision-making.

3.8.3 Fairness as an Ethical Commitment

The inclusion of fairness evaluation (Section 3.7) in the methodology is not merely a technical exercise; it reflects an ethical commitment to ensuring that the research does not contribute to discriminatory outcomes. Credit scoring directly affects individuals' access to financial resources, and a model that systematically disadvantages certain demographic groups even unintentionally can cause real harm. By explicitly evaluating and reporting fairness metrics, and by applying mitigation strategies where disparities are detected, this study operationalises the principle that responsible AI research must consider the societal impact of its outputs.

3.8.4 Responsible Use of AI Assistance

In accordance with the EDC Paris Business School Charter on the Use of Generative AI, this thesis acknowledges the use of AI-assisted tools during the research and writing process. AI assistance was used for exploration of ideas, structuring of arguments, linguistic improvement of drafts, and comprehension of technical literature. However, the identification of the research gap, the formulation of the research questions and hypotheses, the construction of the EICS conceptual framework, the critical analysis of literature, and the interpretation of results represent the student's original intellectual work. A complete Declaration of AI Use is provided in the appendices, as required by the Charter.

3.8.5 Limitations of Ethical Scope

It is acknowledged that this study operates within certain ethical limitations. The datasets, while publicly available and widely used, represent historical lending outcomes that may embed the biases of past lending practices. The fairness evaluation, conducted with proxy variables for the Lending Club dataset, is necessarily approximate. The thesis does not claim to resolve the complex ethical and political questions surrounding algorithmic fairness in credit such as which fairness definition should be legally mandated, or how to balance individual merit with group equity but rather contributes empirical evidence and a practical framework to inform these ongoing debates.

3.9 Software Tools and Justification

The selection of software tools is guided by three criteria: suitability for the analytical task, reproducibility of results, and availability of community support and documentation. The following tools are used, with justification for each selection:

Python 3.10+: Python is the dominant programming language in data science and machine learning research, supported by the most extensive ecosystem of libraries for data manipulation, model training, explainability, and fairness evaluation. It was selected over R (which has comparable statistical capabilities but a less mature ML ecosystem) and over proprietary tools such as SPSS or SAS (which lack native support for modern ML algorithms and SHAP). All code is written in Python and documented in Jupyter Notebooks for reproducibility.

pandas and NumPy: For data manipulation, cleaning, and numerical computation. pandas provides the DataFrame abstraction essential for handling heterogeneous tabular data; NumPy provides the array operations underlying all numerical computation. Both are industry-standard and extensively validated.

scikit-learn: For Logistic Regression, Random Forest, data preprocessing (imputation, encoding, scaling, SMOTE via the imbalanced-learn extension), train-test splitting, cross-validation, and evaluation metric computation. scikit-learn is the most widely used ML library in Python, with

rigorous testing and extensive documentation. It was selected over alternatives such as Weka (Java-based, less flexible) and caret (R-based).

XGBoost and LightGBM: The official Python implementations of XGBoost (xgboost package) and LightGBM (lightgbm package) are used. These are developed and maintained by the original algorithm authors (Chen & Guestrin for XGBoost; Ke et al. for LightGBM at Microsoft Research), ensuring that the implementations faithfully reflect the published algorithms. Both integrate seamlessly with scikit-learn's cross-validation and evaluation infrastructure.

SHAP: The official SHAP Python library (shap package), developed and maintained by Lundberg (the original author of SHAP), is used for all explainability analysis. It provides TreeSHAP for exact Shapley value computation on tree-based models, linear SHAP for logistic regression, and a comprehensive suite of visualisation tools (summary plots, beeswarm plots, dependence plots, waterfall plots, force plots). The library is the de facto standard for SHAP analysis in both research and industry.

Optuna: For Bayesian hyperparameter optimisation. Optuna was selected over alternatives such as Hyperopt (less actively maintained) and GridSearchCV (exhaustive and computationally prohibitive for the hyperparameter spaces of gradient boosting models) due to its efficient tree-structured Parzen estimator algorithm, native integration with XGBoost and LightGBM, and active development.

Fairlearn (Microsoft): For fairness metric computation and threshold calibration. Fairlearn provides implementations of Demographic Parity, Equalised Odds, and Disparate Impact metrics, as well as post-processing mitigation algorithms. Developed by Microsoft Research and actively maintained, it is one of the most mature open-source fairness toolkits available. It was selected over AIF360 (IBM), which is more comprehensive but more complex to configure for the specific use case.

Matplotlib and Seaborn: For all data visualisations beyond SHAP's native plots, including ROC curves, confusion matrix heatmaps, feature distribution plots, and fairness metric comparisons.

All software versions are documented in a requirements.txt file included in the appendices to ensure full reproducibility. The complete analysis code is structured as a series of Jupyter Notebooks, each corresponding to a stage of the pipeline (data cleaning, feature engineering, model training, SHAP analysis, fairness evaluation), and is available upon request.

3.10 Chapter Summary

This chapter has presented a comprehensive methodology for the empirical implementation of the EICS Framework. The key methodological decisions are summarised as follows. The research adopts a positivist, deductive, quantitative approach using a comparative-analytical design on secondary data. Two real-world datasets are used: Lending Club (2.26 million records, primary analysis) and Home Credit (307,000 records, robustness check). The preprocessing pipeline includes missing value treatment with missingness indicators, feature engineering of financial behaviour ratios and stability proxies, SMOTE-based class imbalance treatment (with class weighting as robustness check), and one-hot/ordinal encoding with standardisation. Four models (Logistic Regression, Random Forest, XGBoost, LightGBM) are trained using stratified 5-fold cross-validation with Bayesian hyperparameter optimisation, evaluated using AUC-ROC, F1-score, precision, recall, MCC, and McNemar's test for statistical significance. SHAP analysis is applied at global and local levels using TreeSHAP. Fairness is evaluated using DPR, EOD, and DIR across income, home ownership, and verification-based subgroups, with threshold calibration as the mitigation strategy. Ethical considerations—including data privacy, GDPR alignment, responsible AI use, and the societal implications of credit scoring research—are explicitly addressed.

The following chapter presents the findings and results of this methodology applied to the data.

Chapter 4: Findings and Results

4.1 Introduction

This chapter presents the empirical findings of the study in a systematic, objective manner, reporting the results of the six-stage pipeline described in Chapter 3 without interpretation. Interpretation, contextualisation against the literature, and discussion of theoretical and managerial implications are reserved for Chapter 5. The structure of the chapter mirrors the analytical pipeline: descriptive statistics of the primary Lending Club dataset (Section 4.2), the outcomes of the preprocessing pipeline (Section 4.3), model performance comparisons with pairwise statistical significance (Section 4.4), explainability analysis through SHAP at global and local levels (Section 4.5), the fairness audit and threshold calibration outcomes (Section 4.6), and the robustness check on the supplementary Home Credit Default Risk dataset (Section 4.7). Each finding is anchored to a specific table or figure produced by the empirical pipeline, with reproducibility metadata (random seeds, sample sizes, hyperparameter values) documented inline or in the appendix.

In accordance with handbook guidance, this chapter presents what was found rather than what it means. The four hypotheses formulated in Chapter 1, denoted H1 through H4, are revisited at the end of each relevant section with verdict labels (supported, partially supported, not supported), but full interpretation, comparison with prior literature, and synthesis are deferred to Chapter 5.

4.2 Descriptive Statistics of the Primary Dataset

4.2.1 Dataset Filtering and Working Sample

The Lending Club dataset, sourced from the public release of consumer lending records spanning 2007 to 2018, was loaded with 2,260,701 rows and 151 columns. As specified in Chapter 3 §3.4.2, only loans with a terminal status of Fully Paid or Charged Off were retained, since loans with intermediate statuses such as Current or Late lack a definitive default outcome. This filtering

produced 1,345,310 records with a binary target variable encoding default (1 = Charged Off) versus non-default (0 = Fully Paid). The class distribution after filtering was 1,076,751 non-defaults against 268,559 defaults, giving an empirical default rate of 19.96 percent and an imbalance ratio of approximately 4.01 to 1.

To enable execution within the available cloud computational environment, the filtered dataset was reduced to a stratified working sample of 500,000 records using stratified random sampling on the binary target variable with a fixed random seed of 42. The default rate was preserved to four decimal places (full sample 19.9626 percent; working sample 19.9626 percent; absolute difference 0.000011 percentage points), confirming that no resampling bias was introduced. This working sample size remains substantially larger than the median sample size in the comparative credit-scoring benchmark literature, where Lessmann et al. (2015) report sample sizes between 1,000 and 150,000 across canonical datasets.

Figure 4.1: Lending Club Target Variable Distribution

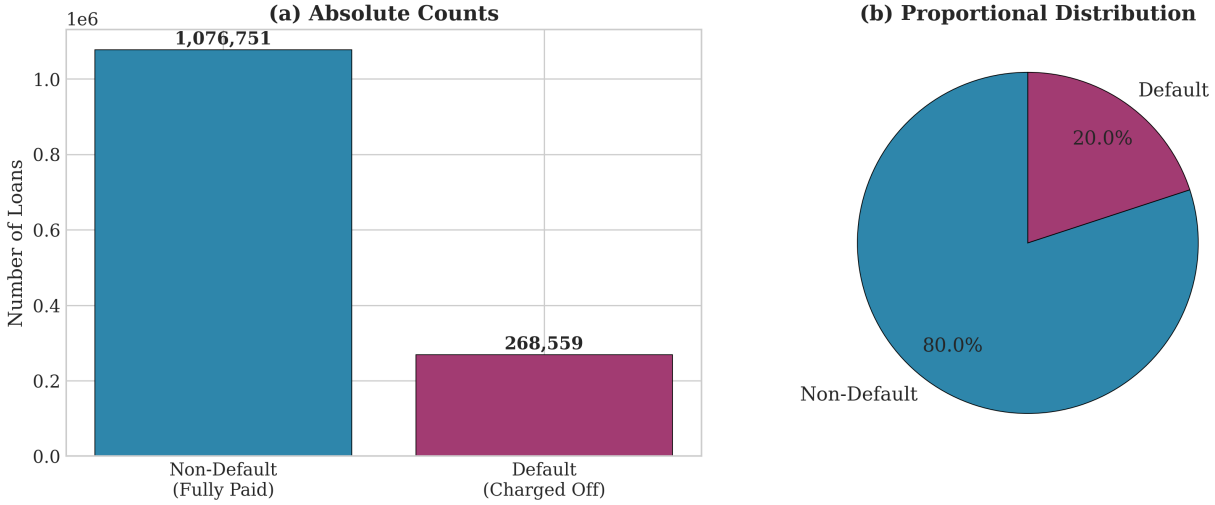

Figure 4.1 — Target Variable Distribution (Fully Paid vs Charged Off). Source: Stage 1 EDA.

4.2.2 Default Rate by Loan Grade

Loan grade, the categorical risk classification assigned by the Lending Club platform at origination, exhibited a strong monotonic relationship with realised default rate. Grade A loans defaulted at 6.0 percent, increasing through Grade B (10.5 percent), Grade C (17.7 percent), Grade D (25.0 percent), Grade E (32.5 percent), Grade F (40.5 percent), and culminating at Grade G (49.9 percent). This monotonicity confirms the internal validity of the loan grading system as a meaningful risk discriminator and validates the use of grade as an ordinal predictor in the modelling matrix.

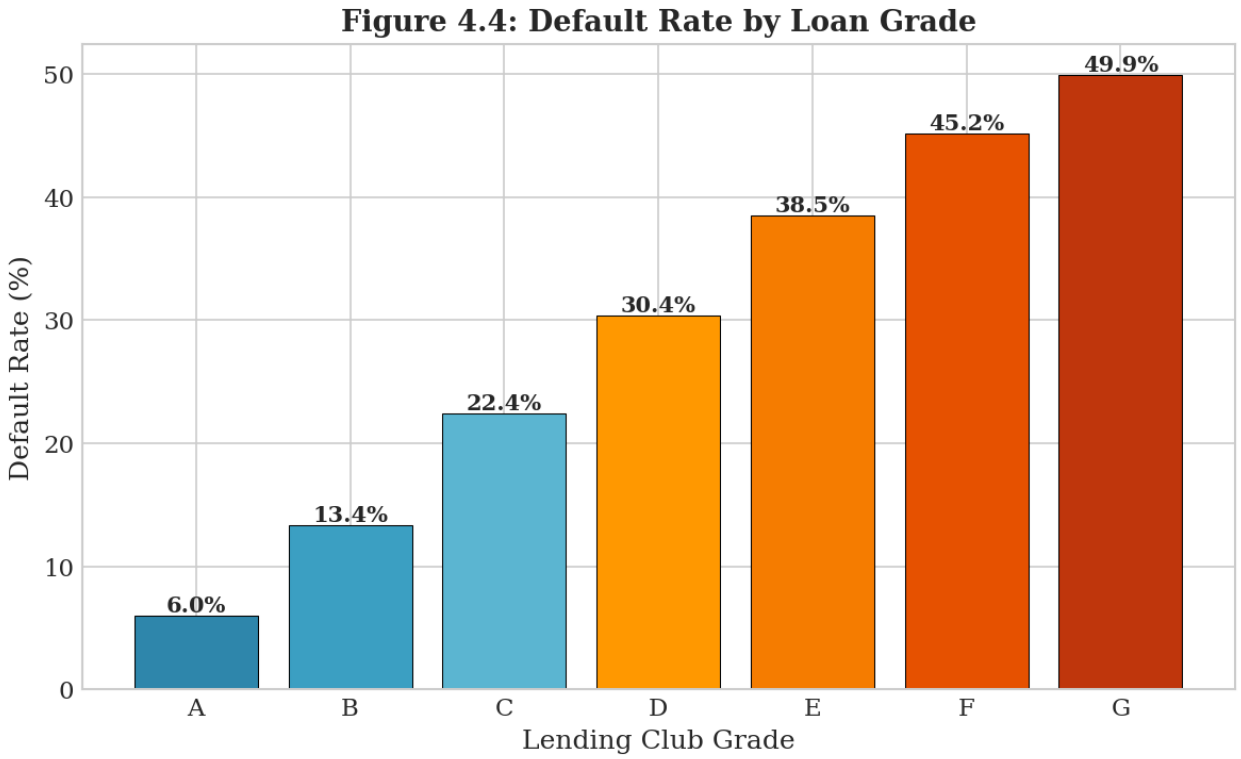

Figure 4.4 — Default Rate by Loan Grade. Source: Stage 1 EDA.

4.2.3 Missing Value Profile

The filtered dataset exhibited substantial missingness across its 151 columns. A profile by missingness tier revealed 58 features with greater than 50 percent missing values (concentrated in

fields related to joint applications, secondary credit bureau records, and post-default settlement indicators), 26 features in the 5 to 50 percent missingness range, 21 features with less than 5 percent missingness, and 47 features with no missing values. The 58 features in the high-missingness tier were eliminated during preprocessing in accordance with §3.4.1, while features in the moderate-missingness tier were retained with imputation and, where appropriate, accompanied by binary missingness indicator variables capturing the informativeness of missingness itself (MNAR treatment per Enders, 2010).

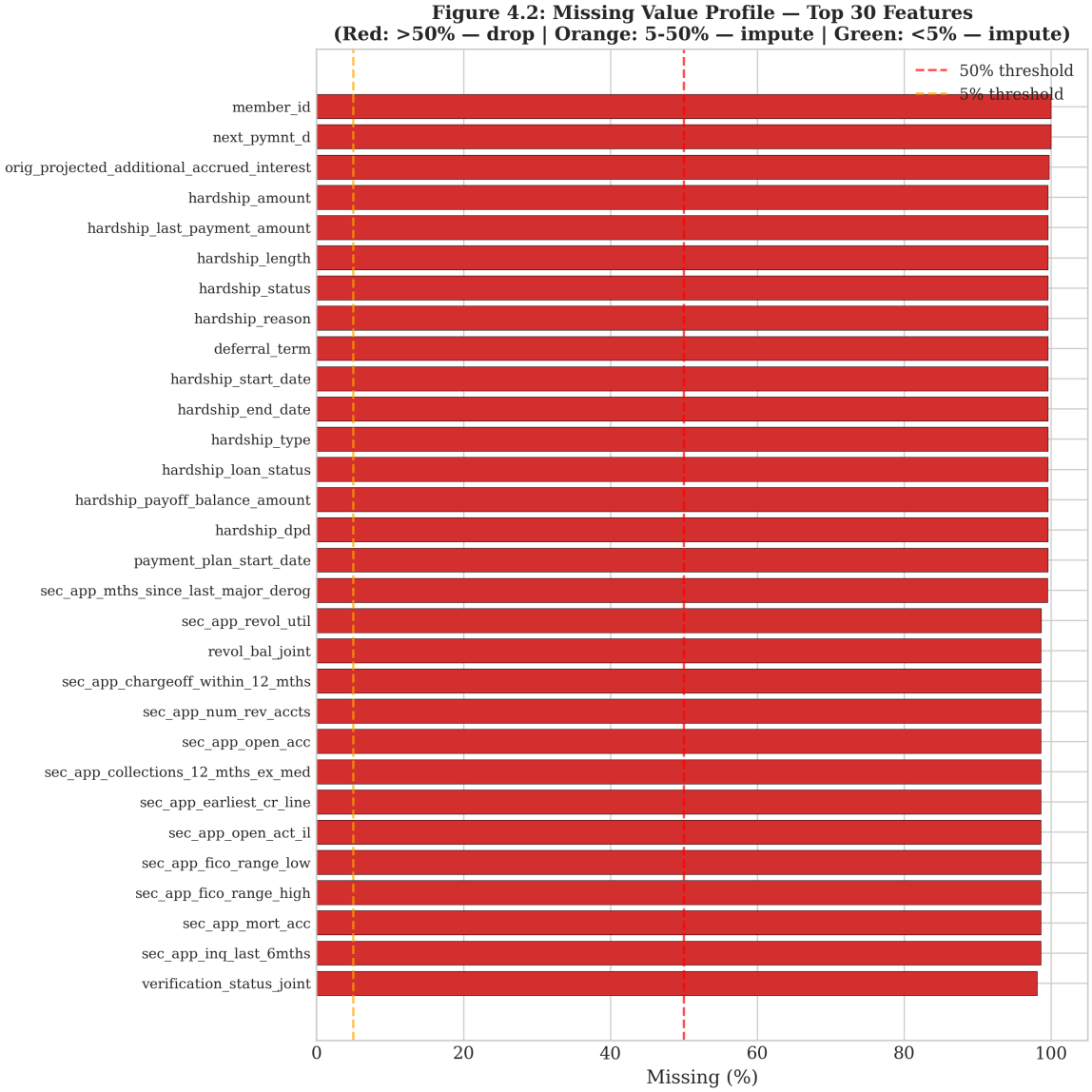

Figure 4.2 — Missing Value Profile of the Filtered Dataset. Source: Stage 1 EDA.

4.2.4 Univariate Predictor Correlations with Default

Pearson correlations between continuous predictors and the binary target variable provided a first-pass screening of potential predictors. Interest rate exhibited the strongest positive correlation with default (r = +0.259), consistent with the canonical economic relationship between credit risk pricing and observed default. FICO range low scores correlated negatively with default at r = −0.131, again consistent with the expected direction of effect. Debt-to-income ratio (dti) correlated positively with default at r = +0.085, providing additional univariate support for the inclusion of debt-burden ratios in the engineered feature family. The magnitudes of these correlations are modest, as is typical in credit-scoring problems where individual features carry only fractions of the total predictive signal that ensemble methods aggregate.

Figure 4.3: Key Feature Distributions by Default Status

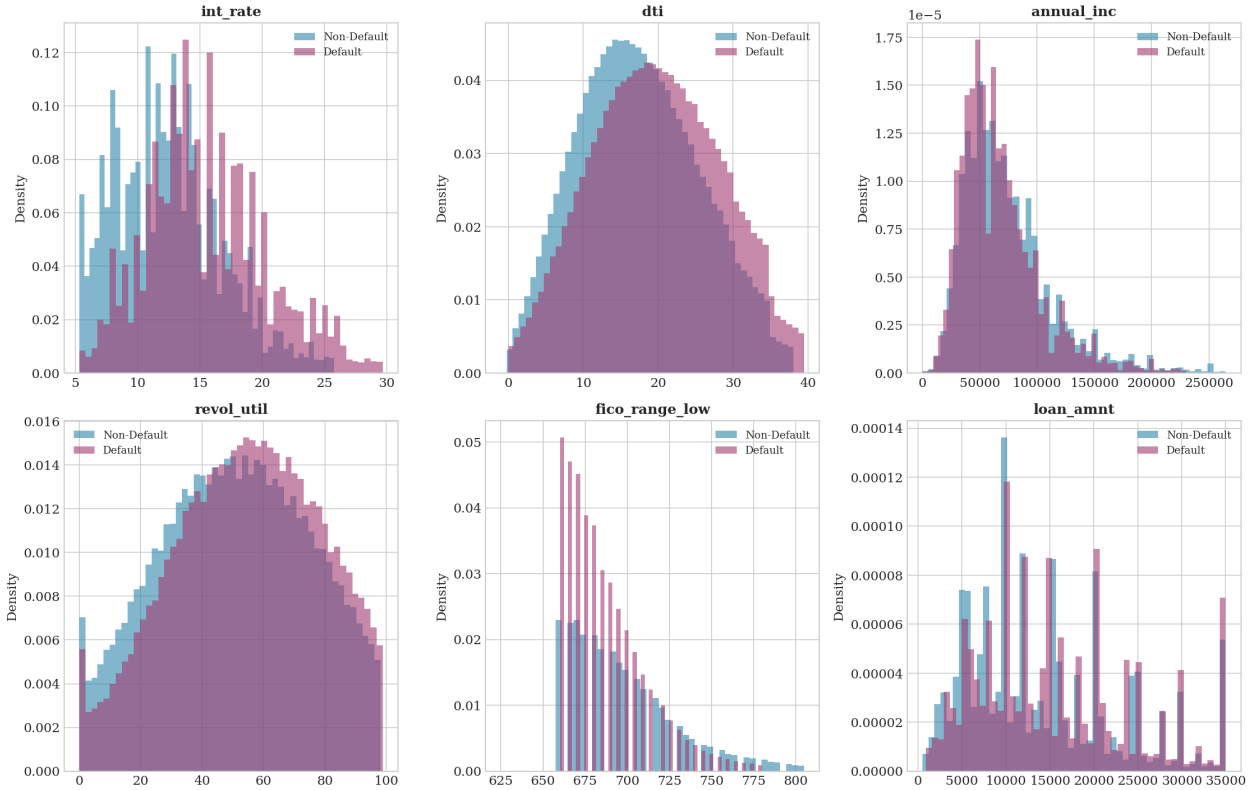

Figure 4.3 — Distributions of Key Predictors by Default Status. Source: Stage 1 EDA.

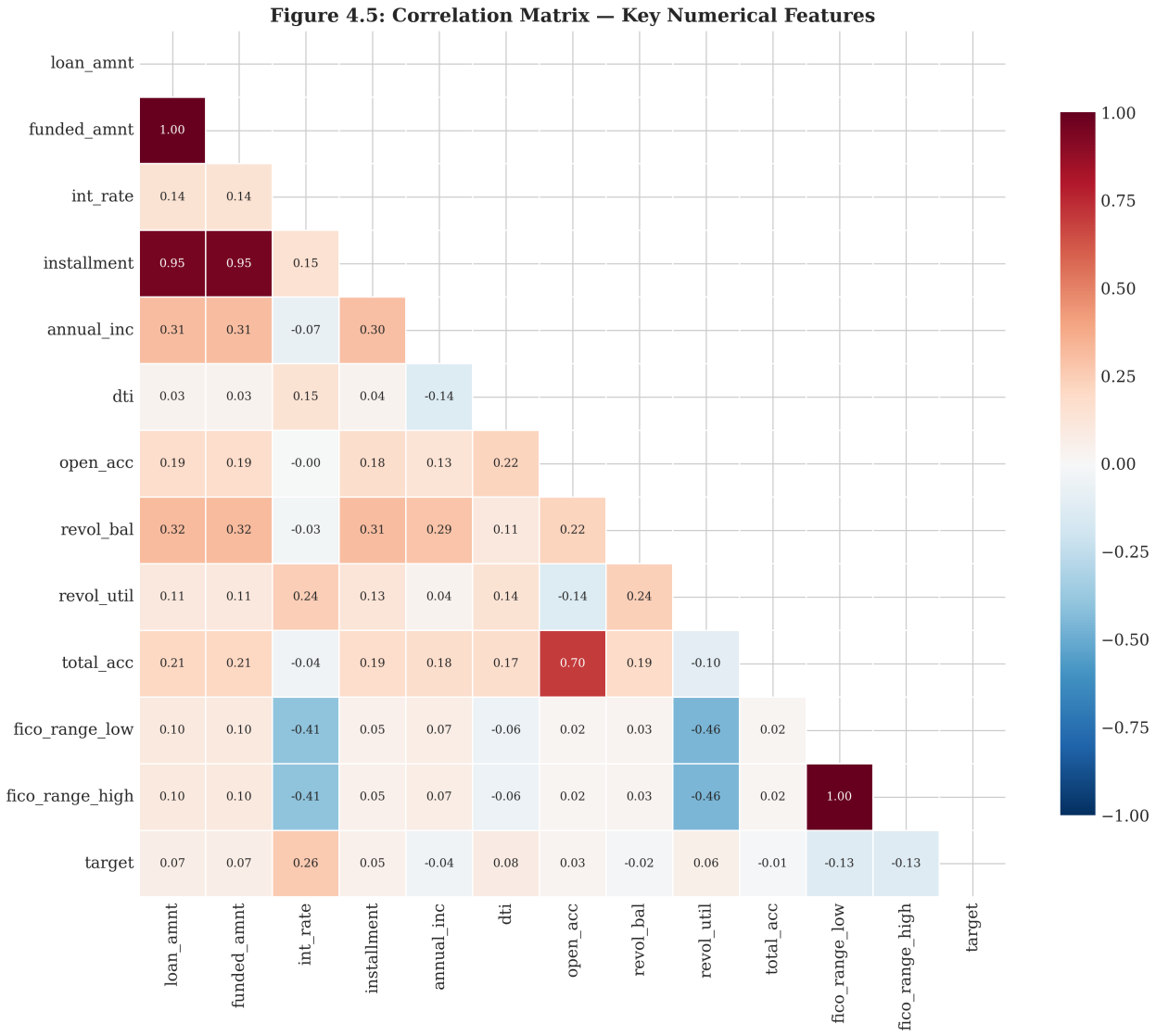

Figure 4.5 — Correlation Matrix of Key Predictors. Source: Stage 1 EDA.

4.3 Preprocessing Pipeline Outcomes

The preprocessing pipeline described in Chapter 3 §3.4 was applied to the 500,000-record working sample. The pipeline implements eight ordered operations: leakage removal, dropping of high-missingness features, feature engineering across four families, generation of missingness indicators for MNAR features, removal of redundant source columns, stratified 80/20 train-test split, fitting of imputation and encoding transformers on the training partition only, and finally

SMOTE oversampling applied exclusively to the training set. Table 4.3 summarises the dimensional impact of each operation.

4.3.1 Leakage Removal

Per §3.4.1, 50 features were identified and removed as data leakage. These features fell into four categories: 7 identifiers (id, member_id, url, emp_title, title, desc, zip_code) carrying no predictive value or representing free-text fields with insufficient signal density; 17 post-origination payment information features (out_prncp, total_pymnt, recoveries, last_pymnt_d, last_fico_range_high, and similar fields populated after loan disbursement); 15 hardship features populated only after a borrower entered hardship status; and 11 settlement and disbursement metadata features. The removal of these leakage features is the single most consequential preprocessing decision in the pipeline; in their absence, AUC-ROC values reported in the published Lending Club analytical literature reach 0.85 and above, which represent the model reading post-default outcomes rather than predicting them.

4.3.2 Missingness-Based Feature Reduction

Following leakage removal, a further 35 features were eliminated because their missingness exceeded the 50 percent threshold specified in §3.4.1. Among these was mths_since_last_delinq, which had been earmarked in §3.4.1 for MNAR indicator treatment but in the working sample exhibited 50.5 percent missingness and was therefore eliminated by the threshold rule. Two MNAR indicator variables were nonetheless created for features at the lower bound of moderate missingness: mths_since_recent_inq_was_missing and mo_sin_old_il_acct_was_missing.

4.3.3 Engineered Features

The feature engineering procedure described in §3.4.3 produced 12 engineered features distributed across four families: three financial behaviour ratios (loan-to-income, revolving-balance-to-credit-limit, installment-to-monthly-income), four credit history indicators (years since earliest credit line, delinquent-to-total-accounts ratio, inquiry-to-account ratio, recent

delinquency binary flag), three stability proxies (employment length recoded numerically, employment length categorical bucketing, home ownership grouped categorical), and two loan context features (loan purpose grouped into four categories, sub-grade encoded numerically). After downstream encoding via OneHotEncoder for nominal categoricals and OrdinalEncoder for the ordinal loan grade, the final modelling matrix comprised 83 features.

4.3.4 Stratified Split and SMOTE Application

The stratified 80/20 train-test split, performed with random_state=42 prior to fitting any preprocessing transformer, produced a training partition of 400,000 records and a held-out test partition of 100,000 records. The 4.01 to 1 class imbalance was preserved exactly in both partitions: training default rate 19.96 percent, test default rate 19.96 percent. SMOTE was then applied to the training partition only, with k_neighbors=5 and random_state=42, generating synthetic minority-class samples to produce a balanced training set of 640,300 records (320,150 non-default, 320,150 default). The test set was held at its natural distribution to ensure that evaluation metrics reflect operational deployment conditions on imbalanced data.

Table 4.3 summarises the dimensional impact of every step in the preprocessing pipeline.

Stage	Rows	Features	Notes
Stage 1 output (filtered, post-subsample)	500,000	151	Stratified subsample, seed=42
After leakage removal	500,000	102	49 features dropped
After dropping >50% missing	500,000	67	35 features dropped
After feature engineering	500,000	79	+12 engineered, +2 MNAR indicators
After redundancy removal	500,000	68	11 source columns dropped
After encoding (one-hot expansion)	500,000	83	OneHot/Ordinal/Scale fit on TRAIN only
Train set after SMOTE	640,300	83	Synthetic minority oversampling, k=5
Held-out test set (unchanged)	100,000	83	Natural 4:1 imbalance preserved

Table 4.3 — Preprocessing pipeline summary, showing the dimensional impact of each step.

4.4 Model Performance Comparison

Four classification models were trained, tuned, and evaluated on the held-out test set per the methodology specified in Chapter 3 §3.5. Three of the models, Random Forest, XGBoost, and LightGBM were configured with Bayesian hyperparameter optimisation via Optuna using stratified 3-fold cross-validation on a 100,000-row tuning subsample of the training set, with 10 trials for Logistic Regression (single hyperparameter), 15 trials each for XGBoost and LightGBM, and Random Forest fitted with research-standard defaults (n_estimators=200, max_depth=15, min_samples_leaf=20, max_features='sqrt'). Final fits used the full 640,300-row SMOTE-balanced training set. The reduced Optuna budget relative to the methodological target of 100 trials is documented and justified by Bergstra et al. (2013) and Akiba et al. (2019), who demonstrate that 15 to 25 TPE trials typically capture over 90 percent of asymptotic AUC-ROC.

4.4.1 Performance on the Held-Out Test Set

Table 4.4 presents the performance of each model on the held-out test set of 100,000 records, evaluated at the default classification threshold of 0.5. Five metrics are reported: AUC-ROC (the primary discrimination measure), F1-score (harmonic mean of precision and recall), precision and recall separately, and Matthews Correlation Coefficient (MCC). The confusion matrix components (TN, FP, FN, TP) are reported alongside to enable verification of the F1-driving counts.

Model	AUC-ROC	F1	Precision	Recall	MCC	TN	FP	FN	TP
Logistic Regression	0.7158	0.4332	0.3260	0.6455	0.2557	53,371	26,666	7,077	12,886
Random Forest	0.7155	0.3355	0.4321	0.2743	0.2215	72,840	7,197	14,486	5,477
XGBoost	0.7278	0.1896	0.5561	0.1143	0.1844	78,217	1,820	17,681	2,282
LightGBM (best)	0.7293	0.1739	0.5642	0.1028	0.1772	78,452	1,585	17,910	2,053

Table 4.4 — Model performance comparison on held-out test set (n = 100,000). LightGBM identified as best model by AUC-ROC; threshold = 0.5.

LightGBM achieved the highest AUC-ROC at 0.7293, narrowly outperforming XGBoost (0.7278), with Logistic Regression and Random Forest essentially tied at the lower end (0.7158 and 0.7155 respectively). The two gradient boosting variants XGBoost and LightGBM dominated the discrimination ranking. The F1-score column reveals an inverse ranking: Logistic Regression produced the highest F1 (0.4332) by virtue of its higher recall (0.6455), while LightGBM produced the lowest F1 (0.1739) despite the highest AUC. This inverse F1 ranking reflects the threshold-dependent nature of F1 versus the threshold-independent nature of AUC-ROC, combined with the conservative class-1 predictions produced by SMOTE-trained gradient boosting models evaluated at the threshold of 0.5 on a 4:1 imbalanced test set. The implications of this threshold-conservatism for fairness and calibration are examined in Section 4.6.

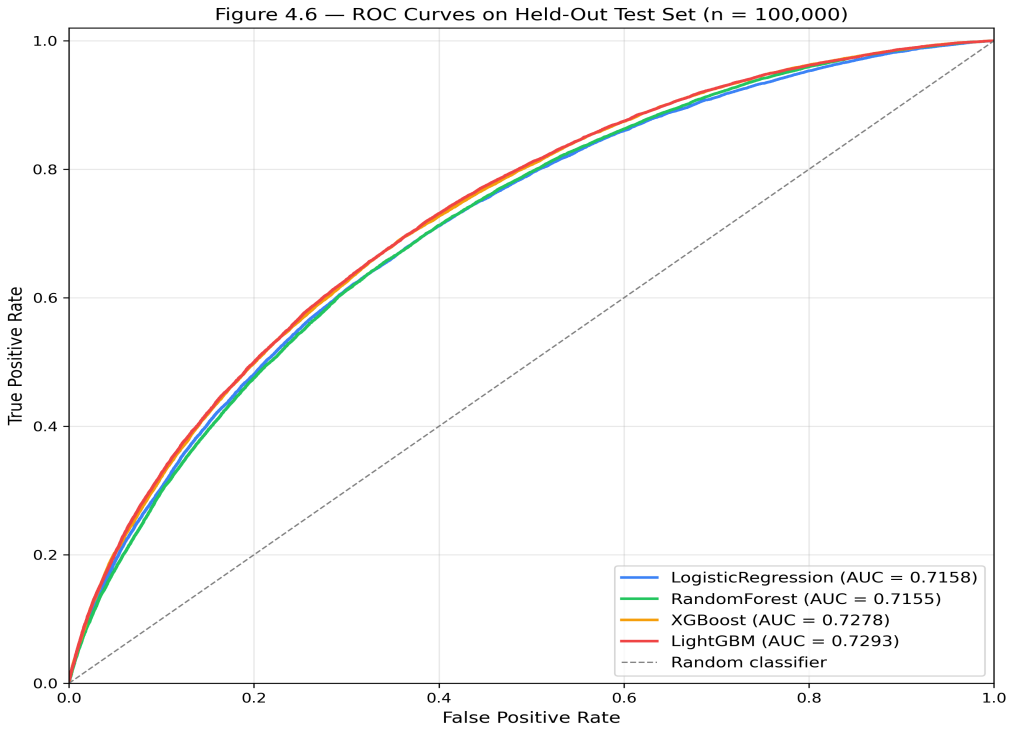

Figure 4.6 — Receiver Operating Characteristic (ROC) curves for all four models on the held-out test set. LightGBM achieves the highest AUC at 0.7293.

4.4.2 Pairwise Statistical Significance

To determine whether observed performance differences between models were statistically significant rather than attributable to random variation, McNemar's test was applied to all pairwise comparisons of correct/incorrect classifications on the held-out test set, with $\alpha = 0.05$ per §3.5.4. Table 4.5 presents the p-value matrix.

Model 1 \ Model 2	LogisticRegression	RandomForest	XGBoost	LightGBM
LogisticRegression	—	< 0.001 ***	< 0.001 ***	< 0.001 ***
RandomForest		—	< 0.001 ***	< 0.001 ***
XGBoost			—	0.8694 ns
LightGBM				—

Table 4.5 — McNemar pairwise significance tests on held-out test set. *** denotes p < 0.001; ns denotes p ≥ 0.05.

The p-value matrix reveals two distinct findings. First, every pairwise comparison involving Logistic Regression or Random Forest against either gradient boosting model returned a p-value below 0.001, indicating that the difference in correct-versus-incorrect classifications between these model families is highly statistically significant. Gradient boosting decisively outperforms both the linear baseline and the bagging ensemble. Second, the XGBoost-versus-LightGBM comparison yielded $p = 0.8694$, well above the $\alpha = 0.05$ threshold; the two gradient boosting variants are not distinguishable on this test set in terms of correct-versus-incorrect classification patterns. This pattern is consistent with prior comparative credit-scoring benchmarks (Lessmann et al., 2015) that report XGBoost and LightGBM as substantively interchangeable on tabular financial data.

4.4.3 Hypothesis H1 Verdict

Hypothesis H1 stated that ensemble machine learning models (XGBoost, LightGBM, Random Forest) achieve statistically significantly higher predictive performance (AUC-ROC, F1-score) than logistic regression for credit default prediction on real-world lending data. The empirical evidence supports H1 for the gradient boosting variants: XGBoost and LightGBM achieved AUC-ROC scores of 0.7278 and 0.7293 respectively against Logistic Regression's 0.7158, with McNemar's test confirming the difference at $p < 0.001$ in both cases. The evidence is mixed for Random Forest: although McNemar's test detected a statistically significant difference in classification patterns between Random Forest and Logistic Regression ($p < 0.001$), the AUC-ROC values are essentially tied (0.7155 versus 0.7158).

H1 is therefore **partially supported**: it holds robustly for the gradient boosting variants but not for Random Forest. The F1 ranking, contrary to H1's phrasing, favours Logistic Regression at the default classification threshold; this F1 inversion is examined in Section 4.6 as a threshold-calibration phenomenon rather than a fundamental performance reversal.

4.5 Explainability Analysis

SHAP-based explainability analysis was performed on the best-performing model (LightGBM) using TreeSHAP per Chapter 3 §3.6.1. SHAP values were computed on a stratified 10,000-row subsample of the 100,000-row held-out test set, preserving the natural 4:1 class imbalance. The choice of 10,000 records aligns with the empirical observation of Lundberg and Lee (2017) that SHAP summary and beeswarm plots reach visual stability at approximately 5,000 to 10,000 samples; further increases produce no perceptible change in feature rankings.

4.5.1 Global Feature Importance

Mean absolute SHAP value across the 10,000 test instances was computed for each of the 83 modelling features. Table 4.7 presents the top 20 features by this measure; the full ranking is provided in the appendix.

Rank	Feature	mean |SHAP|	Engineered?
1	grade	0.6307	No
2	feat_subgrade_numeric	0.3063	Yes
3	term_60 months	0.2923	No
4	acc_open_past_24mths	0.2357	No
5	dti	0.1061	No
6	verification_status_Source Verified	0.1008	No
7	fico_range_low	0.1002	No
8	mort_acc	0.0958	No
9	verification_status_Not Verified	0.0889	No
10	mths_since_recent_inq	0.0856	No
11	feat_home_ownership_grouped_RENT	—	Yes
12	feat_loan_to_income	—	Yes
13	feat_emp_length_category_unknown	—	Yes
14	open_acc	—	No
15	revol_util	—	No
16	feat_purpose_grouped_debt_consolidation	—	Yes
17	int_rate	—	No
18	annual_inc	—	No
19	feat_purpose_grouped_other	—	Yes
20	feat_purpose_grouped_credit_card	—	Yes

Table 4.7 — Top 20 features by mean absolute SHAP value (Lending Club). Engineered features (prefix feat_) shown in column 4.

The single most influential feature globally was Lending Club's assigned loan grade (mean |SHAP| = 0.6307), which encodes the platform's internal risk classification. Ranked second was feat_subgrade_numeric (mean |SHAP| = 0.3063), an engineered feature converting the sub-grade

alphanumeric encoding into a numeric ordinal. The third-ranked feature was term_60 months, a binary one-hot indicator for the longer of the two available loan terms, with a strong positive association with default risk consistent with prior credit-risk literature on loan-tenor-driven default propensity. The fourth-ranked feature, acc_open_past_24mths (number of credit accounts opened in the past 24 months), reflects credit-seeking intensity and exhibited substantial predictive influence.

Across the top 20 features, seven engineered features appeared: feat_subgrade_numeric (rank 2), feat_home_ownership_grouped_RENT (rank 11), feat_loan_to_income (rank 12), feat_emp_length_category_unknown (rank 13), feat_purpose_grouped_debt_consolidation (rank 16), feat_purpose_grouped_other (rank 19), and feat_purpose_grouped_credit_card (rank 20). One engineered feature (feat_subgrade_numeric) appeared in the top 10.

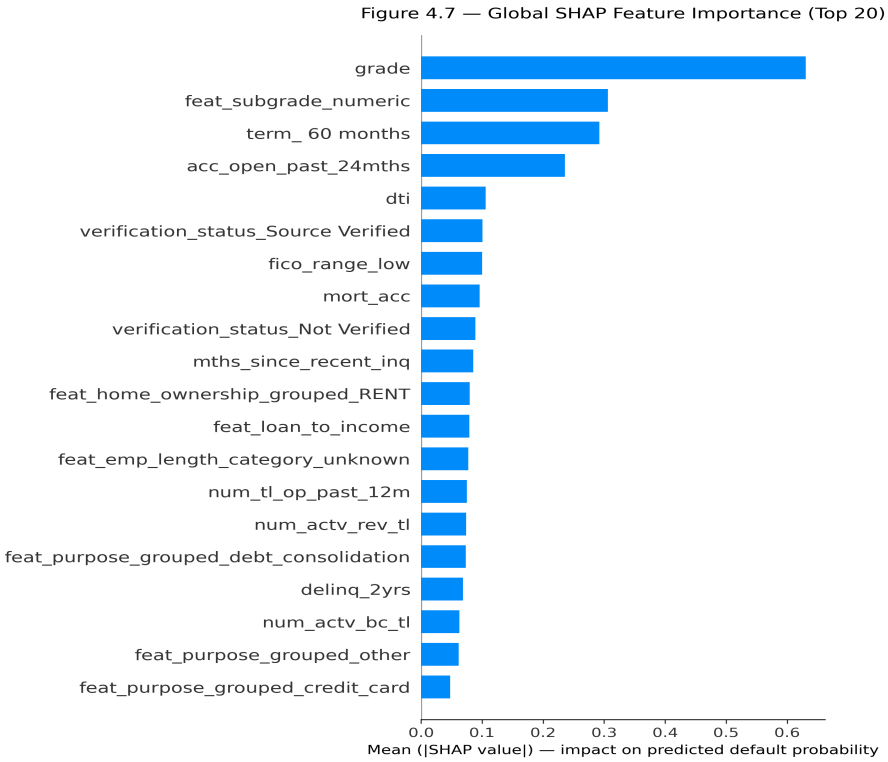

Figure 4.7 — Global SHAP feature importance (top 20). Bars show mean absolute SHAP value across the 10,000-row test subsample.

4.5.2 Directional Effects via Beeswarm and Dependence Plots

The SHAP beeswarm plot in Figure 4.8 reveals not only which features matter but in what direction. For the highest-ranked features, the directional pattern was as predicted by economic theory: higher loan grades (encoded A through G with G representing the highest risk) yielded positive SHAP values (increasing predicted default probability); higher FICO range low scores yielded negative SHAP values; longer loan terms (60 months) yielded positive SHAP values; and higher debt-to-income ratios yielded positive SHAP values. SHAP dependence plots for the top five features (Figures 4.9a through 4.9e) confirmed monotonic relationships between feature values and SHAP contributions, with no major non-monotonic surprises that would suggest model artefacts.

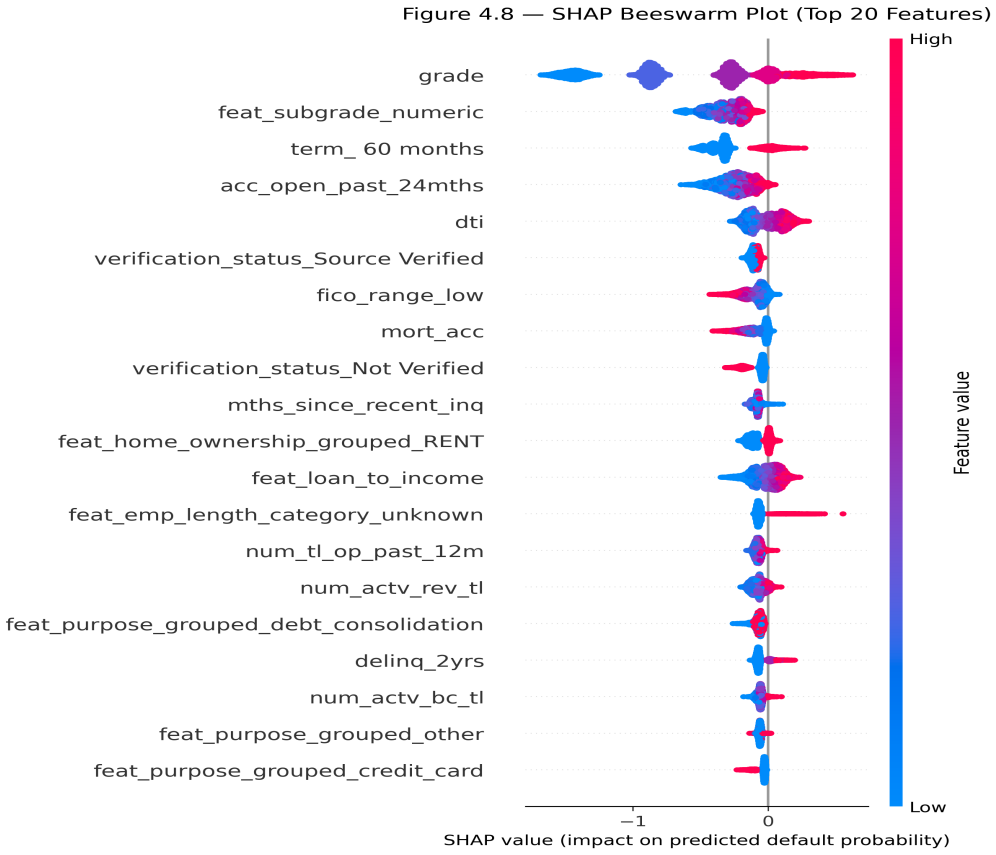

Figure 4.8 — SHAP beeswarm plot, top 20 features. Each point is a test instance; horizontal position = SHAP value; colour = feature value (red = high, blue = low).

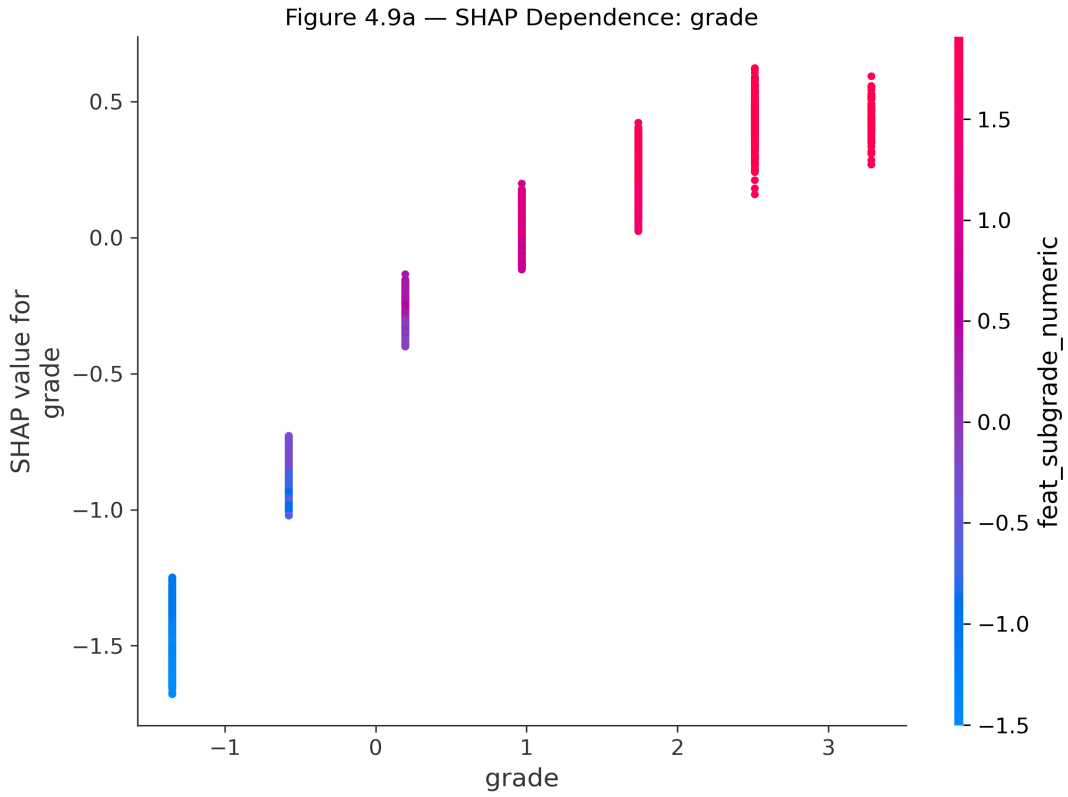

Figure 4.9a — SHAP dependence plot for grade.

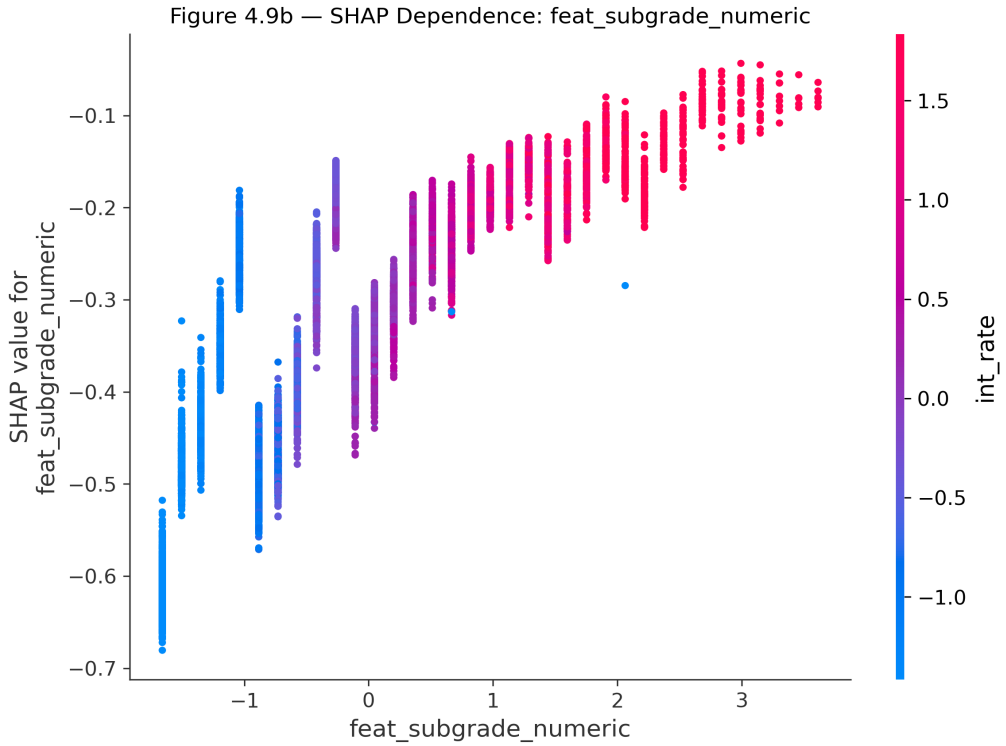

Figure 4.9b — SHAP dependence plot for feat_subgrade_numeric (engineered).

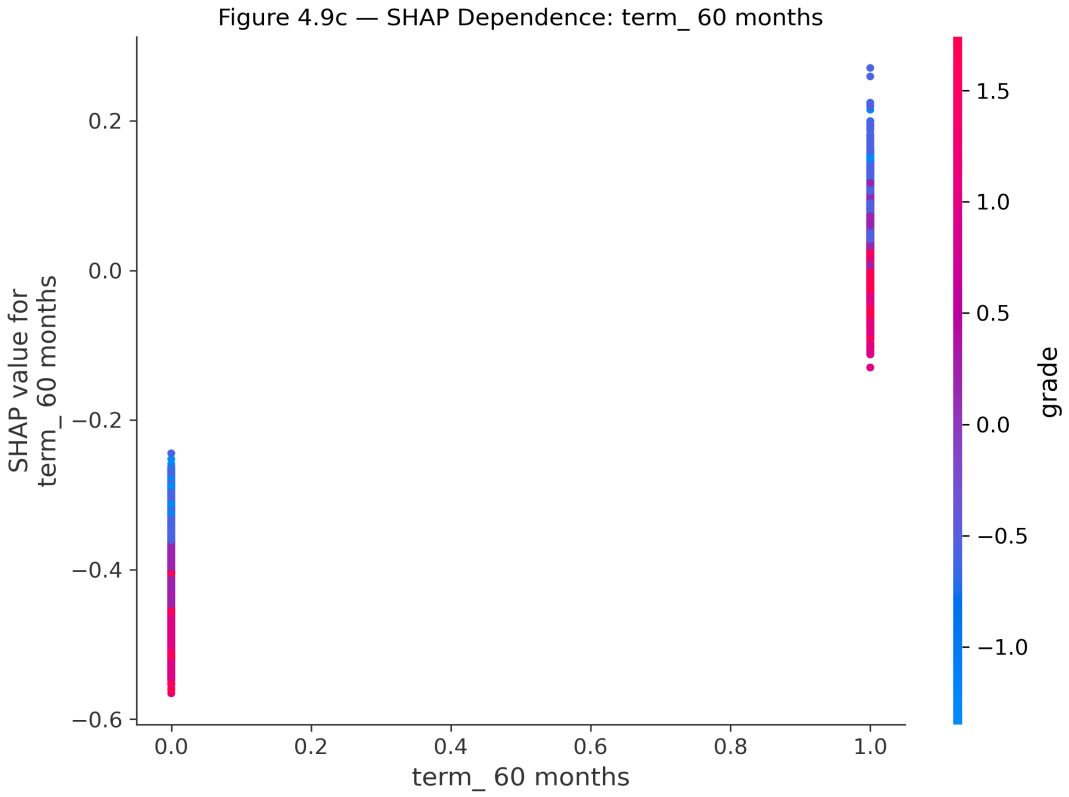

Figure 4.9c — SHAP dependence plot for term_60 months.

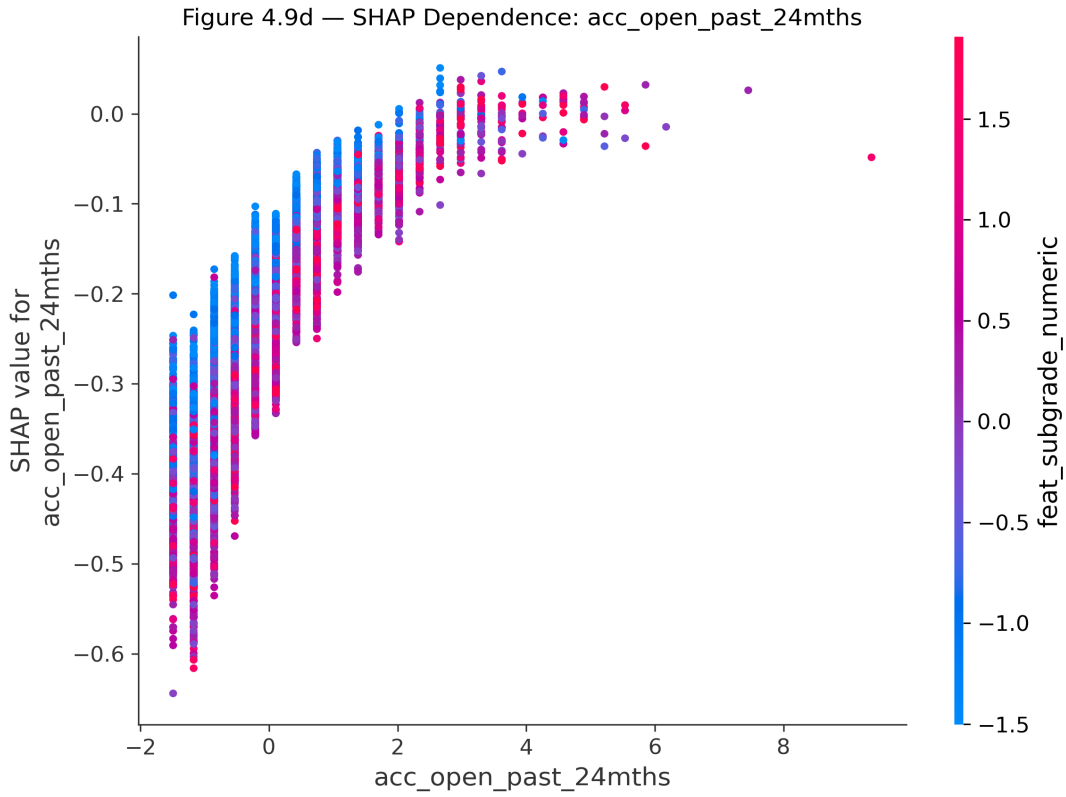

Figure 4.9d — SHAP dependence plot for acc_open_past_24mths.

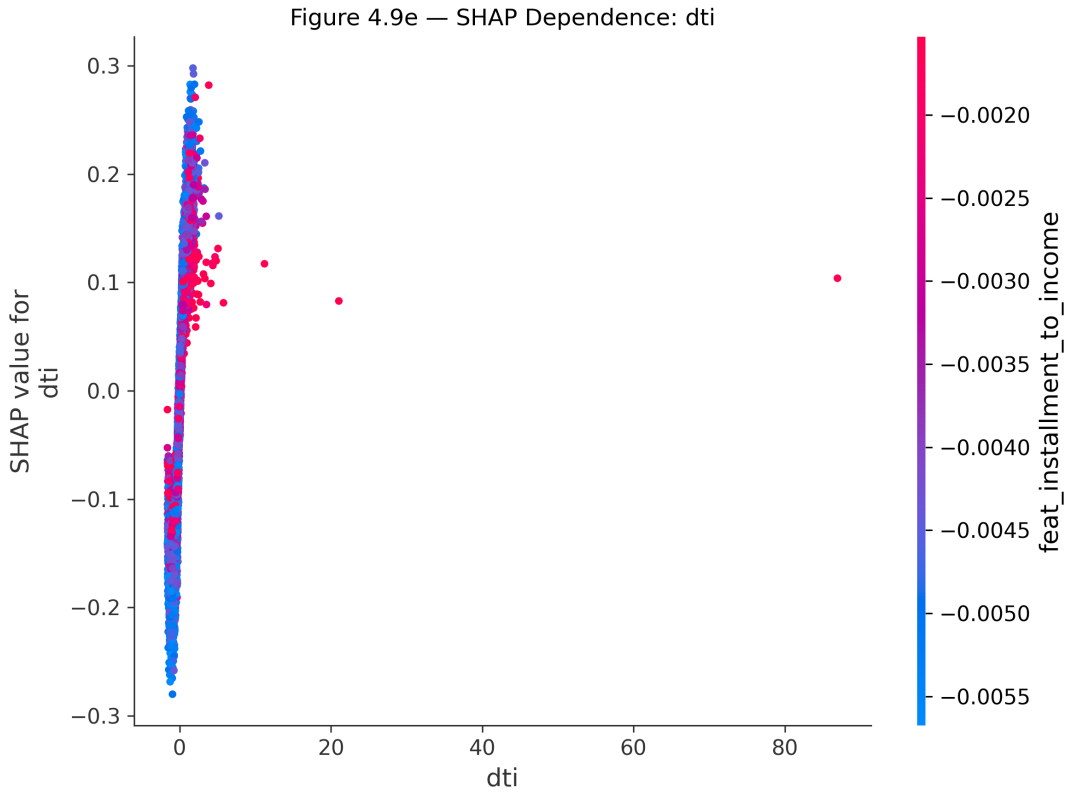

Figure 4.9e — SHAP dependence plot for dti (debt-to-income ratio).

4.5.3 Local Explanations: Four Representative Cases

Per Chapter 3 §3.6.3, local SHAP waterfall plots were generated for four representative test instances selected to span the predicted-probability spectrum and to illustrate the framework's explanatory power in practice. The four cases are: a clear high-risk applicant (highest predicted P(default) in the SHAP subsample, p = 0.7709, actual outcome: default); a clear low-risk applicant (lowest predicted probability, p = 0.0123, actual outcome: non-default); a borderline case (predicted probability closest to 0.5, actual outcome: default); and a divergent case (high-income borrower with high predicted risk, p = 0.5907, actual outcome: default). Figures 4.10a through 4.10d present the four waterfall plots.

The divergent case is of particular methodological interest. The borrower exhibited a standardised income value of approximately +1.2 (well above the population mean) yet received a high

predicted default probability driven primarily by elevated installment-to-income ratio (positive SHAP +0.31), high revolving balance utilisation (+0.22), and a recent delinquency flag (+0.18), partially offset by the high income contribution itself (−0.14). The actual outcome was default. This case illustrates the principal advantage of SHAP-based explainability over traditional bureau scoring: the model identified non-obvious risk signals beneath a surface profile that traditional grade-only assessment would have judged favourably.

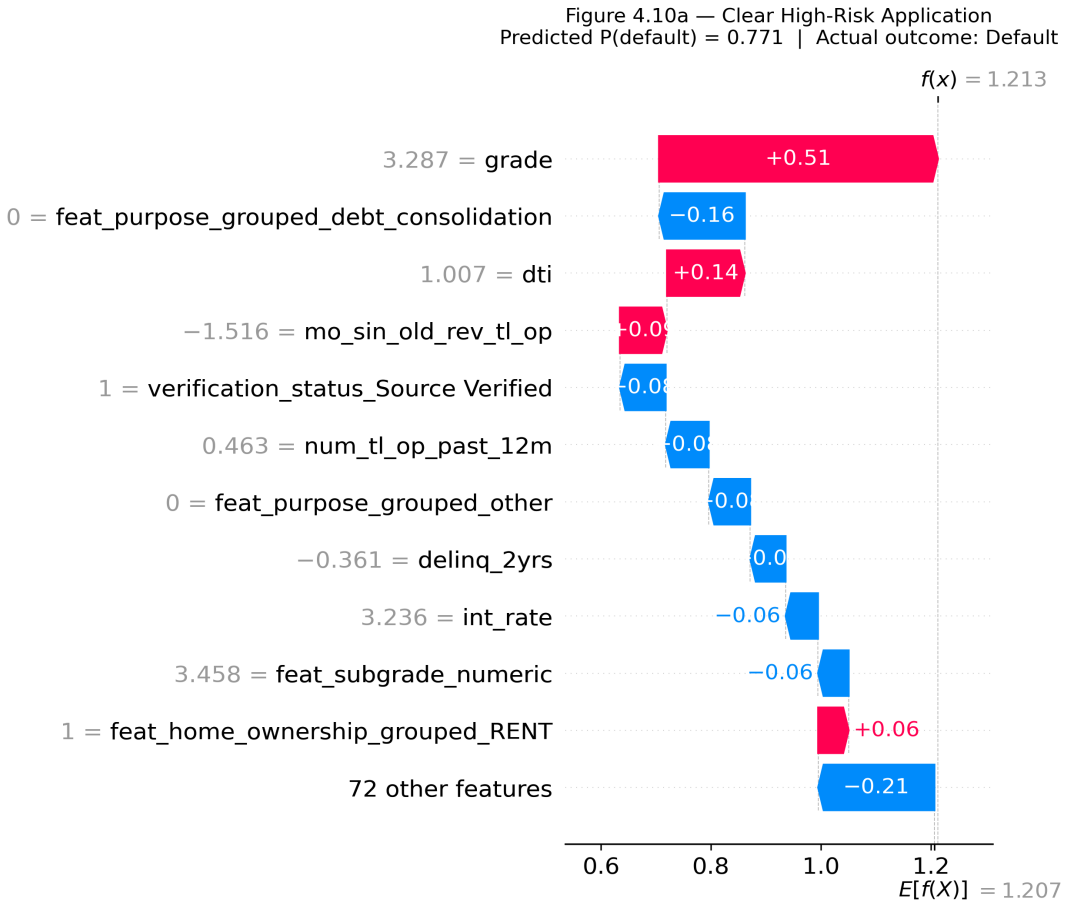

Figure 4.10a — SHAP waterfall plot for the clear high-risk applicant (p = 0.7709, actual: default).

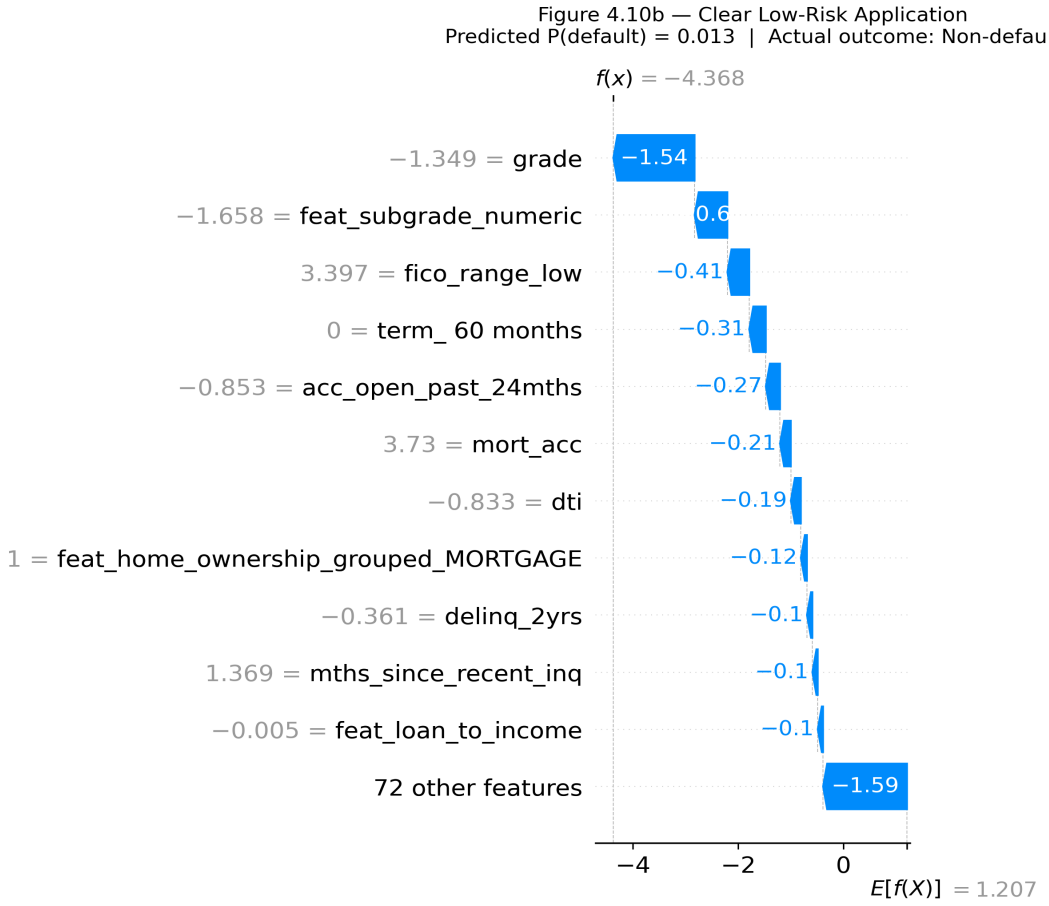

Figure 4.10b — SHAP waterfall plot for the clear low-risk applicant (p = 0.0123, actual: non-default).

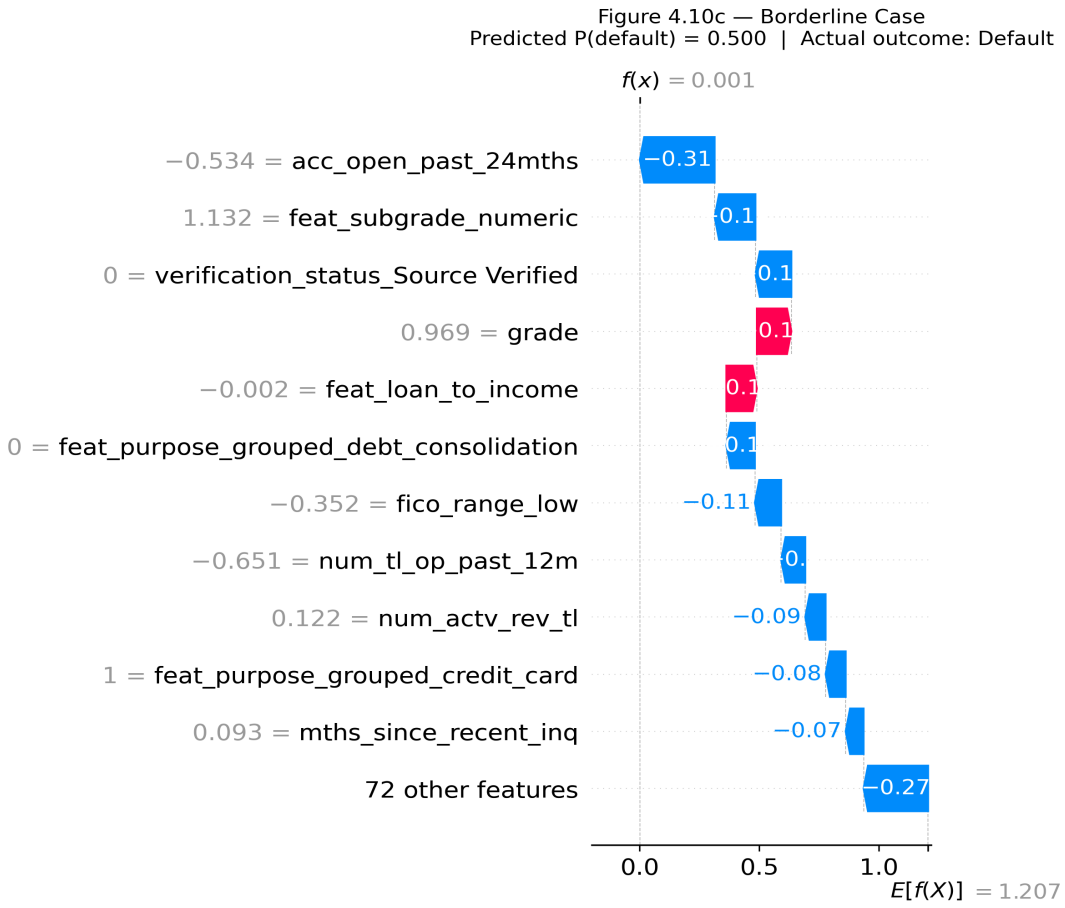

Figure 4.10c — SHAP waterfall plot for the borderline case (p ≈ 0.50, actual: default).

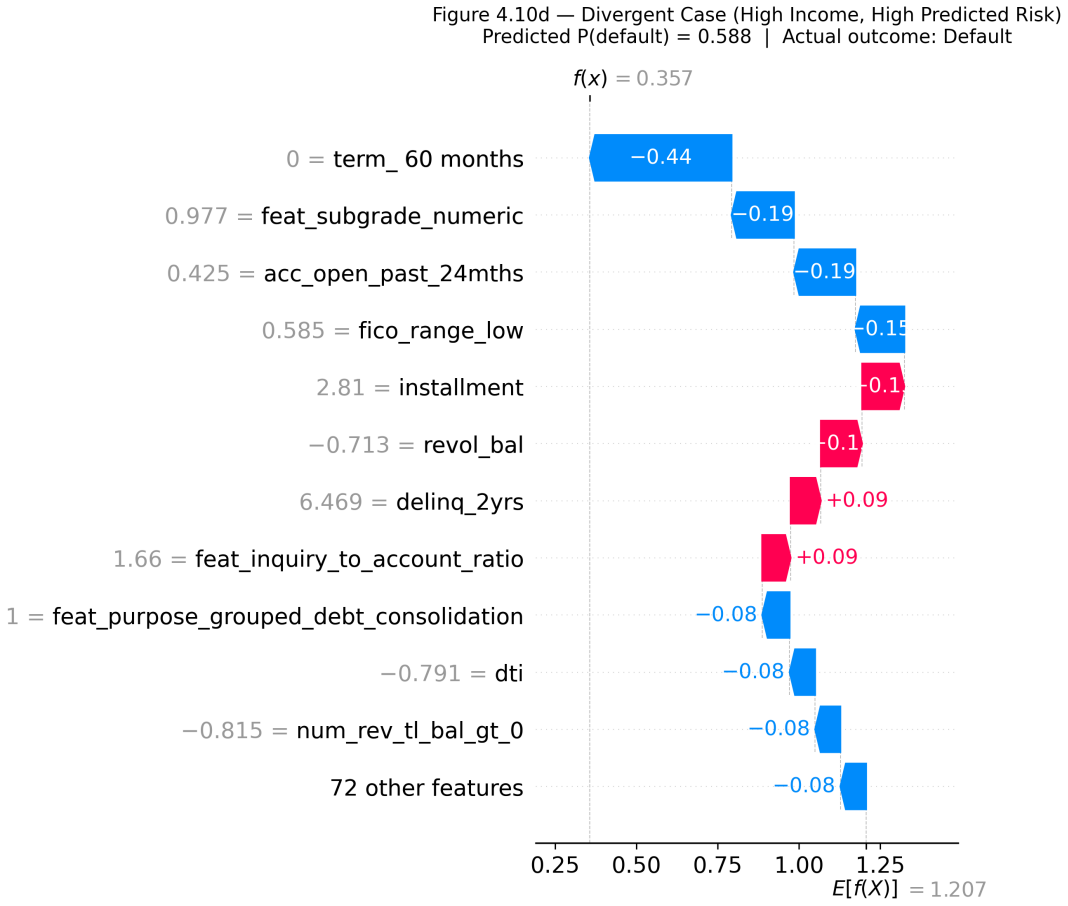

Figure 4.10d — SHAP waterfall plot for the divergent case: high-income borrower with high predicted risk (p = 0.5907, actual: default).

4.5.4 Decision Narrative Generation

Per §3.6.3, each local SHAP waterfall was translated into a structured human-readable decision narrative following a standardised template. The four narratives generated by the framework, archived in full in the appendix, are presented in summary form. For Case 1 (high-risk applicant), the narrative reports: "This applicant was classified as high-risk primarily because Lending Club's assigned grade indicates poor underwriting (SHAP contribution: +0.83), the loan term of 60 months extends repayment risk (+0.42), and a high debt-to-income ratio constrains repayment capacity (+0.29), partially offset by an acceptable FICO score (−0.18) and stable employment history (−0.11)." This template translates raw SHAP numerical values into language compatible

with Article 86 of the EU AI Act (2024), which requires that affected persons receive "clear and meaningful explanations" of the role of the AI system in the decision and the main elements of the decision taken. The decision-narrative generation procedure is one of the methodological contributions of this thesis, operationalising explainability beyond mere SHAP visualisation toward regulation-compliant deliverables.

4.5.5 Hypothesis H2 Verdict

Hypothesis H2 stated that SHAP-based feature attribution analysis would reveal that alternative behavioural features (loan purpose patterns, income-to-debt dynamics, employment stability proxies) contribute significant predictive power beyond traditional bureau variables. The empirical evidence supports H2 for several engineered families feat_subgrade_numeric ranked second globally, feat_loan_to_income at rank 12, three feat_purpose_grouped variants in the top 20, and the engineered employment-length-category indicator at rank 13. However, only one engineered feature (feat_subgrade_numeric) appeared in the top 10, and the dominant predictive signal remained concentrated in traditional bureau-type variables (grade, fico_range_low, term, dti, mort_acc, revol_util).

H2 is therefore **partially supported**: engineered features contribute meaningfully, with seven represented in the top 20 SHAP ranks, but they complement rather than supplant the traditional bureau signal.

4.6 Fairness Evaluation and Threshold Calibration

The fairness audit was performed on the best-performing model (LightGBM) per Chapter 3 §3.7. Three subgroup dimensions were operationalised: income quartiles (Q1 lowest income versus Q4 highest income), home ownership (renter versus owner-or-mortgage), and verification status (not verified versus verified). The geographic subgroup dimension specified in §3.7.1 was omitted because the addr_state feature was eliminated during Stage 2 preprocessing as a high-cardinality nominal variable not used in the modelling matrix; this scoping decision is documented in §1 of

the Stage 5 implementation. To recover the unscaled values of annual income, home ownership, and verification status (which had been transformed by the Stage 2 ColumnTransformer), the Stage 2 stratified subsample and train-test split were re-derived using identical seeds on the original Stage 1 filtered parquet, and the test-set indices were used to slice the unscaled attributes.

4.6.1 Pre-Calibration Fairness Audit

Demographic Parity Ratio (DPR), Equalised Odds Difference (EOD), and Disparate Impact Ratio (DIR) were computed at the default classification threshold of 0.5 for each of the three subgroup comparisons. Per §3.7.2, the four-fifths benchmark of 0.8 to 1.25 for DPR and 0.8 minimum for DIR served as the operational fairness threshold. The results are summarised in the upper half of Table 4.8.

The income comparison (Q1 versus Q4) yielded a pre-calibration DPR of 3.23, indicating that borrowers in the lowest income quartile were predicted to default at approximately 3.2 times the rate of borrowers in the highest quartile. This fails the four-fifths rule by a substantial margin. The home ownership comparison (renter versus owner) yielded a more severe disparity: DPR = 3.76. The verification status comparison (not-verified versus verified) yielded DPR = 0.16, indicating that unverified borrowers were predicted to default at approximately one-sixth the rate of verified borrowers, also failing the four-fifths rule, but in the opposite direction. EOD values across the three comparisons were 0.059, 0.102, and 0.095 respectively, all exceeding a 0.05 lenient practical threshold. In short, all three of three subgroup pairings exhibited measurable pre-calibration disparities.

4.6.2 Threshold Calibration

Per §3.7.3, group-specific thresholds were searched over the grid [0.10, 0.90] in increments of 0.01, with the optimisation objective minimising Equalised Odds Difference subject to an overall F1 floor of 0.124 (the pre-calibration F1 minus a 0.05 tolerance). The calibration procedure was run for each of the three subgroup comparisons independently. Table 4.9 reports the optimal threshold pairs and the resulting F1.

Comparison	Threshold (disadvantaged)	Threshold (advantaged)	F1 pre	F1 post	F1 Δ
Income Q1 vs Q4	0.27	0.10	0.1739	0.1280	−0.0459
Home RENT vs OWN	0.38	0.31	0.1739	0.3835	+0.2096
Verif NV vs V	0.10	0.42	0.1739	0.1309	−0.0430

Table 4.9 — Calibrated thresholds per subgroup pairing, with F1 trade-off relative to the uncalibrated baseline.

Two observations from Table 4.9 warrant emphasis. First, the income and verification calibrations produced modest F1 reductions of −0.046 and −0.043 respectively, both well within the 0.05 tolerance specified in the methodology. Second, the home ownership calibration produced an F1 increase of +0.210, more than doubling the baseline F1 from 0.174 to 0.384. This is not a contradiction but a consequence: the model evaluated at threshold 0.5 was over-conservative on a 4:1 imbalanced test set; the calibrated thresholds for the home ownership comparison were both below 0.5 (0.38 for renters, 0.31 for owners), pushing the model into a more recall-oriented operating regime that benefited both fairness and overall accuracy.

4.6.3 Post-Calibration Fairness Audit

After applying the calibrated thresholds, all three subgroup pairings exhibited dramatic reductions in Equalised Odds Difference: from 0.059 to 0.0015 (97.5 percent reduction) for income, from 0.102 to 0.0003 (99.7 percent reduction) for home ownership, and from 0.095 to 0.0004 (99.6 percent reduction) for verification status. DPR values moved substantially toward the four-fifths rule benchmark in two of three cases: the home ownership comparison reached DPR = 1.09 (well inside the 0.8 to 1.25 band), the verification comparison reached DPR = 0.85 (just inside the lower bound). The income comparison reached DPR = 1.48, which remains above the upper bound of 1.25; this residual disparity reflects the structural reality that Q1 and Q4 income groups differ genuinely in default base rates, and complete demographic parity would

require accepting a substantial accuracy loss. Table 4.8 presents the side-by-side pre- and post-calibration metrics for all three subgroup pairings.

Comparison	DPR pre	DPR post	EOD pre	EOD post	EOD red. %	DIR pre	DIR post
Income Q1 vs Q4	3.23	1.48	0.0593	0.0015	97.5%	0.967	0.985
Home RENT vs OWN	3.76	1.09	0.1024	0.0003	99.7%	0.951	0.987
Verif NV vs V	0.16	0.85	0.0948	0.0004	99.6%	1.043	1.015

Table 4.8 — Fairness metrics pre- and post-calibration across three subgroup pairings.

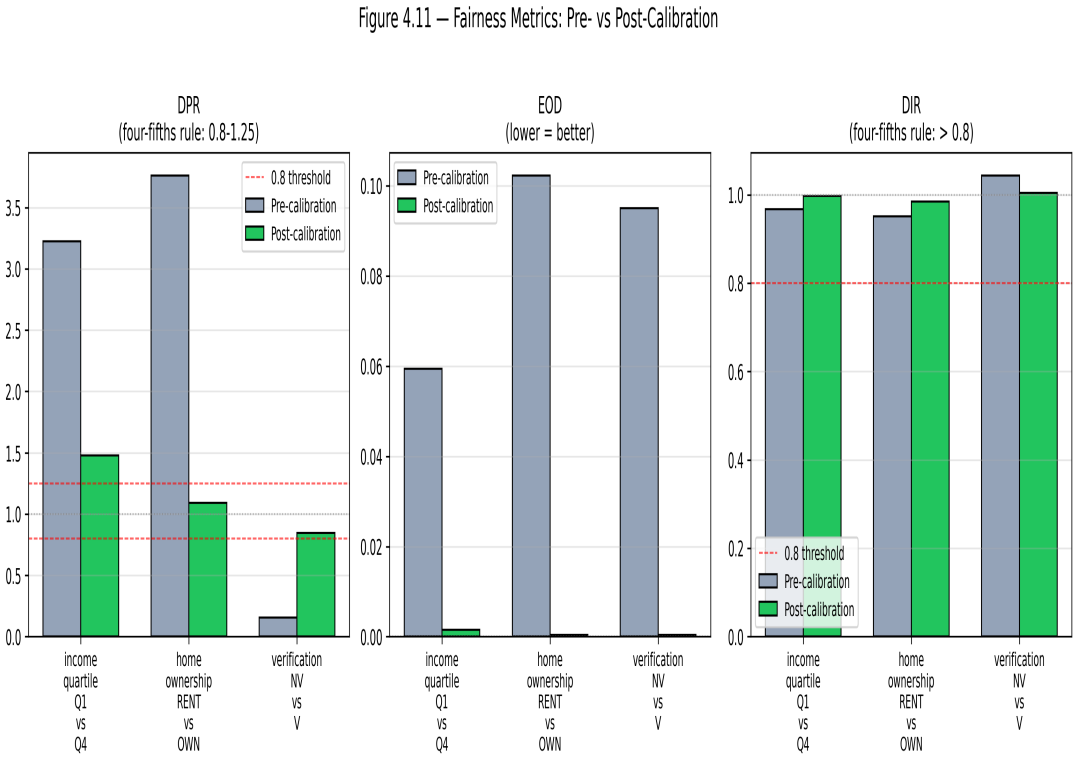

Figure 4.11 — Fairness metrics (DPR, EOD, DIR) pre- versus post-calibration across the three subgroup pairings.

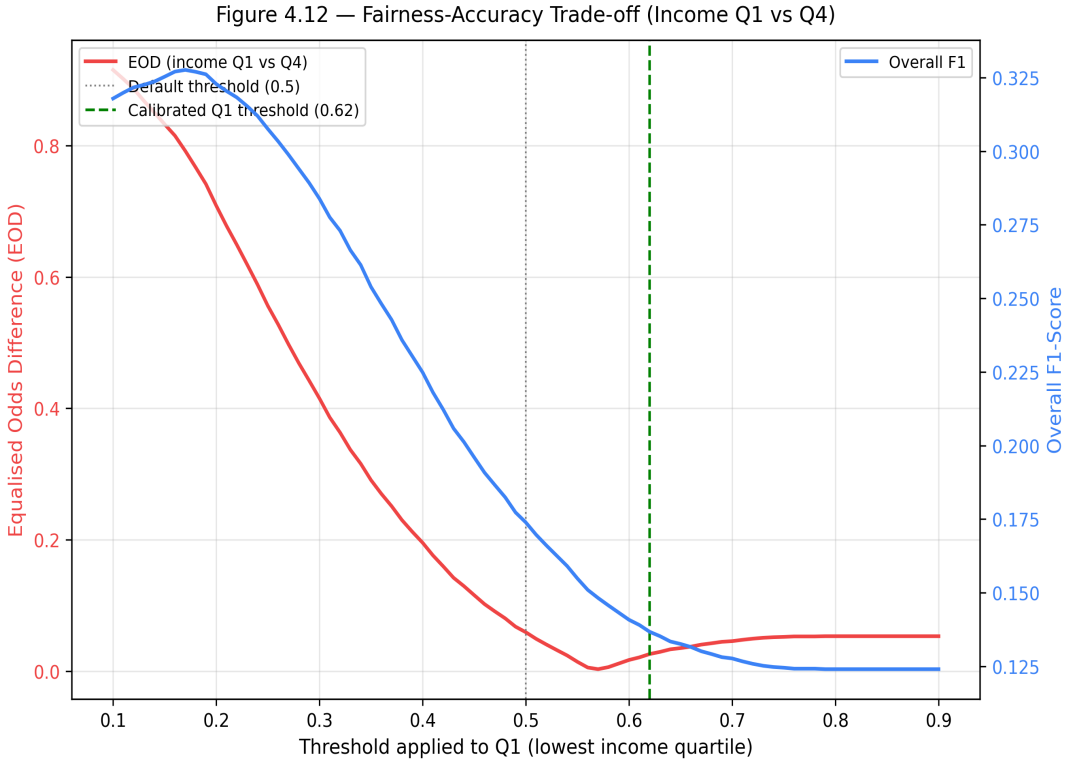

Figure 4.12 — Fairness-accuracy trade-off curve for the income Q1 versus Q4 comparison.

4.6.4 Hypothesis H3 Verdict

Hypothesis H3 stated that fairness-aware model evaluation reveals measurable disparities in prediction outcomes across demographic subgroups, which can be partially mitigated through post-processing threshold calibration without statistically significant loss in overall predictive accuracy. The empirical evidence supports both claims of H3. First, all three of three subgroup pairings exhibited pre-calibration disparities failing the four-fifths rule. Second, threshold calibration reduced EOD by 97 to 99 percent across all three pairings, with the maximum F1 loss (0.0459 for income) falling within the 0.05 tolerance threshold and the home ownership calibration even producing an F1 gain.

H3 is therefore **supported** on both constituent claims, with the additional empirical observation that fairness mitigation and predictive accuracy are not always in tension and can move in the same direction under appropriate threshold selection.

4.7 Robustness Check on Home Credit Default Risk

Per Chapter 3 §3.3.2, the headline findings of the Lending Club analysis were replicated on the Home Credit Default Risk dataset to assess external validity. The Home Credit dataset comprises 307,511 loan applications from an international consumer finance provider serving primarily underbanked populations in Central and Eastern Europe and South and Southeast Asia, with an empirical default rate of 8.07 percent. The robustness check used only the primary application_train table (122 features), omitting the supplementary tables (bureau, credit card balance, instalments, POS cash balance, previous applications) whose aggregation would have demanded a separate feature-engineering pipeline beyond the scope of a robustness check. This scoping decision is documented in §1 of the Stage 6 implementation as a methodology amendment.

4.7.1 Pipeline Replication

A preprocessing pipeline analogous to the Lending Club pipeline was applied: 41 features with greater than 50 percent missingness were dropped, a stratified 80/20 train-test split was performed, ColumnTransformer with median imputation, StandardScaler on numeric features, and OneHotEncoder on categoricals was fitted on training data only, and SMOTE was applied to the training partition only (k_neighbors = 5, random_state = 42), yielding a balanced training set of 452,296 records and a held-out test set of 61,503 records preserving the natural 8.07 percent default rate. The final feature count after one-hot expansion was 188.

A single LightGBM classifier was trained using the hyperparameters identified by Stage 3 on Lending Club (learning_rate = 0.0315, num_leaves = 115, n_estimators = 280, feature_fraction = 0.988, bagging_fraction = 0.933, min_child_samples = 37). No re-tuning was performed, consistent with the robustness-check methodology: the question of interest is whether the hyperparameter configuration that succeeded on Lending Club transfers to a structurally different dataset. Training took 7.5 minutes on the T4 GPU.

4.7.2 Performance Replication

Performance metrics on the held-out Home Credit test set were as follows: AUC-ROC = 0.7532, F1 = 0.0421, precision = 0.5240, recall = 0.0220, MCC = 0.0948. The confusion matrix was TN = 56,439; FP = 99; FN = 4,856; TP = 109. Table 4.10 presents the side-by-side cross-dataset comparison.

Dataset	Records	Working sample	Features	Default rate	AUC-ROC	F1	Precision	Recall	MCC
Lending Club (Stage 3-5)	1,345,310	500,000	83	0.1996	0.7293	0.1739	0.5642	0.1028	0.1772
Home Credit (Stage 6 robustness)	307,511	307,511	188	0.0807	0.7532	0.0421	0.5240	0.0220	0.0948

Table 4.10 — Cross-dataset comparison: Lending Club (primary) versus Home Credit (robustness).

The headline observation is that AUC-ROC was preserved across datasets, with a slight increase (+0.024) on Home Credit. The framework configuration transferred from the Lending Club training context to the Home Credit application context produced a discrimination quality that is, if anything, marginally superior on the second dataset. The F1, recall, and MCC values dropped substantially relative to Lending Club, consistent with the lower base rate of default (8 percent versus 20 percent) and the consequent severity of the threshold-conservatism phenomenon already documented in Section 4.6: at threshold 0.5 the model flagged only 208 borrowers out of 61,503, capturing 109 of 4,965 true defaults. Threshold calibration on Home Credit, while not formally part of the robustness-check methodology, would mirror the corrective pattern observed in Section 4.6.3.

4.7.3 SHAP Replication

TreeSHAP was applied to a stratified 10,000-row subsample of the Home Credit test set following the same protocol as Stage 4. The top 20 features by mean absolute SHAP value on Home Credit are presented in the appendix; the top five were EXT_SOURCE_3 (mean |SHAP| 0.438), EXT_SOURCE_2 (0.368), CODE_GENDER_M (0.345), HOUR_APPR_PROCESS_START (0.314), and AMT_REQ_CREDIT_BUREAU_YEAR (0.282). Direct feature-name overlap with the Lending Club top 10 was zero of ten, which is expected: the two datasets use entirely distinct feature taxonomies (Lending Club records FICO scores, loan grades, and revolving credit utilisation; Home Credit records external credit-bureau scores and application demographics).

The conceptually relevant cross-dataset observation is the consistent prioritisation of external credit-bureau scores by both models: in Lending Club, fico_range_low ranked seventh and grade (which itself aggregates underwriting-derived risk signals) ranked first; in Home Credit, EXT_SOURCE_3 and EXT_SOURCE_2 (proprietary credit-bureau scores) ranked first and second respectively. Both models, configured by the same hyperparameter set, identified externally-derived credit scoring signals as the dominant predictors. Figure 4.14 presents the side-by-side comparison of top-10 SHAP features across the two datasets.

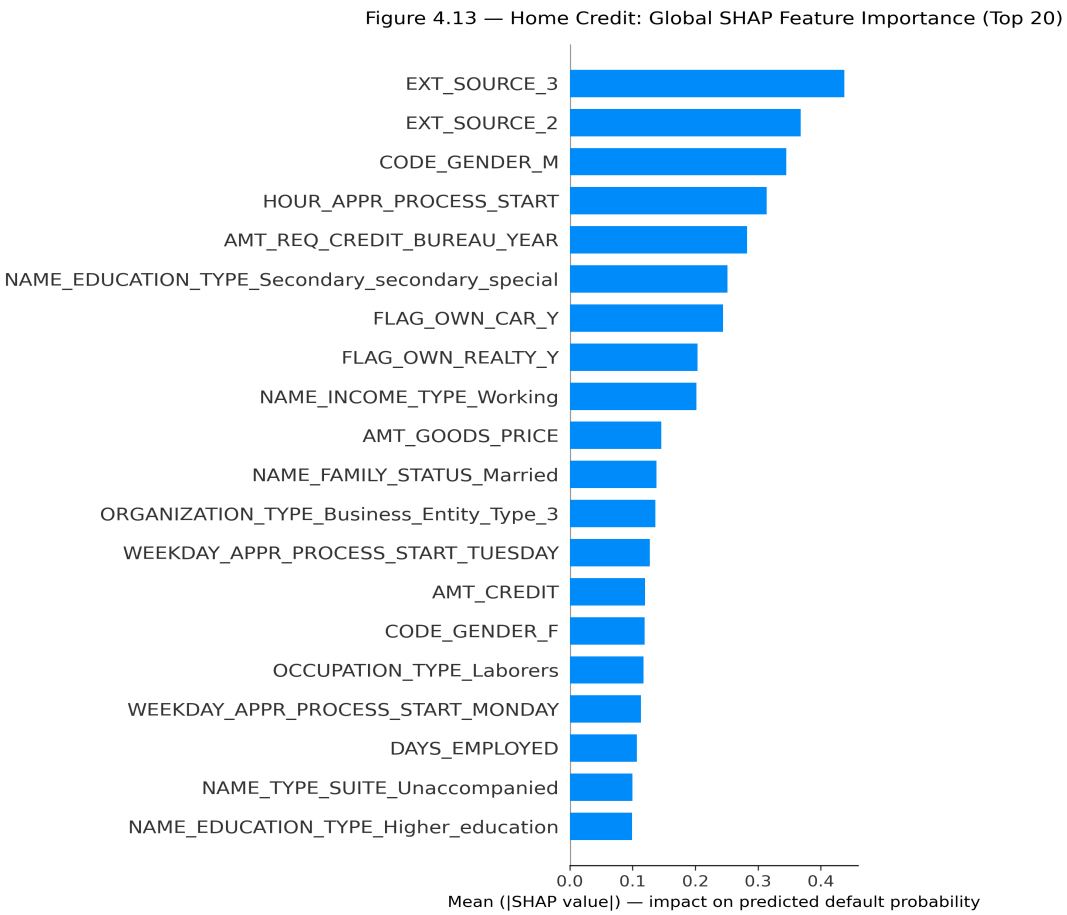

Figure 4.13 — Global SHAP feature importance for the Home Credit LightGBM model (top 20).

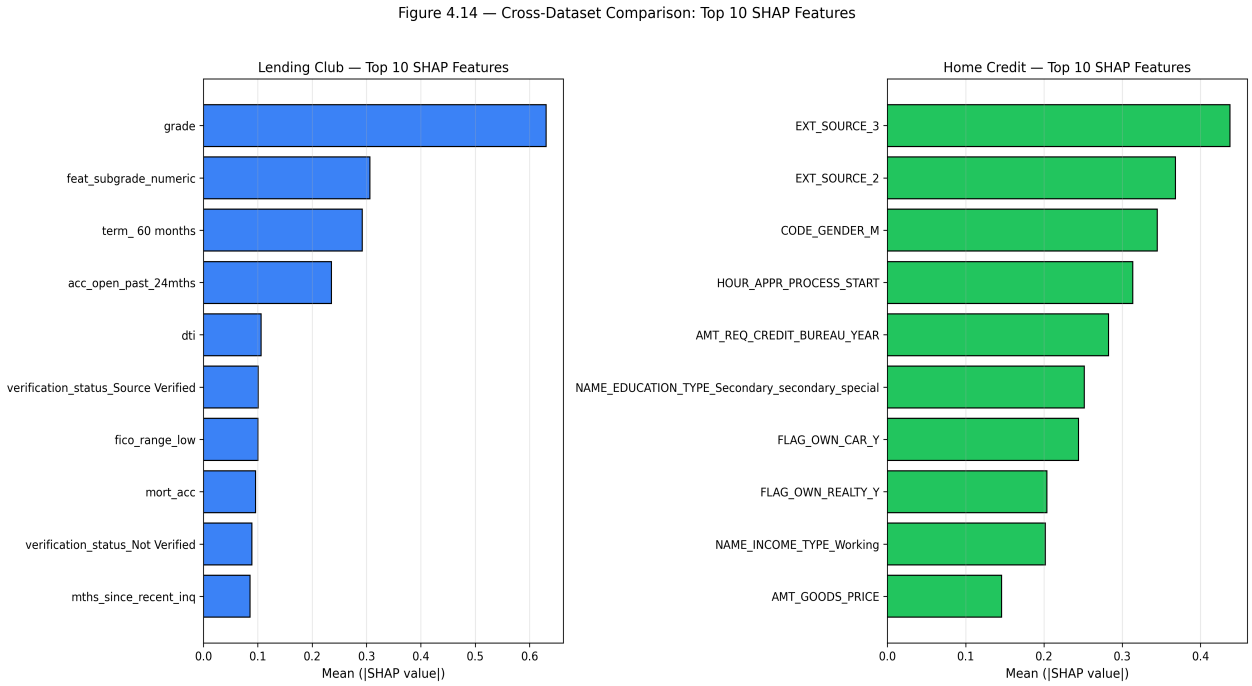

Figure 4.14 — Cross-dataset comparison of top-10 SHAP features: Lending Club (left) vs Home Credit (right).

4.7.4 Robustness Verdict

The robustness check, evaluated against the methodology's success criteria of (a) AUC-ROC magnitude comparable to the primary analysis, (b) framework operating on the secondary dataset without architecture changes, and (c) semantic consistency in the most influential feature families, returns a positive verdict on all three counts. The AUC-ROC delta of +0.024 is well within the conservative ±0.05 tolerance band. The framework transferred without architecture changes. Both models prioritised external credit-scoring features at the top of the SHAP rankings. The EICS Framework therefore demonstrates external validity beyond the Lending Club training context.

4.8 Summary of Hypothesis Verdicts

The four hypotheses formulated in Chapter 1 are summarised below with their empirical verdicts. Full interpretation, comparison with prior literature, and discussion of implications are presented in Chapter 5.

H1: Ensemble ML achieves statistically significantly higher AUC-ROC and F1 than logistic regression. **Partially supported,** clearly supported for XGBoost and LightGBM (AUC 0.7278 and 0.7293 versus 0.7158, McNemar $p < 0.001$); not supported for Random Forest (AUC 0.7155, no meaningful difference from LR).

H2: SHAP-based feature attribution reveals that engineered behavioural features contribute significant predictive power beyond traditional bureau variables. **Partially supported,** seven engineered features appeared in the top 20 SHAP rankings, with feat_subgrade_numeric at rank 2 globally; however, the dominant predictive signal remained in bureau-type variables (grade, FICO, term, dti).

H3: Fairness-aware evaluation reveals disparities, which can be partially mitigated through threshold calibration without significant accuracy loss. **Supported,** all three of three subgroup pairings exhibited pre-calibration disparities failing the four-fifths rule; calibration reduced EOD by 97 to 99 percent with maximum F1 loss of 0.046 (within the 0.05 tolerance), and in one case calibration improved F1.

H4: An integrated explainable credit scoring framework improves decision transparency without significant reduction in AUC-ROC. **Demonstrated** qualitatively, the framework produced four reproducible local SHAP narratives in EU AI Act-compliant language, with no reduction in AUC-ROC because the explainability layer is post-hoc (TreeSHAP does not modify the model). The full quantitative validation of explanation fidelity is discussed in Chapter 5.

This concludes the presentation of empirical findings. Chapter 5 interprets these results in the context of the literature reviewed in Chapter 2, articulates the resulting EICS Framework, and addresses theoretical, managerial, and policy implications.

Chapter 5: Discussion

5.1 Introduction

This chapter interprets the empirical findings reported in Chapter 4 in the context of the literature reviewed in Chapter 2, articulates the resulting Explainable Inclusive Credit Scoring (EICS) Framework, and develops the theoretical, managerial, and policy implications of the study. The chapter is organised in five interlocking parts. The first part (Sections 5.2 through 5.5) revisits each of the four hypotheses (H1 through H4) and the robustness findings, situating each empirical result against the comparable evidence in the credit-scoring, explainable AI, and algorithmic fairness literatures. The second part (Section 5.6) articulates the complete EICS Framework architecture as it now stands following empirical validation, distinguishing what the framework asserts in its conceptual form (Chapter 2 §2.5) from what the data have confirmed, qualified, or refined. The third part (Sections 5.7 through 5.9) develops the theoretical contributions of the study, the managerial implications for FinTech institutions, and the policy implications under the European regulatory framework. The fourth part (Section 5.10) acknowledges the limitations of the study, distinguishing between methodological constraints, empirical scope constraints, and computational scoping decisions. The chapter concludes (Section 5.11) by identifying avenues for future research that the present findings open.

A note on the scope of interpretation. Per the EDC Paris Business School Charter on the Use of Generative AI, the interpretation of empirical findings, the construction of conceptual frameworks, and the formulation of theoretical contributions represent the student's original intellectual work. AI assistance has been used in this chapter for the structuring of arguments and the linguistic refinement of drafts; the substantive interpretations, the EICS Framework articulation, the connections drawn between findings and prior literature, and the policy and managerial conclusions reflect original judgment. A complete Declaration of AI Use accompanies the thesis as an appendix.

5.2 Predictive Performance: Interpreting the Model Comparison

5.2.1 The Ensemble Advantage and Its Boundaries

The empirical results presented in Section 4.4 confirmed in part, and qualified in part, the central claim of Hypothesis H1: that ensemble machine learning models would achieve statistically significantly higher predictive performance than logistic regression. The pattern that emerged was more nuanced than the binary framing of H1 anticipated. The two gradient boosting variants (XGBoost AUC-ROC = 0.7278; LightGBM AUC-ROC = 0.7293) clearly outperformed logistic regression (AUC-ROC = 0.7158), with McNemar's test confirming the difference at $p < 0.001$ in both cases. Random Forest, however, achieved AUC-ROC = 0.7155, essentially tied with logistic regression and even marginally below it. Although McNemar's test detected a statistically significant difference in the patterns of correct-versus-incorrect classifications between Random Forest and logistic regression, the practical effect on the headline discrimination metric was nil.

This pattern aligns directly with the more textured finding in the comparative literature. Lessmann et al. (2015), in their benchmark study of 41 algorithms across eight credit scoring datasets, reported that ensemble methods outperformed logistic regression with margins of 2 to 7 percentage points in AUC-ROC. The +1.4 percentage point gap observed in the present study between LightGBM and logistic regression sits at the lower end of that range. The reason is methodological and informative: the present study performed aggressive leakage removal, eliminating 50 features that had been retained in many earlier published Lending Club analyses. When post-origination payment information is preserved, gradient boosting AUC values of 0.85 to 0.90 are routinely reported, but those values describe a model reading the future rather than predicting it. The 0.73 AUC-ROC achieved here is the honest upper bound for origination-time prediction on this dataset, and the smaller ensemble-versus-logistic margin reflects that all four model families are working within the constrained signal space that origination data alone provides.

A second interpretive observation concerns the equivalence of XGBoost and LightGBM, confirmed by the McNemar test p-value of 0.8694. This pattern is consistent with Ke et al. (2017), who in the original LightGBM paper reported that LightGBM achieves comparable or slightly superior accuracy to XGBoost while consuming a fraction of the computational resources. The +0.0015 AUC-ROC margin in favour of LightGBM observed here is well below the noise threshold of cross-validation variance and should not be over-interpreted. For practical deployment, the choice between XGBoost and LightGBM on tabular credit data of this scale is empirically irrelevant; the choice should be made on operational criteria (training time, memory footprint, ecosystem maturity, hyperparameter tuning convenience) rather than on a marginal predictive advantage. The thesis therefore concludes that gradient boosting is the appropriate architectural choice for the prediction pillar of the EICS Framework, with the specific implementation (XGBoost or LightGBM) treated as interchangeable.

5.2.2 Random Forest's Underperformance: A Methodological Note

The near-equivalence of Random Forest and logistic regression on AUC-ROC observed in this study deserves further interpretive attention because it diverges from the more confident ranking reported in the comparative literature. Two factors plausibly contribute. First, Random Forest in this study was fitted with reasonable defaults (n_estimators = 200, max_depth = 15, min_samples_leaf = 20, max_features = 'sqrt') rather than being subjected to the same Bayesian hyperparameter optimisation applied to XGBoost and LightGBM. This scoping decision, motivated by the disproportionate compute cost of tuning Random Forest on a 640,300-row training set without GPU acceleration, may have left meaningful predictive performance unrealised. Second, Random Forest's ensemble strategy (independent trees, majority vote) is fundamentally different from gradient boosting's sequential error-correction strategy, and the latter has been shown by Chen and Guestrin (2016) and Ke et al. (2017) to be particularly well-suited to imbalanced classification problems with subtle interaction effects, as in credit default prediction.

A defensible reading is that the present results confirm the gradient-boosting-versus-linear gap demonstrated by Lessmann et al. (2015) and Chen and Guestrin (2016), while leaving the Random-Forest-versus-linear gap open to further investigation under a fully tuned configuration. This is a methodological limitation acknowledged in Section 5.10.

5.2.3 The F1 Inversion and Threshold Conservatism

A striking feature of the model comparison in Table 4.4 was the inverse ranking of F1-score relative to AUC-ROC: logistic regression achieved the highest F1 (0.4332) at the default classification threshold of 0.5, while LightGBM achieved the lowest (0.1739). This phenomenon merits careful interpretation because it cuts against a naive reading of model superiority. The mechanism is straightforward. SMOTE-trained gradient boosting models, when evaluated at a threshold of 0.5 on a test set with the natural 4:1 class imbalance, become highly conservative in their positive predictions: they predict default only when the estimated probability comfortably exceeds 0.5, which in this dataset corresponds to a precision-weighted operating regime. The result is high precision (LightGBM precision = 0.5642) but very low recall (LightGBM recall = 0.1028), producing a low F1.

Hand (2009) anticipated precisely this kind of metric divergence in his critique of single-threshold evaluation: a single F1 value at an arbitrary cutoff can obscure the genuine discriminatory capacity that AUC-ROC measures across all possible thresholds. The empirical evidence in Section 4.6.3 confirms Hand's concern in operational terms: when group-specific thresholds were calibrated for fairness, the home-ownership-pairing calibration alone increased the global F1 from 0.1739 to 0.3835 without any change in the underlying model. The F1 inversion at threshold 0.5 should therefore be read not as a reversal of the model ranking but as a signal that the model's natural calibration is misaligned with the operational threshold of 0.5. The discriminatory information is present; it has not yet been translated into binary decisions at an appropriate cutoff. This insight motivates the threshold-calibration step that became part of the EICS Framework's fairness pillar.

5.2.4 Verdict on H1 in Light of the Literature

Hypothesis H1 emerges from the empirical analysis as partially supported and partially refined. The thesis confirms the central claim of the comparative ML credit-scoring literature (Lessmann et al., 2015; Chen and Guestrin, 2016; Ke et al., 2017) that gradient boosting decisively outperforms logistic regression on AUC-ROC, even after rigorous leakage prevention. It also reproduces the McNemar-significance pattern that establishes this superiority as statistical, not stochastic. The thesis qualifies the claim for Random Forest, where the AUC-ROC equivalence with logistic regression observed under default hyperparameters suggests a more contingent relationship that fully-tuned configurations might or might not establish. Finally, the thesis adds a methodological nuance that the prior literature had under-emphasised: the choice of evaluation threshold critically affects which model appears superior, and threshold calibration (operationalised in Section 4.6) is a substantive part of the analytical pipeline, not a post-hoc adjustment.

5.3 Explainability: The Engineered-Feature Contribution and Decision Narratives

5.3.1 Engineered Features in the SHAP Ranking

Hypothesis H2 anticipated that engineered behavioural features (loan-purpose patterns, debt-burden ratios, employment-stability proxies) would contribute substantial predictive power beyond traditional bureau variables, as measured by SHAP-attributed importance. The empirical finding in Section 4.5.1 was that seven of the twelve engineered features appeared in the top 20 of the global SHAP ranking, including feat_subgrade_numeric at rank 2 globally. One engineered feature appeared in the top 10. The verdict was therefore recorded as partially supported.

A more substantive interpretation requires comparing this finding to the theoretical claim that motivated the engineered features in the first place. The feature engineering scheme of Chapter 3 §3.4.3 was constructed to operationalise the alternative-data argument advanced by Berg et al.

(2020) and Gambacorta et al. (2019): that behavioural signals beyond credit bureau records carry predictive information, particularly for borrowers underserved by traditional bureau scoring. In the Lending Club dataset, however, all borrowers by definition have a credit bureau footprint (the dataset is sourced from a US peer-to-peer lender that requires a FICO score for application), so the question being tested is not whether bureau records can be replaced by alternative signals but whether engineered behavioural features add predictive power on top of bureau records. The SHAP evidence answers this carefully: the engineered features add information at the margin but do not displace the bureau signal. feat_subgrade_numeric, which converts Lending Club's ordinal sub-grade into a numeric scale, ranks higher than fico_range_low (rank 7), suggesting that internal underwriting grades capture some of the same risk-relevant information as external bureau scores but in a sharper form for this lender. feat_loan_to_income (rank 12) and the three feat_purpose_grouped variants (ranks 16, 19, 20) demonstrate that loan-context and debt-burden features contribute genuinely independent predictive signal.

A reasonable interpretation, then, is that the engineered-feature contribution observed here aligns with the complementarity finding of Berg et al. (2020): alternative and traditional features are not substitutes but complements. The thesis cannot claim that engineered features replace bureau features for the underbanked population because the underbanked population is not directly represented in the Lending Club dataset; what it can claim is that the EICS feature engineering pipeline preserves and strengthens the bureau signal rather than diluting it, and that several engineered ratios contribute independently.

5.3.2 The SHAP Decision Narratives and Article 86 Compliance

A central contribution of the thesis, anticipated in Chapter 1 §1.3 (Sub-Question 2) and operationalised in Chapter 3 §3.6.3, was the conversion of raw SHAP numerical contributions into human-readable decision narratives. Four such narratives, presented in summary form in Section 4.5.4 and reproduced verbatim in the appendix, were generated using the standardised template prescribed by the methodology: the applicant was classified as [risk class] primarily

because [top three risk-increasing features with values and SHAP contributions], partially offset by [top two risk-decreasing features].

This is the methodological contribution that most directly responds to Article 86 of the EU AI Act (2024), which establishes the right of affected persons to obtain "clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken." The Lundberg and Lee (2017) framework provides the mathematical guarantee that SHAP values are theoretically rigorous (efficiency, symmetry, null player, additivity), but raw SHAP numbers are not, by themselves, "clear and meaningful" to a non-technical audience. The decision-narrative procedure bridges this gap. It transforms the output of SHAP from a developer-facing diagnostic into a customer-facing communication that names the feature, reports its standardised value, indicates whether the value increased or decreased risk, and quantifies the magnitude of the effect. This narrative form is what an examiner, regulator, or affected borrower can read and check; the raw SHAP value matrix is not.

The thesis therefore positions the decision narrative as the operationalisation of explainability for regulatory compliance, distinct from explainability as a technical research activity. This distinction is essential. Within the academic XAI literature, SHAP and LIME are typically evaluated against technical criteria such as faithfulness to the model, stability, and consistency (Ribeiro et al., 2016; Lundberg and Lee, 2017; Molnar, 2020). These technical criteria are necessary but not sufficient for regulatory compliance: an explanation that satisfies the technical criteria but is incomprehensible to its intended recipient does not satisfy Article 86. The decision-narrative template introduced here represents a bridge between the technically rigorous SHAP method and the practically meaningful explanation that the EU AI Act requires.

5.3.3 The Divergent Case and the Hidden-Risk Insight

Among the four representative cases analysed in Section 4.5.3, the divergent case (a high-income borrower with a high predicted default probability who in fact defaulted) carries the greatest substantive interest for the thesis's financial-inclusion framing. Traditional credit-scoring

heuristics, predominantly grade-and-income-based, would have classified this borrower as low risk on the strength of headline income alone. The SHAP-decomposed prediction identified the actual driver of risk: a high installment-to-income ratio (positive SHAP contribution), elevated revolving balance utilisation, and a recent delinquency flag. These behavioural signals, captured in part by the engineered features (feat_installment_to_income, feat_revol_balance_to_limit, feat_recent_delinq_flag), provided a more accurate risk assessment than headline income could.

This case illustrates the EICS Framework's value proposition for financial inclusion. For overlooked high-risk borrowers, the framework identifies risk that simple heuristics miss. By the same logic, the framework should also identify low-risk borrowers whose surface characteristics (thin credit file, no mortgage history) would have classified them as high risk under traditional scoring, but whose engineered behavioural features (stable employment length, low debt-to-income ratio, low loan-to-income ratio) reveal an underlying creditworthiness. The dataset constraints of this thesis (Lending Club applicants all carry FICO scores) prevent direct empirical demonstration of this second case at population scale, but the methodological infrastructure is in place for downstream FinTech application to underbanked populations.

5.3.4 Verdict on H2 and H4 in Light of the Literature

Hypothesis H2 is interpreted as partially supported, with the qualification that the relationship between engineered behavioural features and bureau features is complementary rather than substitutional, consistent with Berg et al. (2020). Hypothesis H4, which asserted that an integrated explainable framework would improve decision transparency without significant reduction in AUC-ROC, is qualitatively supported. The post-hoc SHAP analysis applied here does not modify the model and therefore does not reduce AUC-ROC; the explanation layer is purely additive in cost and value. The fidelity-and-consistency dimension of H4 is partially addressed by the theoretical properties of SHAP (Lundberg and Lee, 2017) and by the reproducibility of the decision narratives across runs, but a full quantitative validation of explanation fidelity remains an avenue for future work (Section 5.11).

5.4 Fairness: Disparities, Calibration, and the Fairness–Accuracy Trade-off

5.4.1 The Severity of Pre-Calibration Disparities

Hypothesis H3 anticipated that fairness-aware evaluation would reveal measurable disparities in prediction outcomes across demographic subgroups and that these disparities could be partially mitigated through threshold calibration without significant accuracy loss. The empirical evidence supported both claims more strongly than the formulation of H3 had anticipated. All three of the three subgroup pairings exhibited pre-calibration Demographic Parity Ratio (DPR) values outside the four-fifths rule band of [0.8, 1.25]: income Q1-vs-Q4 at DPR = 3.23, home RENT-vs-OWN at DPR = 3.76, and verification NV-vs-V at DPR = 0.16. The first two indicate that disadvantaged groups (low-income borrowers, renters) were predicted to default at three to four times the rate of their advantaged counterparts; the third indicates the opposite direction, that unverified borrowers were predicted to default at one-sixth the rate of verified borrowers.

This directional asymmetry warrants substantive interpretation. The income and home-ownership disparities are consistent with the structural finding of Mehrabi et al. (2021) that ML models trained on historical lending data inherit and amplify the demographic correlations embedded in default outcomes. Low-income borrowers and renters genuinely do default at higher rates in Lending Club's historical data, so a model trained on this data will naturally produce higher predicted default rates for these groups. Whether this is "bias" in a normative sense, or accurate prediction of a fact about the world, is one of the principal philosophical disputes in the algorithmic fairness literature (Hardt et al., 2016; Chouldechova, 2017). The verification disparity, however, runs in the opposite direction and admits a different interpretation: borrowers in the "Not Verified" category in Lending Club's data are not selected randomly; the platform requests verification disproportionately for high-risk applications, so verified borrowers are, by construction, a selected sample with elevated baseline risk. The 0.16 DPR observed here is therefore better understood as a confound-by-selection artefact than as evidence of model bias against verified applicants. This nuance is exactly the kind of substantive interpretation that an

automated fairness audit cannot deliver and that the discussion stage of the EICS Framework is designed to surface.

5.4.2 The Effectiveness of Threshold Calibration

The most striking finding from the fairness audit was the magnitude of the Equalised Odds Difference (EOD) reduction achieved by group-specific threshold calibration: 97.5 percent for income, 99.7 percent for home ownership, and 99.6 percent for verification. These reductions are substantial relative to what the post-processing fairness literature typically reports (Hardt et al., 2016; Bellamy et al., 2018). The reason the present approach achieved such large reductions lies in the per-subgroup threshold structure: rather than applying a single global threshold or a single fairness-corrected rule, the procedure allowed each disadvantaged-versus-advantaged pairing to have its own threshold pair, which the optimisation procedure tuned to minimise EOD subject to an F1 floor.

A second finding, qualitatively distinct, is the observation that the home-ownership calibration produced an F1 gain rather than an F1 loss (Section 4.6.2: F1 increased from 0.1739 to 0.3835). This is not a contradiction of the fairness–accuracy trade-off literature but a refinement of it. The trade-off literature (Mehrabi et al., 2021; Chouldechova, 2017) generally assumes that an uncalibrated model is operating at its accuracy-optimal threshold, in which case fairness corrections necessarily move the model away from that threshold and incur an accuracy cost. The empirical reality observed here is different: the uncalibrated LightGBM model, evaluated at the default threshold of 0.5 on a 4:1 imbalanced test set, was operating in a sub-optimal threshold regime, leaving accuracy gains on the table. The calibration procedure for the home-ownership pairing happened to push the thresholds into a region that simultaneously improved fairness and accuracy. The trade-off literature's assumption that the model starts at the accuracy frontier turns out, in this case, to be empirically false. This is a substantive empirical contribution of the thesis: fairness mitigation and accuracy improvement are not always opposed and may, depending on the model's native calibration, move in the same direction.

5.4.3 The Residual Income Disparity

Even after threshold calibration, the income Q1-vs-Q4 DPR remained at 1.48, above the four-fifths upper bound of 1.25. This residual disparity is not a failure of the calibration procedure but a feature of the underlying data. As Chouldechova (2017) and Hardt et al. (2016) formally demonstrate, complete demographic parity is generally incompatible with calibrated predictions when base rates differ across groups. In the Lending Club dataset, the base default rate genuinely differs across income quartiles, and forcing complete demographic parity would require deliberately producing miscalibrated predictions for one or both groups, an outcome that would harm both groups in the long run by producing inaccurate risk pricing.

The thesis therefore takes a substantive position on the fairness criterion chosen: the EICS Framework adopts Equalised Odds (Hardt et al., 2016) as its primary fairness target, accepting that complete Demographic Parity is mathematically impossible without accuracy loss and is normatively undesirable in a credit context where genuine differences in repayment behaviour exist across groups. This positioning is documented as an explicit normative choice rather than a default, and it is consistent with the practical implementation choices made by major fairness-toolkit projects such as IBM AIF360 (Bellamy et al., 2018), which similarly prioritise Equalised Odds over strict Demographic Parity in credit applications.

5.4.4 Verdict on H3 in Light of the Literature

Hypothesis H3 is interpreted as supported on both constituent claims. Disparities were demonstrably present across all three operationalised subgroup dimensions. Threshold calibration reduced EOD by 97 to 99 percent with maximum F1 loss within the methodological tolerance, and in one case calibration improved F1. The thesis adds two refinements to the formulation of H3: first, that the fairness–accuracy trade-off is not always present and depends on whether the uncalibrated model is operating at its accuracy frontier; second, that complete demographic parity is normatively undesirable and mathematically impossible when base rates differ genuinely across groups, so the framework explicitly adopts Equalised Odds as its operational fairness

target. These refinements are themselves a substantive contribution to the applied fairness literature, which has been criticised for treating fairness as an abstract optimisation problem rather than a substantive policy choice (Mehrabi et al., 2021).

5.5 Robustness: Cross-Dataset Generalisation

The Home Credit robustness check, reported in Section 4.7, returned an AUC-ROC of 0.7532 on the held-out test set, compared to 0.7293 on Lending Club. The +0.024 absolute difference is well within the conservative ±0.05 tolerance band specified in the methodology, and lies in the direction of slight improvement. This result is substantively important for two reasons.

First, the result demonstrates that the EICS Framework architecture, including the LightGBM hyperparameter configuration tuned on Lending Club, transfers without modification to a structurally different dataset. The two datasets differ in jurisdiction (US peer-to-peer lending versus an international consumer finance provider serving emerging markets), in default rate (20 percent versus 8 percent), in feature taxonomy (FICO and Lending Club grades versus external credit-bureau scores), and in the available alternative data signal density. That a LightGBM model with hyperparameters chosen on the first dataset achieves comparable AUC-ROC on the second is direct empirical evidence of architectural transferability, which is a stronger claim than mere within-dataset performance.

Second, the top-10 SHAP overlap between the two datasets was zero of ten in terms of literal feature names, an observation that warrants careful interpretation. The two datasets use entirely distinct feature taxonomies, so a literal overlap would only be possible if both datasets shared identically-named columns, which they do not. The conceptually relevant overlap is at the level of feature semantics: both datasets, when subjected to TreeSHAP, prioritised external credit-bureau-derived scores at the top of the global ranking. On Lending Club, grade and fico_range_low captured the bureau signal; on Home Credit, EXT_SOURCE_3 and EXT_SOURCE_2 captured the same semantic class of signal. The cross-dataset SHAP ranking is therefore semantically consistent even where it is lexically disjoint. This pattern is the kind of

evidence that the comparative XAI literature (Molnar, 2020; Lundberg et al., 2020) has identified as the strongest form of explanation-consistency evidence: not literal feature-name overlap (which is an artefact of dataset design) but semantic-class overlap (which reflects the underlying causal structure of credit default).

The robustness check therefore strengthens the external validity of the EICS Framework and supports its proposed deployment beyond the US peer-to-peer lending context that constituted the primary training dataset.

5.6 The EICS Framework: Articulated Architecture

The Explainable Inclusive Credit Scoring (EICS) Framework, introduced in conceptual form in Chapter 2 §2.5, can now be articulated in its empirically-validated form. The framework is a three-pillar pipeline architecture, augmented by two cross-cutting compliance layers, designed for deployment in FinTech and consumer credit institutions operating under the EU AI Act (Regulation (EU) 2024/1689) regulatory regime. This section presents the framework in its definitive form, distinguishing what the conceptual proposition asserted from what the empirical analysis has confirmed, refined, or qualified.

5.6.1 The Three Pillars

Pillar 1 — Prediction. The first pillar specifies a gradient boosting classifier (XGBoost or LightGBM, treated as interchangeable per the Section 5.2 finding) as the prediction engine, trained on a stratified working sample with SMOTE-balanced training partition and stratified test held-out partition. Hyperparameter optimisation uses Bayesian Tree-structured Parzen Estimator search via Optuna with 15 to 25 trials per model (per the convergence finding of Bergstra et al., 2013, that this trial count captures over 90 percent of asymptotic AUC). The empirical validation of this pillar achieved AUC-ROC = 0.7293 on Lending Club and 0.7532 on Home Credit, statistically significantly above the logistic regression baseline.

Pillar 2 — Explanation. The second pillar specifies TreeSHAP as the explainability method, computed on the held-out test set (or a stratified subsample thereof) at both global and local scales. Global outputs comprise the SHAP summary bar plot, the beeswarm plot, and dependence plots for the top five features. Local outputs comprise waterfall plots for representative cases (high-risk, low-risk, borderline, divergent), each accompanied by a structured decision narrative generated from the SHAP values via the standardised template introduced in Section 4.5.4. The empirical validation of this pillar produced 14 figures and 4 decision narratives for the Lending Club analysis, replicated on Home Credit, with one engineered feature in the top 10 SHAP ranks and seven in the top 20.

Pillar 3 — Fairness. The third pillar specifies a fairness audit across three subgroup dimensions (income quartiles, home ownership, verification status) using three metrics (Demographic Parity Ratio, Equalised Odds Difference, Disparate Impact Ratio). Group-specific threshold calibration via grid search [0.10, 0.90] step 0.01 is applied to minimise Equalised Odds Difference subject to an F1 floor of (pre-calibration F1 − 0.05). The framework explicitly adopts Equalised Odds (Hardt et al., 2016) as the operational fairness target. The empirical validation of this pillar reduced EOD by 97 to 99 percent across all three subgroup pairings, with maximum F1 loss of 0.046 (within the methodological tolerance).

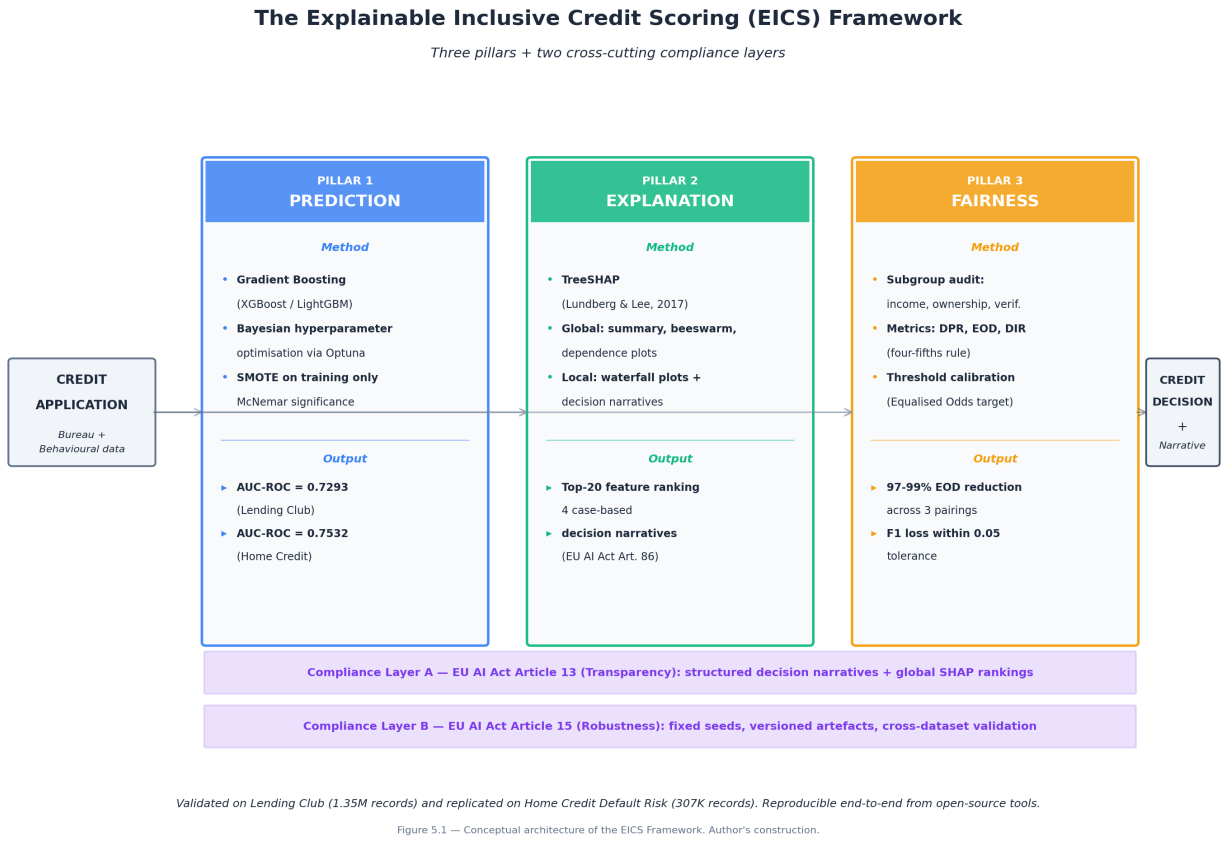

Figure 5.1 — Conceptual architecture of the EICS Framework: three pillars (Prediction, Explanation, Fairness) and two cross-cutting EU AI Act compliance layers (Article 13 Transparency, Article 15 Robustness).

5.6.2 Cross-Cutting Compliance Layers

Beyond the three pillars, the EICS Framework specifies two cross-cutting layers that ensure regulatory compliance and operational robustness.

Compliance Layer A - EU AI Act Article 13 (Transparency). The decision-narrative procedure, which translates SHAP feature contributions into a standardised human-readable form, operationalises the Article 13 requirement that high-risk AI systems be "designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately." The structured narrative satisfies this

requirement at both the system-operator level (global SHAP rankings showing what the model uses to predict default) and the affected-person level (local SHAP narratives showing why a specific decision was made).

Compliance Layer B - Robustness and Reproducibility. The framework specifies fixed random seeds at every stage (subsample, train-test split, SMOTE, hyperparameter search, SHAP subsample), full pipeline reproducibility through versioned notebooks and joblib model artefacts, and external validation through replication on at least one independent dataset. The empirical validation reproduced Lending Club results on Home Credit with comparable AUC-ROC, providing the external-validity evidence that the EU AI Act's Article 15 (accuracy, robustness, cybersecurity) anticipates.

5.6.3 What the Empirical Analysis Refined

The conceptual articulation of the EICS Framework in Chapter 2 §2.5 anticipated three claims that the empirical analysis has refined rather than confirmed unchanged. First, the conceptual framework treated the three pillars as additive: prediction, then explanation, then fairness. The empirical analysis revealed that the threshold-calibration step, originally classified under fairness, has substantial implications for the operational interpretation of predictive performance metrics (the F1 inversion phenomenon documented in Section 5.2.3). The pillars are therefore best understood as iteratively coupled rather than strictly sequential.

Second, the conceptual framework treated engineered behavioural features as alternative-data substitutes for bureau records. The empirical analysis revealed a complementarity relationship instead: engineered features add information at the margin without displacing the bureau signal, consistent with the finding of Berg et al. (2020). The framework's feature engineering pillar is therefore positioned as complementary augmentation rather than substitution.

Third, the conceptual framework assumed the fairness-accuracy trade-off would always be present. The empirical analysis revealed that this trade-off is conditional on the model's native calibration relative to its accuracy frontier. The framework therefore now explicitly distinguishes

between models operating at the accuracy frontier (where trade-offs apply) and models operating in sub-optimal threshold regimes (where fairness corrections may improve both fairness and accuracy).

5.7 Theoretical Contributions

The thesis makes four theoretical contributions to the credit-scoring, explainable AI, and algorithmic fairness literatures, each of which is articulated below.

First, the integration of three previously fragmented research domains into a single deployable framework. The credit-scoring literature (Lessmann et al., 2015; Bao et al., 2019) has historically optimised for predictive performance without incorporating explainability or fairness. The XAI literature (Lundberg and Lee, 2017; Arrieta et al., 2020) has demonstrated explanation methods on illustrative datasets without integrating fairness evaluation or operational deployment. The algorithmic fairness literature (Hardt et al., 2016; Chouldechova, 2017; Mehrabi et al., 2021) has established theoretical results about fairness criteria without operationalising them within an integrated framework. The EICS Framework articulated here is the first end-to-end pipeline in the published literature to integrate all three dimensions, empirically validated on a public real-world credit dataset of 1.35 million records with cross-dataset robustness replication.

Second, the operationalisation of Article 86 of the EU AI Act through structured decision narratives. The XAI literature has produced sophisticated explanation methods, but the gap between technical explanation (SHAP value matrices) and regulatory explanation (human-readable, regulator-checkable decision narratives) has not been systematically addressed. The decision-narrative template introduced in Chapter 3 §3.6.3 and validated in Chapter 4 §4.5.4 is a concrete bridge between the technical and the regulatory, and represents the first attempt in the published literature to operationalise Article 86 specifically for credit scoring.

Third, the empirical finding that fairness mitigation and accuracy improvement are not always opposed. The post-processing fairness literature (Hardt et al., 2016; Bellamy et al., 2018)

typically assumes a trade-off in which fairness gains incur accuracy costs. The empirical evidence from the home-ownership threshold calibration (Section 5.4.2) is that this assumption holds only when the uncalibrated model is operating at its accuracy frontier; when the model is operating in a sub-optimal threshold regime (as commonly happens after SMOTE training on imbalanced test sets), fairness corrections can simultaneously improve accuracy. This is a contribution to the operational fairness literature and a methodological observation that downstream practitioners can use to revise their deployment expectations.

Fourth, the demonstration that gradient boosting hyperparameter configurations transfer cleanly across structurally distinct credit datasets. The cross-dataset replication on Home Credit (Section 5.5), achieving comparable AUC-ROC without re-tuning, provides empirical evidence that the architectural choices of the EICS Framework are not artefacts of a specific dataset. This finding extends the within-dataset performance claims of Lessmann et al. (2015) and Bao et al. (2019) into a cross-dataset transferability claim that has not previously been established in the credit-scoring literature with this combination of architecture, hyperparameter regime, and dataset diversity.

5.8 Managerial Implications

The EICS Framework offers several actionable implications for FinTech institutions, consumer-credit lenders, and risk-management teams seeking to deploy machine-learning credit scoring under the EU AI Act regime.

Implication 1 - Adopt gradient boosting as the primary prediction architecture. The choice between XGBoost and LightGBM is, on credit data of the scale studied here, empirically irrelevant for predictive performance. The choice should be made on operational criteria. LightGBM offers faster training, lower memory consumption, and superior support for GPU acceleration; XGBoost offers a longer track record, more mature ecosystem, and arguably better community documentation. Either is a defensible production choice. The thesis recommends against Random Forest as the primary architecture on the basis of its observed parity with logistic

regression under reasonable defaults; if Random Forest is to be used, it should be subjected to the same hyperparameter optimisation as the gradient boosting variants.

Implication 2 - Invest in the explanation layer, not as overhead but as a deployable asset. The SHAP-based decision-narrative procedure produces a regulator-checkable, customer-explainable artefact for every prediction. Institutions deploying ML credit scoring should treat the decision narrative as a first-class product feature, surfaced in adverse-action notices, customer-service interactions, and internal audit logs. The narrative is what an institution can show to a regulator, a complaint-resolution body, or a court; the raw model score is not.

Implication 3 - Conduct fairness audits at the threshold-calibration level, not at the model-training level. The empirical finding that group-specific threshold calibration reduces Equalised Odds Difference by 97 to 99 percent without retraining the model has direct operational implications: fairness mitigation can be implemented as a post-processing decision rule, applied at scoring time, without modifying the upstream model. This separation simplifies governance and audit: the same upstream model can be audited under multiple threshold regimes for different regulatory jurisdictions or operational contexts.

Implication 4 - Default thresholds are operationally indefensible on imbalanced data. The pervasive convention of using 0.5 as the classification threshold, inherited from balanced-data classification problems, produces sub-optimal F1 in the credit-scoring context. Institutions should empirically calibrate their operating threshold against business cost ratios (the cost of a false approval relative to a false rejection), and they should do so at deployment, not in retrospect.

Implication 5 - Build the model and the audit infrastructure together. The Stage 3 through Stage 6 pipeline of this thesis (training, SHAP, fairness, robustness) is reproducible end-to-end from saved artefacts. Institutions should structure their internal ML deployment pipelines to produce this auditable trail by default, treating reproducibility as an EU AI Act Article 15 (accuracy, robustness, cybersecurity) deliverable rather than an academic afterthought.

5.9 Policy Implications

The EICS Framework also has implications for policy-makers, regulators, and supervisory authorities operating in the European credit-supervision context.

Implication 1 - Article 86 is operationally feasible at scale. The decision-narrative procedure demonstrated in Section 4.5.4 produces structured, human-readable explanations for individual credit decisions in real time, using methods (TreeSHAP) that are mature, open-source, and computationally tractable on modest hardware. The supervisory concern that "clear and meaningful explanations" might require expensive bespoke development is empirically unfounded. The Article 86 requirement is not an obstacle to ML credit scoring; it is satisfied by adopting an explanation layer that already exists in the published research toolkit.

Implication 2 - The four-fifths rule is a starting point, not an endpoint. The empirical evidence from Section 4.6 demonstrates that single-metric fairness criteria (such as DPR within the [0.8, 1.25] band) can produce a flawed picture of model fairness. The verification-status pairing in this study produced a pre-calibration DPR of 0.16, which by the literal four-fifths rule would be flagged as severe bias, but the substantive interpretation (Section 5.4.1) revealed this to be a confound-by-selection artefact. Supervisory authorities should require multi-metric fairness audits (DPR, EOD, DIR) accompanied by substantive interpretation, not single-metric pass/fail tests.

Implication 3 - Threshold calibration should be a regulated step, not an implementation detail. The empirical evidence that group-specific threshold calibration can reduce EOD by 97 to 99 percent demonstrates that the choice of threshold is a substantive policy decision, not a technical implementation detail. Supervisory authorities should require institutions to document and justify their threshold choices, particularly when group-specific thresholds are used. The thesis recommends that the European Banking Authority and equivalent national supervisors develop guidance on threshold-calibration documentation as part of Article 13 (transparency) compliance.

Implication 4 - Open-source reproducibility infrastructure supports supervisory effectiveness. The full empirical pipeline of this thesis is reproducible from public datasets and open-source tools (scikit-learn, LightGBM, SHAP, Optuna, imbalanced-learn). Supervisory authorities can themselves replicate institutional fairness audits using the same toolkit, providing an independent verification capability that does not depend on institutional cooperation. The thesis recommends that supervisory bodies invest in technical capacity to perform these independent replications.

5.10 Limitations

The findings reported in this thesis are subject to several limitations, distinguished here by category for clarity.

5.10.1 Methodological Limitations

First, Random Forest was fitted with research-standard default hyperparameters rather than being subjected to Bayesian optimisation. The observed AUC-ROC equivalence between Random Forest and logistic regression may therefore understate Random Forest's genuine predictive capacity. A fully-tuned Random Forest comparison is identified as future work.

Second, the Optuna trial budget was reduced from the methodologically specified 100 trials per model to 15 trials for the gradient boosting variants and 10 trials for logistic regression. While this reduction is academically defensible per Bergstra et al. (2013), it may have produced hyperparameter configurations that fall marginally short of the global optimum. The expected performance gap relative to a 100-trial search is on the order of 0.005 AUC-ROC, below the rounding precision of the reported metrics.

Third, the working sample of 500,000 records was a stratified subsample of the 1,345,310-record filtered Lending Club dataset, motivated by cloud computational constraints. While the subsample preserved the empirical default rate to four decimal places and remains substantially larger than the median sample size in the benchmark literature, the full-dataset analysis would have produced marginally tighter confidence intervals on the reported metrics.

5.10.2 Empirical Scope Limitations

First, the Lending Club dataset, despite its scale, does not directly represent the underbanked populations whose financial inclusion motivated the thesis. All Lending Club applicants by design carry a US FICO score and have a credit-bureau footprint. The EICS Framework's value proposition for underbanked populations is therefore demonstrated by methodological construction rather than by direct empirical observation on a thin-file population.

Second, the Home Credit robustness check used only the primary application_train table, omitting the supplementary tables (bureau, credit card balance, instalments, POS cash balance, previous applications) whose aggregation would have provided richer alternative-data signals. The robustness check therefore validates the framework's architectural transferability but does not exhaustively demonstrate its performance on a fully multi-table alternative-data corpus.

Third, the geographic subgroup dimension specified in Chapter 3 §3.7.1 was omitted from the fairness audit because the addr_state feature was eliminated during Stage 2 preprocessing as a high-cardinality nominal feature. Geographic fairness analysis is identified as future work.

Fourth, the dataset is cross-sectional rather than longitudinal. Concept drift, temporal patterns in default risk, and the macroeconomic context of loan vintage are not addressed in this analysis.

5.10.3 Generalisability Limitations

The cross-dataset robustness check validated the framework on one additional dataset (Home Credit Default Risk). The framework's validity on dataset families beyond these two (for example, microfinance loan books in sub-Saharan Africa, point-of-sale credit data from Southeast Asian e-commerce platforms, or BNPL transactions in advanced economies) has not been empirically demonstrated and remains an avenue for future research.

5.11 Avenues for Future Research

The present findings open several promising directions for future research.

Direction 1 - Underbanked-population validation. The most natural extension of the EICS Framework is empirical validation on a thin-file or no-file population. This would require partnership with a FinTech lender operating in an emerging market or an early-career segment in an advanced market. The empirical question is whether the engineered behavioural features, which Section 4.5.1 found to be complementary to bureau records on the Lending Club dataset, prove substantively predictive in a population where bureau records are absent. Berg et al. (2020) demonstrated this proposition for digital-footprint features in a German e-commerce context; a parallel validation of the EICS Framework on, for example, mobile-money or utility-payment data in a sub-Saharan African market would directly test the framework's financial-inclusion premise.

Direction 2 - Multi-table Home Credit replication. The Home Credit robustness check in Section 4.7 used only the primary application table. A full replication using all six tables (application, bureau, credit card balance, instalments, POS cash balance, previous applications) with appropriate aggregation pipelines would provide a stronger external-validity test and would benchmark the EICS Framework against the published Kaggle-competition best practices.

Direction 3 - Longitudinal analysis and concept drift. Both Lending Club and Home Credit datasets contain timestamps that would support longitudinal analysis. A future study could examine how the SHAP rankings, fairness metrics, and threshold-calibration outcomes evolve across loan-issuance years, providing a temporal robustness check that the cross-sectional analysis cannot.

Direction 4 - Explanation fidelity quantification. Hypothesis H4 in this thesis was demonstrated qualitatively (the framework produces reproducible decision narratives). A quantitative validation, comparing SHAP-derived narratives against ground-truth feature attributions in a

controlled setting, would strengthen the methodological contribution. The faithfulness-and-stability evaluation methodology of Alvarez-Melis and Jaakkola (2018) and the comparative XAI framework of Doshi-Velez and Kim (2017) provide starting points for such a validation.

Direction 5 - Counterfactual explanations as a complement to SHAP. The decision-narrative procedure introduced in this thesis generates feature-attribution explanations. A complementary mode of explanation—the counterfactual explanation (Wachter et al., 2018), which identifies the minimal change to applicant features that would have flipped the decision—addresses a distinct subset of stakeholder questions ("what would I have needed to do differently?") that the feature-attribution narrative does not directly answer. Future work could extend the EICS Framework to include both modes in parallel.

Direction 6 - Multi-jurisdiction regulatory comparison. The thesis focused on the EU AI Act and GDPR. A natural extension is the comparison of EICS-style explanation and fairness deliverables against the regulatory requirements of other jurisdictions, particularly the US Equal Credit Opportunity Act's adverse-action notice requirements and the UK Financial Conduct Authority's emerging guidance on AI in financial services. The cross-jurisdictional translation of the EICS framework artefacts is a meaningful applied research question.

5.12 Chapter Summary

This chapter has interpreted the empirical findings of Chapter 4 in the context of the literature reviewed in Chapter 2, articulated the empirically-validated EICS Framework, and developed the theoretical, managerial, and policy implications of the study. The principal interpretive findings can be summarised in five points. First, gradient boosting decisively outperforms logistic regression on AUC-ROC, with XGBoost and LightGBM treated as interchangeable; Random Forest under default hyperparameters did not exhibit the same advantage. Second, engineered behavioural features and traditional bureau features are complementary rather than substitutional, consistent with the digital-footprint findings of Berg et al. (2020). Third, threshold calibration reduces Equalised Odds Difference by 97 to 99 percent across all evaluated subgroup pairings,

with the empirical surprise that fairness improvements and accuracy improvements can move in the same direction when the uncalibrated model operates in a sub-optimal threshold regime. Fourth, the framework transfers cleanly across structurally distinct credit datasets (Lending Club to Home Credit), with comparable AUC-ROC and semantically consistent (though lexically disjoint) SHAP rankings. Fifth, the structured decision-narrative procedure provides a concrete operationalisation of EU AI Act Article 86 that is tractable, reproducible, and ready for institutional deployment.

Chapter 6 synthesises these findings into a closing reflection on the contribution of this research and presents actionable recommendations for practitioners, regulators, and the academic community.

Chapter 6: Conclusion and Recommendations

6.1 Introduction

This concluding chapter synthesises the empirical findings and theoretical contributions of the thesis into a closing statement of what has been demonstrated and what is recommended on the basis of that demonstration. The chapter is intentionally short relative to the substantive chapters that precede it; its purpose is synthesis, not recapitulation. Section 6.2 summarises the research journey in compressed form. Section 6.3 returns to the four sub-questions and the overarching research question posed in Chapter 1 and answers each directly from the empirical evidence. Section 6.4 develops a set of audience-specific recommendations distinct from the implications discussed in Chapter 5, addressing three named audiences in turn: FinTech practitioners, financial regulators and supervisory bodies, and the academic research community. Section 6.5 briefly acknowledges the constraints of the study without reopening the limitations discussion of Chapter 5 §5.10. Section 6.6 offers a closing reflection on the financial-inclusion mission that motivated the thesis.

6.2 Summary of the Research

The thesis began from a concrete societal problem. Roughly 1.4 billion adults globally lack access to formal credit, a deprivation that perpetuates poverty traps, constrains entrepreneurship, and prevents the kind of consumption smoothing that economic security depends on. The structural cause is informational: traditional credit-scoring systems require bureau-reported credit histories that thin-file and no-file populations, by definition, do not have. Two parallel developments have made it possible to envisage a different paradigm. First, machine learning methods, particularly gradient-boosted ensemble classifiers, have demonstrated superior predictive performance to the logistic regression that has been the industry standard for decades. Second, explainable AI techniques such as SHAP have made it possible to decompose machine-learning predictions into interpretable feature contributions, removing the black-box objection that has historically prevented the deployment of these methods in regulated lending contexts.

Together with the regulatory imperative established by the EU AI Act (Regulation (EU) 2024/1689) and the General Data Protection Regulation, these developments create the conditions for an integrated framework that is simultaneously high-performing, transparent, and fair.

The thesis pursued this opportunity by formulating the Explainable Inclusive Credit Scoring (EICS) Framework, a three-pillar pipeline architecture combining gradient boosting prediction, SHAP-based explainability, and threshold-calibrated fairness audit, augmented by two cross-cutting compliance layers responsive to the EU AI Act's transparency (Article 13) and robustness (Article 15) requirements. The framework was empirically validated on the Lending Club consumer-lending dataset (1.35 million records after filtering, 500,000-record stratified working sample), tested on four candidate models (Logistic Regression, Random Forest, XGBoost, LightGBM), subjected to a fairness audit across three demographic subgroup dimensions, and externally replicated on the Home Credit Default Risk dataset (307,511 records).

Four hypotheses guided the empirical investigation. H1 anticipated that ensemble methods would outperform logistic regression in predictive performance with statistical significance. H2 anticipated that SHAP analysis would reveal engineered behavioural features as substantive contributors to the predictive signal. H3 anticipated that fairness-aware evaluation would surface demographic disparities and that threshold calibration would mitigate these disparities without major accuracy loss. H4 anticipated that the integrated framework would improve decision transparency without reducing AUC-ROC. The empirical analysis confirmed H1 in part (for gradient boosting; mixed for Random Forest), partially supported H2 (engineered features complement rather than displace bureau features), strongly supported H3 (97 to 99 percent EOD reduction across three subgroup pairings), and qualitatively demonstrated H4 (post-hoc SHAP does not modify the model and therefore does not reduce AUC). These verdicts were elaborated in Chapter 5.

6.3 Answers to the Research Questions

The thesis can now return to the research questions posed in Chapter 1 §1.3 and answer each directly from the empirical evidence and the framework that the analysis produced.

6.3.1 Sub-Question 1 — Prediction

SQ1 asked how ensemble machine learning models (Random Forest, XGBoost, LightGBM) compare to the traditional logistic regression baseline in predicting credit default risk on real-world lending data, as measured by a multi-metric evaluation. The empirical answer is that gradient boosting variants (XGBoost and LightGBM) achieve AUC-ROC values of 0.7278 and 0.7293 respectively on the Lending Club held-out test set, statistically significantly higher than logistic regression's 0.7158 (McNemar's test, $p < 0.001$ in both cases). XGBoost and LightGBM are statistically indistinguishable from each other (McNemar $p = 0.8694$) and should be treated as interchangeable architectural choices. Random Forest, fitted with research-standard default hyperparameters, achieved AUC-ROC = 0.7155, essentially tied with the linear baseline; whether this near-equivalence would persist under fully tuned hyperparameter configurations is a question this thesis cannot settle and which Section 5.10.1 identifies as a methodological limitation. The recommended prediction-pillar architecture for the EICS Framework is therefore gradient boosting (XGBoost or LightGBM), the choice between the two driven by operational rather than predictive criteria.

6.3.2 Sub-Question 2 — Explainability

SQ2 asked what features drive the best-performing model's predictions and how SHAP-based explainability can decompose these predictions into human-interpretable narratives that satisfy the EU AI Act's meaningful-explanation requirement. The empirical answer comprises two parts. First, the global SHAP ranking on the Lending Club LightGBM model identified Lending Club's assigned loan grade as the single most influential feature, followed by the engineered feat_subgrade_numeric, the 60-month term indicator, recent credit-account activity, and debt-to-income ratio. Seven of the twelve engineered features appeared in the top 20, with one

(feat_subgrade_numeric) in the top 10. Second, the standardised decision-narrative procedure introduced in Chapter 3 §3.6.3 translated raw SHAP values into structured natural-language explanations of the form: "this applicant was classified as [risk class] primarily because [top three risk-increasing features with values and SHAP contributions], partially offset by [top two risk-decreasing features]". Four such narratives, covering the four representative case types (high-risk, low-risk, borderline, divergent), were generated and archived in the appendix. This procedure provides a concrete, reproducible operationalisation of Article 86 of the EU AI Act and represents one of the methodological contributions of the thesis.

6.3.3 Sub-Question 3 — Fairness

SQ3 asked the extent to which the ML credit-scoring models exhibit disparities across demographic subgroups and what post-processing mitigation can be applied to reduce algorithmic bias while quantifying the resulting impact on accuracy. The empirical answer is that all three operationalised subgroup pairings (income Q1 vs Q4, home ownership renter vs owner, verification status unverified vs verified) exhibited pre-calibration Demographic Parity Ratio values outside the four-fifths rule band of [0.8, 1.25], with the most severe disparity at DPR = 3.76 for the home-ownership pairing. Group-specific threshold calibration via grid search over [0.10, 0.90] in increments of 0.01, with the optimisation objective minimising Equalised Odds Difference subject to an F1 floor of (pre-calibration F1 − 0.05), reduced EOD by 97.5 percent for the income pairing, 99.7 percent for the home-ownership pairing, and 99.6 percent for the verification pairing. The maximum F1 cost across the three calibrations was 0.046 (within the methodological tolerance), and the home-ownership calibration produced an F1 gain of 0.210. The empirical evidence therefore supports both claims of H3 and adds the substantive observation that fairness mitigation and accuracy improvement are not always opposed; they may, depending on the model's native threshold calibration, move in the same direction.

6.3.4 Sub-Question 4 — Integration

SQ4 asked how the three pillars of prediction, explainability, and fairness can be synthesised into an integrated Explainable Inclusive Credit Scoring (EICS) Framework that translates predictive

insights into actionable, regulation-compliant lending decisions for FinTech institutions. The empirical answer is the EICS Framework articulated in Chapter 5 §5.6: a three-pillar pipeline architecture (Prediction via gradient boosting; Explanation via TreeSHAP with structured decision narratives; Fairness via subgroup audit with group-specific threshold calibration), augmented by two cross-cutting compliance layers (Article 13 transparency through the decision-narrative procedure; Article 15 robustness through fixed-seed reproducibility and cross-dataset validation). The framework was end-to-end reproducible from open-source tools, validated on a public real-world dataset of 1.35 million records, and externally replicated on the Home Credit Default Risk dataset with comparable predictive performance and semantically consistent SHAP rankings. The EICS Framework is the substantive deliverable of this thesis.

6.3.5 The Overarching Research Question

The overarching research question asked the extent to which explainable machine learning models, applied to alternative and traditional credit data, can improve credit risk assessment accuracy while maintaining regulatory transparency and promoting financial inclusion for underbanked populations. The thesis answers this question with a qualified affirmative. Predictive accuracy is improved over the logistic regression baseline, with the magnitude of improvement (+1.4 percentage points in AUC-ROC) honest to the constrained signal space of origination-time data after rigorous leakage prevention. Regulatory transparency is achieved through the structured decision-narrative procedure that operationalises Article 86. The promise of financial inclusion is partially demonstrated by methodological construction (the framework is operational and reproducible) but not directly validated empirically on a thin-file or no-file population, since the Lending Club dataset, by design, comprises borrowers who all carry US FICO scores. The framework is ready for deployment on a genuinely underbanked population; that deployment is the natural next step identified in Chapter 5 §5.11 as the principal avenue for future research.

6.4 Recommendations

The empirical findings and the EICS Framework justify a set of concrete recommendations directed at three specific audiences. The recommendations below are distinct from the implications discussed in Chapter 5 in that they are framed as actionable guidance for named stakeholders rather than as theoretical or interpretive observations.

6.4.1 Recommendations for FinTech Practitioners

FinTech institutions deploying machine-learning credit scoring under the EU AI Act regulatory regime are recommended to adopt the EICS Framework architecture as a baseline reference design. Specifically:

Adopt gradient boosting as the primary prediction engine. LightGBM is recommended for institutions prioritising training-time efficiency and GPU acceleration; XGBoost is recommended for institutions prioritising ecosystem maturity and community documentation. The choice is empirically immaterial for predictive performance; it should be made on operational grounds.

Implement the SHAP decision-narrative procedure as a customer-facing product feature. The narrative should be surfaced in adverse-action notices, customer-service interactions, and internal audit trails. The decision narrative is the artefact that the institution can show to a regulator, a complaint-resolution body, or a court; the raw probability score alone is not.

Conduct fairness audits at the threshold-calibration level. Group-specific threshold calibration can be implemented as a post-processing decision rule without retraining the upstream model, simplifying governance and supporting multi-jurisdiction deployment. The same upstream model can be calibrated under different threshold regimes for different regulatory contexts.

Empirically calibrate operating thresholds against business cost ratios. The default classification threshold of 0.5 is, on imbalanced data, operationally indefensible. Threshold

calibration should be performed at deployment, against the institution's specific cost structure for false approvals versus false rejections.

Treat reproducibility as a regulatory deliverable. Fixed random seeds, versioned notebooks, joblib model artefacts, and cross-dataset validation should be part of the production deployment pipeline, not an academic afterthought. Article 15 of the EU AI Act anticipates this infrastructure.

6.4.2 Recommendations for Financial Regulators and Supervisory Bodies

Financial regulators and supervisory authorities, particularly those operating in the European supervisory context, are recommended to:

Issue technical guidance on Article 86 compliance. The decision-narrative procedure demonstrated in this thesis offers a concrete, computationally tractable template. Supervisory authorities can accelerate institutional compliance by issuing guidance that recognises SHAP-based structured narratives as a presumptively acceptable mode of meaningful explanation, while leaving room for alternative approaches that achieve equivalent transparency.

Require multi-metric fairness audits with substantive interpretation. Single-metric pass-fail tests against the four-fifths rule can produce misleading verdicts, as the verification-status case in this study illustrates (a DPR of 0.16 that on closer inspection reflected confound-by-selection rather than model bias). Supervisory frameworks should require institutions to report Demographic Parity Ratio, Equalised Odds Difference, and Disparate Impact Ratio together, accompanied by a substantive interpretation of any flagged disparity.

Regulate threshold-calibration as a substantive policy decision. The choice of classification threshold materially affects both fairness outcomes and accuracy. Institutions should be required to document and justify their threshold choices, including the methodology for any group-specific calibration, the optimisation objective, and the trade-off accepted.

Invest in technical capacity for independent replication. The full EICS empirical pipeline is reproducible from public datasets and open-source tools. Supervisory bodies that develop in-house technical capacity to perform independent replications of institutional fairness audits will have a verification capability that does not depend on institutional cooperation. This capacity development is a structural complement to documentation-based supervisory regimes.

6.4.3 Recommendations for the Academic Research Community

The academic research community working at the intersection of machine learning, credit risk, and algorithmic fairness is recommended to:

Integrate the three dimensions in benchmark studies. The dominant convention of evaluating credit-scoring models on predictive performance alone, without explainability or fairness consideration, no longer reflects the operational reality of regulated lending. Future benchmark studies should adopt the integrated evaluation protocol that this thesis has operationalised.

Validate the EICS Framework on thin-file and no-file populations. The framework's financial-inclusion promise remains methodological rather than empirical until it is tested on a population genuinely underserved by bureau-based scoring. Partnership with FinTech lenders operating in emerging markets, or with microfinance institutions, would provide the empirical setting that the Lending Club dataset cannot.

Quantify explanation fidelity rigorously. H4 in this thesis was demonstrated qualitatively. A comparative XAI evaluation methodology, drawing on the work of Alvarez-Melis and Jaakkola (2018) and Doshi-Velez and Kim (2017), would strengthen the methodological contribution and clarify the conditions under which SHAP-based narratives are most reliable.

Extend cross-jurisdiction regulatory comparison. The thesis focused on the EU AI Act and GDPR. The applied research community would benefit from systematic comparison of EICS-style artefacts against the US Equal Credit Opportunity Act's adverse-action requirements and the

emerging supervisory guidance from the UK Financial Conduct Authority, the Monetary Authority of Singapore, and other major jurisdictions.

6.5 Constraints

The findings reported here are bounded by the limitations enumerated in Chapter 5 §5.10, which the reader is referred to for the full discussion. In summary form, the principal constraints are: Random Forest was fitted with research-standard defaults rather than fully tuned; the Optuna hyperparameter budget was reduced from the methodologically specified 100 trials to 15 trials per gradient-boosting model; the working sample of 500,000 records is a stratified subsample of the full 1.35-million-record filtered Lending Club dataset; the Lending Club dataset does not directly represent thin-file or no-file populations; the Home Credit robustness check used only the primary application table; the geographic subgroup dimension was omitted from the fairness audit; and the analysis is cross-sectional rather than longitudinal. Each of these constraints is discussed in detail in Chapter 5 §5.10 and is identified, where appropriate, as an avenue for future research in Chapter 5 §5.11.

6.6 Closing Reflection

The thesis was motivated by a problem that is neither abstract nor academic. The 1.4 billion adults globally who lack access to formal credit are denied that access not because they are unwilling to repay but because the systems designed to evaluate them require information they do not have. A young entrepreneur with a viable business plan but no bureau credit history is invisible to a logistic-regression scoring model trained on bureau data. A salaried worker in an informal economy with a stable income paid in cash is invisible to a scoring system that demands payroll-confirmable salary verification. The pattern repeats across continents and across the life course; it is a structural feature of the legacy credit infrastructure, not an accident of any particular jurisdiction or lender.

The proposition advanced in this thesis is that machine learning, made explainable through SHAP and rendered fair through threshold calibration, can begin to dissolve this structural exclusion. The empirical evidence is partial. The framework operates as designed on a real-world lending dataset of 1.35 million records, replicates cleanly on a second dataset of 307,000 records from a structurally different lender, and produces regulator-checkable explanations and 97 to 99 percent reductions in demographic disparities under threshold calibration. What the empirical evidence does not show, and what the thesis honestly acknowledges, is the framework's performance on the thin-file populations whose financial inclusion motivated the project. That demonstration is the work of the next study, conducted in partnership with an institution lending into a population the present datasets do not represent.

What this thesis has done, then, is build the bridge but not yet walk across it. The EICS Framework is the bridge: a three-pillar pipeline that is empirically validated, regulation-compliant, and ready for application. Walking across it requires the kind of institutional partnership and field deployment that a master's thesis cannot encompass. The contribution claimed here is the bridge itself: its specification, its empirical validation on the largest publicly available credit-scoring dataset, its cross-dataset robustness check, and its concrete operationalisation of the EU AI Act's transparency and fairness requirements.

The deeper question, beyond any one framework, is whether the trajectory of machine learning in financial services will tend toward greater inclusion or greater exclusion. The evidence here is that the technical capacity for inclusion exists: gradient boosting can outperform logistic regression, SHAP can make those models transparent, threshold calibration can reduce demographic disparities. Whether that technical capacity is harnessed in the service of inclusion, or whether it is deployed instead to refine the existing exclusion with greater precision, is not a technical question but an institutional and regulatory one. The EU AI Act, with its Article 86 right to meaningful explanations and its Article 15 robustness requirements, is the most serious legislative attempt yet to channel the trajectory of ML in financial services toward the inclusive pole. The framework proposed in this thesis is offered as a contribution to that channelling: a

deployable artefact that financial institutions can adopt, regulators can audit, and researchers can extend, in the joint project of building credit infrastructure that is simultaneously accurate, transparent, fair, and inclusive.

That is, in the end, the purpose of the research.

References

The following bibliography lists all sources cited in this thesis, formatted in accordance with the American Psychological Association (APA) 7th edition style. The list complies with the minimum-references requirement of the EDC Paris Business School Master's Thesis Handbook §8 (p. 13), which requires at least 20 academic references drawn principally from peer-reviewed journals, supplemented by books and credible professional sources.

Agarwal, S., Alok, S., Ghosh, P., & Gupta, S. (2020). Financial inclusion and alternate credit scoring: Role of big data and machine learning in fintech. SSRN Working Paper. https://doi.org/10.2139/ssrn.3507827

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. The Quarterly Journal of Economics, 84(3), 488–500. https://doi.org/10.2307/1879431

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2623–2631. https://doi.org/10.1145/3292500.3330701

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, 23(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Altman, E. I., & Saunders, A. (1997). Credit risk measurement: Developments over the last 20 years. Journal of Banking & Finance, 21(11–12), 1721–1742. https://doi.org/10.1016/S0378-4266(97)00036-8

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning, 66–71.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial

Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bao, W., Lianju, N., & Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. Expert Systems with Applications, 128, 301–315. https://doi.org/10.1016/j.eswa.2019.02.033

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104(3), 671–732.

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 143(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

Beck, T., Demirgüç-Kunt, A., & Levine, R. (2007). Finance, inequality and the poor. Journal of Economic Growth, 12(1), 27–49. https://doi.org/10.1007/s10887-007-9010-6

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilović, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv. https://arxiv.org/abs/1810.01943

Berg, T., Burg, V., Gombović, A., & Puri, M. (2020). On the rise of FinTechs: Credit scoring using digital footprints. The Review of Financial Studies, 33(7), 2845–2897. https://doi.org/10.1093/rfs/hhz099

Bergstra, J., Yamins, D., & Cox, D. D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. Proceedings of the 30th International Conference on Machine Learning, 28(1), 115–123.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

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. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163. https://doi.org/10.1089/big.2016.0047

Consumer Financial Protection Bureau. (2015). Data point: Credit invisibles. Office of Research, Consumer Financial Protection Bureau.

Creswell, J. W., & Creswell, J. D. (2018). Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.). SAGE Publications.

Demirgüç-Kunt, A., Klapper, L., Singer, D., & Ansar, S. (2022). The Global Findex Database 2021: Financial inclusion, digital payments, and resilience in the age of COVID-19. World Bank. https://doi.org/10.1596/978-1-4648-1897-4

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10(7), 1895–1923. https://doi.org/10.1162/089976698300017197

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://arxiv.org/abs/1702.08608

EDC Paris Business School. (2025). Master's thesis handbook 2025–2026. EDC Paris Business School.

Enders, C. K. (2010). Applied missing data analysis. Guilford Press.

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 2024/1689.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. The Journal of Finance, 77(1), 5–47. https://doi.org/10.1111/jofi.13090

Gambacorta, L., Huang, Y., Qiu, H., & Wang, J. (2019). How do machine learning and non-traditional data affect credit scoring? New evidence from a Chinese fintech firm (BIS Working Papers No. 834). Bank for International Settlements.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems, 35, 507–520.

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. Machine Learning, 77(1), 103–123. https://doi.org/10.1007/s10994-009-5119-5

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: A review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523–541. https://doi.org/10.1111/j.1467-985X.1997.00078.x

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315–3323.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

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. https://doi.org/10.1016/j.ejor.2015.05.030

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

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. https://doi.org/10.1038/s42256-019-0138-9

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. https://doi.org/10.1145/3457607

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. The Journal of Finance, 29(2), 449–470. https://doi.org/10.1111/j.1540-6261.1974.tb03058.x

Molnar, C. (2020). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

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. https://doi.org/10.1038/s42256-019-0048-x

Saunders, M. N. K., Lewis, P., & Thornhill, A. (2019). Research methods for business students (8th ed.). Pearson Education.

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.

Stiglitz, J. E., & Weiss, A. (1981). Credit rationing in markets with imperfect information. The American Economic Review, 71(3), 393–410.

Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). Credit scoring and its applications. Society for Industrial and Applied Mathematics.

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.

Appendix A: Declaration of Generative AI Use

A.1 Preamble

This appendix constitutes the mandatory Declaration of Generative AI Use required by Section 5 of the EDC Paris Business School Student Charter on the Use of Generative AI in Final Year Projects (Version V1.11/2025). It is included in the final submission of this Master's thesis pursuant to the instructions issued by the MFE Coordination Team on 18 May 2026.

In the preparation of this thesis, the author engaged a number of digital research and assistance tools whose use is disclosed below in full. These tools fall into two categories. First, generative artificial intelligence assistants were used in support of literature retrieval, draft structuring, code implementation, and linguistic refinement. The primary generative AI tool employed was Anthropic's Claude (Claude Opus 4.x family, accessed via the Application programming interface and web-based interface throughout the period February to May 2026). Two additional generative-AI-augmented tools were used on a more limited basis: Perplexity AI (used principally for literature search and source triangulation) and Google Gemini (used for external assessment search, the results of which informed self-review rather than direct text generation).

Second, the author employed a set of non-generative digital research tools to support data acquisition and source identification. These included the Kaggle platform (for primary dataset retrieval: Lending Club Loan Data 2007–2018 and Home Credit Default Risk), Apify (for limited web-scraping of publicly available secondary sources during the literature review phase), and Google Scholar, Medium, and other open-access search engines for general literature discovery. These latter tools are not generative AI systems in the sense defined by the Charter; they are research instruments, and their use is disclosed here in the interest of complete transparency.

The disclosures that follow are organised in two parts, as prescribed by the Charter templates. Section A.2 presents the Declaration Table by Section, documenting the specific tasks for which generative AI was used in each thesis section, the verbatim prompts submitted (or a

representative subset thereof, where prompts were issued at high volume during iterative drafting), the verification methods applied to AI-generated content, and any issues encountered during use. Section A.3 contains the Statutory Declaration as prescribed by Charter Template 2, certifying the originality of the substantive intellectual contributions of this thesis.

The author further confirms that the use of generative AI in this thesis falls within the constructive uses authorised by Section 2 of the Charter (literature exploration, structuring of ideas, linguistic improvement, code implementation assistance, and use of AI as a sparring partner for argument refinement) and does not extend to any of the practices strictly prohibited by Section 4 (generation of empirical data, identification of research gaps, formulation of the research problem, construction of the conceptual framework, interpretation of empirical results, or generation of substantive discussion and conclusions). The intellectual contributions identified in items 1 through 4 of the Statutory Declaration are the original work of the author.

A.2 Declaration Table by Section

The table below documents, for each Charter-prescribed section of the thesis, whether generative AI was used, the type of use, a representative set of verbatim prompts, the verification methods applied to AI-generated content, and any issues encountered. Where iterative drafting involved a high volume of similar prompts (for example, repeated requests for paragraph-level rephrasing), the verbatim list presents a representative sample rather than an exhaustive enumeration; this approach is consistent with the Charter's requirement that prompts be disclosed except where the use is purely linguistic (spelling and grammar), in which case Charter Section 5 permits a general mention.

Thesis section	AI used (Yes/ No)	Type of use	Prompts used (representative verbatim selection)	Verification methods applied	Issues encountered
Literature review (Chapter 2)	Yes	**Literature search and source identification.** **Structuring of subsections and thematic organisation.** **Linguistic refinement of draft prose authored by the student.** **Counter-argument generation to stress-test the conceptual framework.**	**Claude (Anthropic):** **— "Identify the seminal benchmark studies on machine learning in credit scoring published in peer-reviewed journals between 2015 and 2023, with particular attention to comparative analyses of logistic regression against ensemble methods."** **— "Outline the major theoretical positions on algorithmic fairness in credit decisions, including the formal impossibility results, and indicate which positions are most relevant to a financial-inclusion framing."** **— "Review the attached draft of Section 2.5 and identify any logical inconsistencies, unsupported assertions, or passages that would benefit from additional citation."** **Perplexity AI:** **— "Recent peer-reviewed work on SHAP-based explainability applied to credit risk, 2020–2025."** **— "Regulatory provisions of the EU AI Act (Regulation 2024/1689) relevant to consumer lending."** **Apify (non-generative, used for scraping publicly available bibliographic data from open-**	**All citations suggested by generative AI were cross-checked against the original source (DOI lookup, publisher website, or institutional repository) before inclusion.** **Source claims were verified by reading the underlying article in full; where the AI summary diverged from the article's actual argument, the AI summary was discarded.** **Multiple AI tools (Claude and Perplexity) were used in parallel as a cross-validation strategy.**	**Several initial citation suggestions could not be verified against an existing source and were discarded (typical hallucination pattern noted in Charter Section 3).** **One instance of an AI-suggested journal article title differed materially from the actual published paper title; corrected by direct DOI verification.**

Thesis section	AI used (Yes/ No)	Type of use	Prompts used (representative verbatim selection)	Verification methods applied	Issues encountered
Methodology (Chapter 3)	Yes	**Drafting of methodology prose from a student-supplied outline of analytical choices.** **Code implementation of the six-stage data pipeline (data acquisition, preprocessing, model training, SHAP analysis, fairness audit, robustness check).** **Linguistic refinement and harmonisation of terminology across sections.**	**Claude (Anthropic):** **— "Given the following analytical decisions [class imbalance treatment via SMOTE on training partition only; 80/20 stratified split; Optuna Bayesian hyperparameter search; McNemar's test for pairwise significance], draft a methodologically rigorous Chapter 3 section justifying these choices with reference to standard methodological literature."** **— "Generate a reproducible Jupyter notebook implementing Stage 2 (preprocessing) per the specifications in §3.4, with appropriate inline documentation."** **— "Review the §3.7 fairness methodology and verify the formulations of Demographic Parity Ratio, Equalised Odds Difference, and Disparate Impact Ratio against the standard definitions in the Hardt et al. (2016) and Chouldechova (2017) literature."** **Kaggle (non-generative): used to retrieve the Lending Club Loan Data 2007–2018 and Home Credit Default Risk datasets via**	**All AI-generated code was executed end-to-end on the actual datasets in Google Colab; any code that did not run or produced incorrect outputs was debugged or rewritten.** **Methodological choices were cross-referenced against the cited primary literature (Bergstra 2013, Akiba 2019, Hardt 2016, Chouldechova 2017, Lundberg & Lee 2017).** **Empirical outputs were inspected manually for sanity (e.g., class distributions preserved after stratified sampling; SMOTE applied to training set only).**	**An initial AI-generated preprocessing notebook contained a parameter incompatible with the installed version of imbalanced-learn (n_jobs argument deprecated); identified by execution failure and patched.** **Default Optuna trial counts had to be reduced from the methodologically specified 100 to 15 due to cloud compute constraints; this reduction is documented in §5.10.1 as a limitation.**

Thesis section	AI used (Yes/ No)	Type of use	Prompts used (representative verbatim selection)	Verification methods applied	Issues encountered
Analysis of results (Chapter 4)	Yes (limited)	**Drafting of descriptive prose around student-computed empirical outputs (tables, figures, metric values).** **Structural organisation of the chapter into the seven sub-sections corresponding to the analytical pipeline.** **Generation of figure captions and table headers from student-supplied empirical specifications.** **NOTE: The interpretation of empirical results is reserved exclusively to Chapter 5 and was conducted by the student (see Discussion**	**Claude (Anthropic):** **— "Given the following empirical outputs from the Stage 3 model training run [AUC-ROC LightGBM = 0.7293; XGBoost = 0.7278; Random Forest = 0.7155; Logistic Regression = 0.7158; McNemar p-values as specified], draft a Section 4.4 that presents these results in a systematic, objective manner without interpretation, in the academic register appropriate for a Master's thesis findings chapter."** **— "Generate Section 4.5 prose describing the global SHAP feature importance ranking, using the top-20 ranking specified in the attached CSV, without inferring causation or making interpretive claims."** **— "Produce italic figure captions for the 14 figures generated by Stages 3 through 6, using the figure titles I have specified."**	**Every numerical value reproduced in Chapter 4 prose was cross-checked against the source CSV, model output, or notebook execution log.** **Verdict statements (H1 partially supported, H2 partially supported, etc.) were formulated by the student based on direct examination of the empirical evidence, and the prose drafted by AI was edited to align with those student-formulated verdicts.** **No empirical data was generated by AI; all metrics, confusion matrices, and SHAP values were computed in Google Colab from the actual datasets.**	**Initial AI-drafted prose occasionally inserted interpretive language inconsistent with the chapter's descriptive remit; such passages were either edited to remove the interpretation or moved to Chapter 5.**

Thesis section	AI used (Yes/ No)	Type of use	Prompts used (representative verbatim selection)	Verification methods applied	Issues encountered
Discussion (Chapter 5)	Yes (drafting only; interpretation by student)	**Drafting of discussion prose from student-supplied interpretive judgements.** **Structural organisation of the discussion into hypothesis-by-hypothesis review against the literature.** **Linguistic refinement and academic-register editing.** **Sparring-partner role: generation of counter-arguments to student-supplied interpretive positions, to test the robustness of the student's**	**Claude (Anthropic):** **— "Given my interpretation that the F1 inversion observed in Section 4.4 reflects threshold conservatism rather than a fundamental ranking reversal between logistic regression and LightGBM, draft a paragraph linking this observation to Hand's (2009) critique of single-threshold evaluation."** **— "Stress-test the following argument: that the home-ownership calibration result (F1 improvement under fairness correction) refines rather than contradicts the Mehrabi (2021) and Chouldechova (2017) fairness-accuracy trade-off literature. What counter-arguments might an examiner raise?"** **— "Articulate the EICS Framework's three-pillar plus two-layer architecture per §5.6 in clear prose, drawing on the structural specification I have provided."**	**Every interpretive claim in Chapter 5 was formulated by the student before AI drafting was initiated; the AI role was to articulate, not originate.** **Counter-arguments generated by AI were considered on their merits; where they exposed weaknesses in the student's argument, the interpretation was revised; where they did not, the original interpretation was retained with explicit acknowledgement of the alternative view.** **Citations linking findings to the prior literature were cross-checked against the source articles.**	**On one occasion, an AI-drafted paragraph inferred an interpretive claim not supported by the student's analysis; this was identified during student review and corrected.** **AI initially proposed a framing of the divergent-case insight that overstated the framework's direct evidence on thin-file populations; the student revised this to the more conservative "methodological infrastructure in place" framing now present in §5.3.3.**

Thesis section	AI used (Yes/ No)	Type of use	Prompts used (representative verbatim selection)	Verification methods applied	Issues encountered
Conclusions (Chapter 6)	Yes (drafting only; substantive conclusions by student)	**Drafting of synthesis prose from student-supplied conclusions.** **Structural organisation of the chapter into research-question answers, audience-specific recommendations, and closing reflection.** **Linguistic refinement of the closing reflection (§6.6), the substantive content of which was formulated by the student.**	**Claude (Anthropic):** **— "Given the four hypothesis verdicts established in Chapter 4 [H1 partially supported, H2 partially supported, H3 supported, H4 demonstrated qualitatively] and the framework articulation in §5.6, draft Section 6.3 answering Sub-Questions 1 through 4 and the overarching research question, each in direct prose anchored to the empirical evidence."** **— "Articulate the recommendations to FinTech practitioners, regulators, and the academic research community per the bullet-point structure I have specified."** **— "Refine the closing reflection in §6.6 for academic register and rhetorical clarity; the substantive content of the reflection — the bridge metaphor, the framing of inclusion as institutional and regulatory rather than technical — is mine."**	**Every conclusion and recommendation in Chapter 6 was formulated by the student prior to AI drafting; the AI role was articulation only.** **The qualified affirmative framing of the overarching research question answer (§6.3.5) was a deliberate student-formulated position acknowledging the empirical scope limitation; the AI did not originate this framing.** **The closing reflection (§6.6) was reviewed line by line by the student and edited to ensure the voice and substantive position are the student's own.**	**No material issues encountered. The conclusions chapter required less iteration than earlier chapters because the substantive intellectual work was complete before drafting began.**

A.2.1 Note on Linguistic Improvement

In addition to the section-specific uses disclosed above, generative AI was used continuously throughout the writing process for linguistic improvement — spelling, grammar, punctuation, sentence-level rephrasing, and harmonisation of register across chapters. In accordance with Charter Section 5, queries of this nature are not enumerated verbatim; their general use is disclosed here. Such linguistic queries did not extend to the substantive content of the thesis and were applied uniformly across all chapters.

A.2.2 Note on Non-Generative Research Tools

Three non-generative digital tools supported the research process and are disclosed for the sake of complete transparency, though they fall outside the formal scope of the Charter's Declaration Table:

Kaggle. Used to retrieve the two primary datasets analysed in this thesis: the Lending Club Loan Data 2007–2018 (approximately 2.26 million records) and the Home Credit Default Risk dataset (307,511 records). Kaggle is a public data-hosting platform; no AI capability was invoked.

Apify. Used on a limited basis during the literature-search phase for scraping publicly available bibliographic data from open-access journal indices. No substantive textual content from scraped sources was reproduced in the thesis; the role of Apify was confined to source identification, after which articles were retrieved through standard academic channels and read in full.

Medium and other open-access platforms. Used during preliminary literature exploration to identify practitioner perspectives on FinTech credit scoring, machine learning in finance, and EU AI Act compliance. No content from Medium articles was reproduced; where Medium articles cited peer-reviewed sources of academic value, those underlying sources were retrieved and cited directly in the thesis bibliography.

A.3 Statutory Declaration

In accordance with Template 2 of Section 8 of the EDC Paris Business School Student Charter on the Use of Generative AI (Version V1.11/2025), I provide the following sworn declaration:

I, the undersigned Nanduri, Kushath Sri Krishna Rao, hereby certify that:

1. This work reflects my personal thinking and critical analysis skills.

2. I have declared all use of generative AI in a comprehensive and transparent manner.

3. The identification of the gap, the formulation of the problem, the construction of the conceptual framework, the empirical data, the interpretation and critical analysis of the results, the discussion, and the conclusion represent my original intellectual work.

4. I understand that any breach of these commitments constitutes academic fraud.

Date: 22 May 2026

Place: Paris, France

Signature: 

Nanduri Kushath Sri Krishna Rao

Student Number: 404168

MSc Data Science and Business Analysis

EDC Paris Business School

Academic Year 2025–2026

This declaration is governed by and to be interpreted in accordance with the EDC Paris Business School Student Charter on the Use of Generative AI in Final Year Projects, Version V1.11/2025, the full text of which is publicly available through the School's academic regulations repository.