INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Personal Credit Risk Assessment with Machine Learning: Balancing Performance, Fairness, and Interpretability

Caroline Dalcomuni de Moura

Master of Data Science,

Supervisor:
Dr. Diana Elisabeta Aldea Mendes, Associate Professor,
ISCTE-IUL

Supervisor:
Dr. Sérgio Moro, Full Professor,
ISCTE-IUL

September, 2025

Department of Quantitative Methods for Management and Economics

Department of Information Science and Technology

Personal Credit Risk Assessment with Machine Learning: Balancing Performance, Fairness, and Interpretability

Caroline Dalcomuni de Moura

Master of Data Science,

Supervisor:
Dr. Diana Elisabeta Aldea Mendes, Associate Professor,
ISCTE-IUL

Supervisor:
Dr. Sérgio Moro, Full Professor,
ISCTE-IUL

September, 2025

Acknowledgement

First and foremost, I thank God for the strength and faith that carried me through every step of this journey. To my parents, for their unconditional love, constant support, and for always believing in my potential, without you, reaching this stage would not have been possible. I am also profoundly grateful to my supervisors, Professor Diana Mendes and Professor Sérgio Moro, whose guidance, dedication, and trust were essential to the realization of this work. This thesis simply would not exist without their support.

To my husband, the love of my life and my greatest source of encouragement, thank you for your patience, affection, and for never letting me give up. From the countless nights you stayed awake helping me set up virtual environments and debugging scripts, to being my personal "Python professor," your motivation made this achievement possible.

My heartfelt thanks also go to my best friends, Sofia and Arthur, for their determination, inspiration, and unwavering support, and to my dear friend Beatriz L., who stood by me through countless late nights, projects, and kind words, always reminding me that I would make it to the finish line.

Finally, I am grateful to my former employer, CASAFARI, and my current organization, Bitsight, for fostering the work–life balance that allowed me to pursue this academic goal. To all those who, in one way or another, contributed to the completion of this stage, I extend my most sincere gratitude.

Resumo

Este estudo investiga a aplicação de técnicas de aprendizagem supervisionada Machine Learning (ML) na avaliação de crédito pessoal, com especial atenção à sua relevância para populações sub-bancarizadas e às potenciais implicações para o setor bancário português. Devido a restrições de privacidade no acesso a dados bancários reais, recorreu-se a um conjunto de dados sintéticos da plataforma LendingClub. Embora tenha origem num contexto de peer-to-peer lending, o dataset foi reinterpretado sob a ótica da banca comercial e mapeado, tanto quanto possível, com os critérios de avaliação de solvabilidade definidos pelo sistema bancário.

A investigação seguiu a metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining), abrangendo compreensão dos dados, pré-processamento, análise de correlação e seleção de variáveis, modelação e avaliação. Foram testados diversos algoritmos de aprendizagem supervisionada em cenários de classes equilibradas e não-equilibradas, refletindo a predominância de casos de não-incumprimento nos mercados de crédito. Os métodos de *ensemble*, em particular o LightGBM e o AdaBoost, superaram os modelos tradicionais em termos de recall, accuracy e AUC-ROC, demonstrando forte capacidade discriminatória na identificação de incumprimentos. A Regressão Logística, usada como referência, manteve relevância devido à sua interpretabilidade e alinhamento regulatório.

Para além da performance técnica, o estudo evidencia a importância da transparência e da equidade na avaliação de crédito. A análise de importância das variáveis confirmou que os modelos de ML priorizam critérios próximos dos do Banco de Portugal, reforçando a sua pertinência institucional.

Combinando rigor metodológico e reflexão sobre inclusão financeira, a investigação mostra como o ML pode complementar práticas tradicionais em Portugal, aumentando a precisão preditiva e promovendo acesso mais justo ao crédito.

Palavras-chave: Avaliação de Crédito, *Machine Learning*, Inclusão Financeira, Risco de Inadimplência, Populações sub-bancarizadas, Portugal

Abstract

This study investigates the application of machine learning (ML) techniques in personal credit assessment, with particular attention to their relevance for underbanked populations and their implications for the banking sector. Due to strict privacy and governance restrictions on access to real banking data, a synthetic dataset from the LendingClub platform was employed. While originating in a peer-to-peer lending context, the dataset was reinterpreted through the lens of commercial banking and mapped, as far as possible, against the creditworthiness assessment criteria defined by the Bank of Portugal.

The research followed the Cross-Industry Standard Process for Data Mining (CRISP-DM), covering data understanding, preprocessing, feature selection, modeling, and evaluation. Multiple supervised learning algorithms were tested under both unbalanced and balanced class distributions, reflecting the prevalence of non-default cases in real credit markets. Ensemble methods, particularly LightGBM and AdaBoost, outperformed traditional models in recall, accuracy, and AUC-ROC, demonstrating strong discriminatory capacity in identifying defaults. Logistic Regression, used as a benchmark, retained value due to its interpretability and regulatory alignment.

Beyond technical performance, the study emphasizes transparency and fairness in credit scoring. Feature importance analysis confirmed that ML models prioritize variables aligned with Portuguese credit assessment practices, reinforcing their institutional relevance. By combining methodological rigor with a reflection on inclusiveness, this research contributes to ongoing debates on responsible credit risk modeling. The findings suggest that ML can complement traditional approaches in Portugal, improving predictive accuracy while supporting more equitable access to credit.

Keywords: Credit Assessment, Machine Learning, Financial Inclusion, Credit Risk, Underbanked Populations, Portugal

Index

List of Figures

List of Tables

Abbreviation List

AI: Artificial Intelligence

ANN: Artificial Neural Network

ANOVA: Analysis of Variance

API: Application Programming Interfaces

APRC: The Annual Percentage Rate of Charge

AUC-ROC: Area Under the Curve – Receiver Operating Characteristic

AutoML: Automated Machine Learning

CART: Classification and Regression Trees

CRISP-DM: Cross-Industry Standard Process for Data Mining

CSMB: Conta de Serviços Mínimos Bancários

DNN: Deep Neural Network

DSTI: The Debt Service-to-Income

DT: Decision Tree

EBA: European Banking Authority

ESIS: European Standardised Information Sheet

FPR: False Positive Rate

F1-score: Harmonic mean of Precision and Recall

GDPR: General Data Protection Regulation

INE: Instituto Nacional de Estatística

IQR: The interquartile range

IT: Information technology

IRC: Corporate Income Tax

IRS: Personal Income Tax

KNN: k-Nearest Neighbors

LASSO: Lasso regression

LIGHTGBM: Short for Light Gradient Boosting Machine

LIME: Local Interpretable Model-agnostic Explanations

LR: Logistic regression

LSTM: Long Short-Term Memory

LTV: The loan-to-value ratio

MCC: Matthews Correlation Coefficient

MI: Mutual Information

ML: Machine learning

NB: Naive Bayes

OASP: out-of-court arrears settlement procedure

PRAP: The Pre-Arrears Action Plan

SHAP: Shapley Additive Explanations

SIS: Standardised Information Sheet

SJR: SCImago Journal Rank

SQRT: Square Root

SVM: Support Vector Machine

SMOTE: The Synthetic Minority Over-sampling Technique

TT(Sec): Training Time (in seconds) or Total Time (in seconds)

U.S.: United States

XAI: Explainable Artificial Intelligence

XGBOOST: eXtreme Gradient Boosting

CHAPTER 1

Introduction

Access to credit is a cornerstone of modern economies, enabling households and businesses to finance consumption, investment, and growth. At the center of this system lies the credit contract, an agreement between lender and borrower that requires careful alignment between the type of credit granted and the borrower's repayment capacity (Bank of Portugal, 2025[1]). Credit, however, remains one of the most profitable yet riskiest financial services (Kamalloo & Abadeh, 2014). Default risk, the possibility that a borrower fails to meet repayment obligations, makes accurate assessment essential for both lenders and regulators.

In view of this risk, it is essential for lenders to require adequate guarantees to secure the fulfillment of the financial obligation. Conversely, borrowers are expected to demonstrate their capacity to repay the total amount borrowed. Credit approval requires financial institutions to assess the applicant's ability to meet repayment obligations. The specific steps, data inputs, and regulatory criteria involved in this process are discussed, with emphasis on definitions and standards set by the Bank of Portugal.

Traditional credit analysis, while widely used, has significant limitations. Conventional models often struggle to capture the complexity of borrower behavior, leading to inaccurate classifications, exclusion of certain populations, and reduced transparency in decision-making (Yao *et al.*, 2022; Rajesh *et al.*, 2023). These shortcomings particularly affect the so-called underbanked or "thin-file" borrowers: individuals with minimal credit histories who are not inherently high-risk but lack the traditional indicators used in conventional scoring systems. As a result, they are often denied credit despite some potential repayment capacity (Muñoz-Cancino *et al.*, 2022).

The financial exclusion of these groups has prompted growing interest in alternative approaches. Mhlanga (2021) emphasizes that exclusion is not only a matter of repayment risk but also of information asymmetry, limited collateral, and structural barriers that disproportionately affect women, youth, and small businesses. Machine learning (ML) and artificial intelligence (AI) are identified as transformative tools capable of integrating alternative data sources, such as digital footprints, mobile usage, or social media interactions into credit scoring. By broadening the informational base, ML models can enhance predictive

[1] https://clientebancario.bportugal.pt/en/what-they-are-and-types-loans

accuracy, improve transparency, and expand access to credit (Muñoz-Cancino *et al.*, 2022; Mhlanga, 2021). At the same time, financial institutions and regulators are increasingly attentive to the risks of algorithmic bias. Without safeguards, ML systems may perpetuate or even exacerbate discrimination, undermining the very goal of inclusion (Kozodoi *et al.*, 2021). This makes interpretability and fairness central requirements. Explainable Artificial Intelligence (XAI) techniques, such as LIME and SHAP, have emerged to address the "black-box" problem of such algorithms, reinforcing trust and regulatory compliance (Yao *et al.*, 2022).

Against this backdrop, the present study explores the evolving landscape of credit assessment in Portugal, focusing specifically on personal (consumer) credit. It investigates how ML models can balance three key dimensions: predictive accuracy, interpretability, and social impact. Particular attention is given to underbanked populations, with the aim of identifying approaches that can improve credit access without compromising financial stability. Furthermore, the framework developed herein is intended to be adaptable to similar contexts across other European countries and potentially on a global scale. By examining the challenges currently faced by underbanked populations with credit institutions in Portugal, this research aims to foster critical reflection on opportunities for improvement, innovation, and targeted solutions that promote a more inclusive, yet less risky, credit lending practice.

The research design of this study follows the Cross-Industry Standard Process for Data Mining (CRISP-DM), a proven methodology that outlines the typical phases of a project, the tasks within each phase, and the relationships between them, thereby providing a structured framework for data-driven analysis (Tounsi, Anoun, & Hassouni, 2020). The methodology encompasses a comprehensive literature review, the use of a synthetic dataset inspired by real-world banking practices, constructed collaboratively by contributors on the Kaggle.com, data preparation and feature engineering, and the evaluation of multiple machine learning classifiers under both balanced and imbalanced target scenarios. Logistic regression is employed as a benchmark, with ensemble methods such as LightGBM, Random Forest, and AdaBoost tested as alternatives.

Although the study primarily focuses on the credit assessment process related to consumer credit within Portuguese banking institutions, it may also draw on references from mortgage credit practices or specific aspects of credit evaluation within the broader European Union context. While these adjacent topics may indirectly influence credit assessment predictions, they remain beyond the core scope of this research. It is also important to acknowledge that the study's findings are subject to limitations, particularly those arising from the availability and reliability of operational data, as well as the inherent uncertainty associated with credit risk modeling.

The structure of this dissertation is as follows: Chapter 2 presents the literature review conducted for this study, along with a comprehensive overview of the business context related to consumer credit assessment in Portugal. Chapter 3 explores the methodological framework applied in the practical component of this thesis, the characteristics of the dataset used, and the data preparation process. Chapter 4 presents and discusses the results, highlighting their implications for practice and theory. Finally, Chapter 5 concludes the study by summarizing key findings, reinforcing the relevance of the research, acknowledging its limitations, and offering recommendations for future work.

What distinguishes this study from prior research on credit risk assessment is its multidimensional approach, combining predictive accuracy, financial inclusion, and model transparency within a single framework. Although machine learning has already been applied to credit evaluation, much of the literature focuses primarily on technical performance while underscoring the difficulty of interpreting results from complex models. This study addresses that gap by employing the PyCaret framework, which enables systematic comparison across algorithms and enhances interpretability, thereby responding to one of the main criticisms of machine learning in this field. In addition, whereas many studies adopt a broad perspective on credit markets, often encompassing corporate lending, mortgage credit, or international settings, this research concentrates specifically on personal credit. This thematic narrowing provides a more precise examination of how advanced machine learning models can improve credit access for underbanked populations while ensuring results remain transparent, actionable, and aligned with the social and regulatory realities of the banking sector.

CHAPTER 2

Business Understanding

Understanding credit assessment and its full process is fundamental to laying the groundwork for examining the current practices adopted by banks and credit institutions. This chapter aims to provide a comprehensive overview of the credit assessment process and how it is conducted by Portuguese credit institutions in the context of personal loan approvals. For consistency with the terminology used by the Bank of Portugal, this type of credit will be referred to as consumer credit throughout the study. This chapter is organized into two main sections: (i) a review of academic literature on contemporary credit modeling techniques, risk mitigation, and financial inclusion; and (ii) a detailed examination of the credit assessment process, with particular emphasis on the regulatory frameworks shaping the Portuguese context, the current situation of the country's unbanked population, and how financial inclusion is addressed.

This knowledge is equally vital for formulating effective recommendations to improve the predictive accuracy of credit evaluations. Enhancing these capabilities enables institutions to better assess creditworthiness, increase profitability, refine client segmentation, and mitigate default and late-payment risks. Methods used to quantify default risk are then examined, establishing a baseline for model-improvement strategies, including identifying the most impactful features in loan decisions and exploring approaches that raise predictive accuracy without adding excessive complexity that would hinder interpretability.

The rise of Artificial Intelligence (AI) and Machine Learning (ML) has introduced transformative capabilities in credit risk assessment, improving predictive accuracy, automating decisions, and expanding financial access (Filchenkov *et al.*, 2021; Florez-Lopez & Ramon-Jeronimo, 2014). Advanced models, such as deep neural networks, ensemble methods, and hybrid algorithms, outperform traditional techniques by capturing nonlinear patterns in large, complex datasets (Robinson & Sindhwani, 2024). Emerging techniques such as federated learning enable collaborative training without compromising privacy (Oualid *et al.*, 2023), while Explainable AI (XAI) strengthens transparency and regulatory trust (Yao *et al.*, 2022).

As noted by Feng *et al.* (2019), the adoption of AI-driven models is motivated by their ability to uncover complex, nonlinear borrower behaviors, enabling more accurate, and potentially fairer credit decisions, particularly for underbanked populations. In this context, the goal of fair ML is to ensure predictions meet statistical fairness criteria, typically assessed via inter-group differences across sensitive attributes such as gender, religion, or ethnicity (Kozodoi *et al*., 2021). Without safeguards, algorithms may perpetuate structural biases and exclude the very groups targeted by financial inclusion initiatives. Models must therefore balance predictive performance with fairness constraints, and training-time adjustments that optimize fairness alongside conventional model-fit metrics are commonly referred to as fairness processors (Kozodoi *et al.*, 2021).

Taken together, these insights position the present study at the intersection of predictive accuracy, fairness, and financial inclusion. By focusing on personal credit in the Portuguese context, the research addresses a critical area of banking practice and contributes to broader debates on responsibly extending credit to underbanked populations, highlighting the dual role of AI and ML in improving credit risk management efficiency and advancing the societal objective of financial inclusion.

2.1. Literature Review

Analyzing the characteristics of credit granting, the impact of resulting decisions arising from it, and the role of Portuguese credit institutions requires preliminary research to frame the broader context and situate financial inclusion. This stage was guided by three key research questions, which directed the literature review process, including database selection, keyword definition, and inclusion and exclusion criteria for article selection. The research questions guiding this study are as follows:

1. Does the use of machine learning contribute to reducing the risk of default in financial institutions?
2. How can more machine learning models be implemented in the credit assessment process, in a way that maintains or improves decision-making effectiveness?
3. In what ways can the application of machine learning promote financial inclusion among underbanked populations?

By adopting this structured research approach, the chapter aims to provide a comprehensive overview of the state of the art on how machine learning techniques can improve credit assessment practices, mitigate default risk, and promote financial inclusion in the context of credit granting.

In this initial research phase, Scopus, Web of Science, and Google Scholar were chosen as the primary databases, given their recognized relevance in academic research and to ensure broad coverage of relevant documents. Scopus is widely used for its extensive scope and rigorous peer-review system, which guarantees quality and consistency (Soomro *et al.*, 2024; Oualid *et al.*, 2023). Web of Science (WoS) is valued for its stringent curation process and focus on high-quality publications, ensuring thematic relevance and minimizing bias (Sadok *et al.*, 2022). Google Scholar, in turn, offers accessibility and a vast repository of scholarly articles, although its inclusion of grey literature may compromise quality. To address this limitation, only the first 20 pages of results were considered, as the most relevant sources typically appear within them.

To enhance the research process and clarify the search criteria, three initial keywords were defined: Machine Learning, Avaliação de Crédito (Credit Assessment), and Inclusão Financeira (Financial Inclusion). While Machine Learning was already in English, the other two were initially in Portuguese to capture relevant work and documentation in the country's native language, given the study's focus on the Portuguese banking sector. A preliminary search was conducted to explore the relationship between these terms. However, most results appeared due to the keyword Machine Learning, with limited relevance to the other two concepts. This discrepancy highlighted the need to standardize the search language; therefore, all keywords were translated into English to improve the precision and relevance of the results, as shown in Table 2.1.

Table 2.1. Preliminary Terms

Topic	**Contains at least one of the following terms**
Machine Learning	machine learning
Credit Assessment	credit OR evaluation OR risk
Financial Inclusion	financial OR inclusion

After incorporating English expressions into the search process, the results became more comprehensive. A preliminary review of titles, abstracts, and keywords from the initial articles also revealed additional relevant terms previously overlooked, such as creditworthiness, credit scoring, loan evaluation, default risk, and fairness. These findings refined and expanded the list of search terms used in the final research query, as presented in Listing 2.1.

```
("credit risk" OR "loan evaluation" OR "credit scoring" OR
"creditworthiness" OR "creditworthiness assessment" OR "default risk")
AND ("machine learning" OR "artificial intelligence") AND ("bank
customers" OR "banks" OR "financial institutions") AND ("financial
inclusion" OR "inclusion" OR "fairness”)
```

Listing 2.1. Query used for database searches.

For the purposes of this study, articles published within a ten-year period, from 2014 to 2024, were considered. The choice of this timeframe reflects the period following the recovery from the global financial crisis of 2008, which significantly disrupted credit markets and financial stability.

For this study, articles published between 2014 and 2024 were considered. This timeframe reflects the period after the 2008 global financial crisis, which severely disrupted credit markets and financial stability. In response to the subprime crisis, the European Council[2] introduced international standards such as Basel III, aimed at strengthening banking resilience, stabilizing credit dynamics, and reinforcing regulatory frameworks. These rules ensure that European banks can continue operating during economic shocks while supporting sustainability and preparing for the green and digital transitions, an essential step toward consolidating the Banking Union. As noted by Orlova (2021), these frameworks not only reinforced capital adequacy requirements but also required improvements in risk management practices and IT systems to enhance financial stability.

The Zotero reference management tool was used to organize the documents retrieved from the selected databases, supporting duplicate identification, screening, and review. Regarding the inclusion criteria for the final selection of articles are detailed in Table 2.2.

[2] https://www.consilium.europa.eu/en/press/press-releases/2024/05/30/basel-iii-reforms-new-eu-rules-to-increase-banks-resilience-to-economic-shocks/

Table 2.2. Inclusion and Exclusion Criteria for Literature Selection

Criterion	**Justification**
The study must focus on the dynamics of credit granting by commercial banks.	Excluding topics such as portfolio management, financial investments, or insurance ensures that the analysis remains focused on credit granted by commercial banks. Furthermore, avoiding studies related to credit scoring for credit card issuance helps to maintain the study's scope.
Practical application of the proposed models.	Studies that successfully tested the proposed models and reported the results contribute to the practical validation and relevance of the evaluated methods.
Studies that focus exclusively on traditional assessment methods were excluded.	The article must present solutions that incorporate machine learning and artificial intelligence methods, as these align with the objective of advancing the state of the art.
Exclusion was applied to studies that focused on areas unrelated to individual credit.	Studies analyzing credit granting for specific sectors that fall outside the defined scope, such as agricultural, corporate credit or mortgages, were excluded. This ensures that the focus remains on individual credit.
SCImago Journal Ranking (SJR).	For journal articles, the ranking provided by the SCImago Journal Rank (SJR) was used, with priority given to those classified in Quartiles 1 and 2 within categories such as Computer Science, Artificial Intelligence, and Finance.

As of December 2024, the search process yielded 607 records from Scopus, 301 from Web of Science (WoS), and 250 from Google Scholar, the latter corresponding to all articles retrieved from the first 20 pages of results. After duplicate removal, the remaining articles underwent a structured screening protocol: title assessment, keyword inspection, and abstract analysis. This filtering produced 208 candidate articles. Their introductions were then examined in detail, and studies not meeting the inclusion criteria were excluded, leaving 147 peer-reviewed articles eligible for review. From these, 37 articles were prioritized for full-text analysis based on scientific relevance and journal quality, specifically those published in SCImago Journal Rank (SJR) Quartiles 1 and 2. The final corpus includes studies on machine learning techniques for credit risk assessment, credit scoring, and financial inclusion, ranging from conventional algorithms to advanced ensemble models and explainable AI frameworks. The complete screening and selection workflow is shown in Figure 2.1.

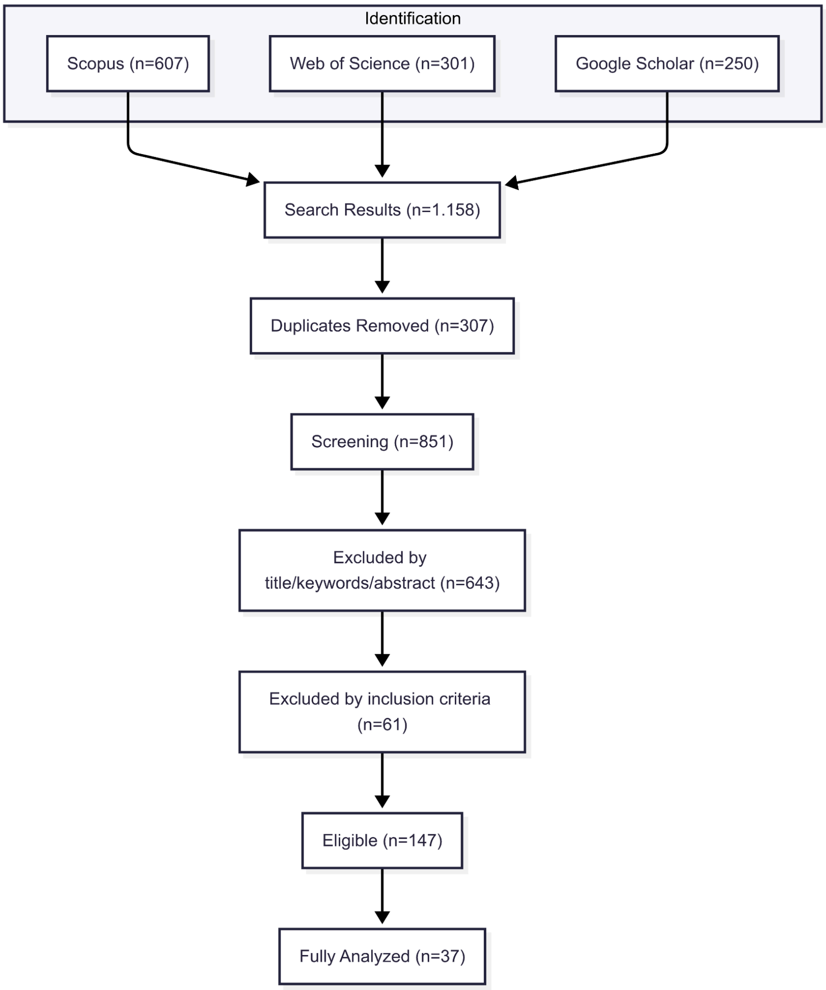

Figure 2.1. Diagram of the document collection and selection process

This literature review underscores the critical role of credit granting in the financial ecosystem, demonstrating its significant impact on multiple stakeholders. For financial institutions, especially banks, it remains a primary source of revenue. At the same time, credit is a vital financial resource for individuals, serving as an indicator of economic growth, particularly in developing economies, by enabling access to liquidity for personal needs, investments, or projects otherwise unattainable with available resources (Muñoz-Cancino *et al.*, 2022). Recent studies highlight the increasing relevance of advanced machine learning techniques, such as deep neural networks, ensemble methods, and hybrid approaches, in improving credit risk prediction. Robinson and Sindhwani (2024) show that Random Forest and Gradient Boosting outperform traditional models in capturing complex financial behaviors, while Oualid *et al.* (2023) demonstrate that federated learning preserves privacy without compromising predictive performance. Other researchers, including Yao *et al.* (2022) and Filchenkov *et al.* (2021), stress the importance of Explainable Artificial Intelligence (XAI) as a key factor in fostering trust among decision-makers.

Despite these advancements, significant challenges remain. ML models are often limited by algorithmic bias (Li *et al.*, 2023; Florez-Lopez & Ramon-Jeronimo, 2014), and their applicability in emerging markets, where structured financial data is scarce, remains underexplored. To address these gaps, Filchenkov *et al.* (2021) highlight techniques such as autoencoders and semi-supervised learning as promising approaches to mitigate data scarcity and improve model robustness.

The subsequent sections provide a detailed examination of the core concepts from the reviewed literature, with attention to technological developments, operational implementations, persistent challenges in credit risk assessment frameworks, as well as the question of fairness within this process. The analysis covers the methodologies underlying both traditional and contemporary credit scoring systems, emphasizing their predictive capabilities, structural assumptions, and practical limitations. Particular focus is given to state-of-the-art machine learning approaches, especially regarding data sparsity, heterogeneity, and algorithmic bias. Building on this foundation, the discussion turns to the specific characteristics of credit assessment in the Portuguese financial sector, evaluating how advanced technologies could be calibrated to operate under conditions of limited data availability and infrastructural variability. By situating these insights within the broader discourse on financial inclusion for underbanked populations and institutional risk management, the analysis aims to provide a rigorous and context-aware perspective on the current and future trajectory of credit risk modeling.

2.1.1. State of the Art in Credit Risk Modeling with Machine Learning

Credit risk prediction is a fundamental element of decision-making in the financial sector, enabling institutions to estimate the likelihood of borrower default with precision. In recent years, machine learning (ML) techniques have gained prominence due to their capacity to process high-dimensional data and capture complex, nonlinear relationships among variables. Conventional approaches, however, often rely on single classifiers, which struggle to represent the intricate nature of financial behavior, resulting in limited generalizability and reduced predictive accuracy. A further challenge in credit risk modeling is the class imbalance common to most datasets, where non-defaulting borrowers greatly outnumber defaulters. This imbalance can bias models toward the majority class, reducing sensitivity to the minority class, the actual defaults. To mitigate this issue, studies have explored methodological strategies such as resampling techniques (e.g., Synthetic Minority Over-sampling Technique – SMOTE), ensemble learning architectures like Random Forest and AdaBoost, and interpretability techniques such as Shapley Additive exPlanations (SHAP), aimed at improving both predictive performance and transparency while ensuring adequate representation of borrower profiles (Aruleba & Sun, 2024).

Building on these advancements, recent studies show that machine learning models such as Naive Bayes (NB), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees (DT), and K-Nearest Neighbors (KNN) consistently outperform traditional statistical methods like Logistic Regression (RL), which struggle with nonlinearities and complex interactions (Feng *et al.*, 2019). Techniques including decision trees, neural networks, and ensemble classifiers are widely adopted in high-dimensional contexts for their robustness and adaptability (Moral-Garcia & Abellan, 2023; Xu *et al.*, 2020). Deep learning approaches, in particular, demonstrate significant improvements in default prediction by leveraging both historical and behavioral data (Tripathi *et al.*, 2021; Bulut & Arslan, 2024), while also enhancing risk segmentation and adapting to different data availability contexts. Hybrid models that combine supervised learning with other techniques have proven effective, especially with limited data or low-default portfolios (Kamalloo *et al.*, 2014). In large-scale scenarios, deep learning systems outperform other approaches, offering improved scalability and accuracy (Yemmanuru *et al.*, 2024). As highlighted by Moral-Garcia and Abellan (2023), classifier diversity is critical to boosting performance, with decision trees standing out for their ability to incorporate probabilistic uncertainty into predictions.

As highlighted by Aruleba and Sun (2024), despite the notable performance gains of machine learning models in credit risk prediction, a critical limitation remains: their lack of interpretability. The opaque or “black-box” nature of many algorithms, particularly deep learning models, challenges the financial sector, where transparency and accountability are essential for regulatory compliance and stakeholder trust. With regard to deep learning models, Muñoz-Cancino *et al.* (2022) also call attention the execution time and computational complexity, noting that applying their proposed Graph Neural Network (GNN) methodology, a type of deep learning model, to the datasets required about 15,500 minutes of processing, due not only to the large dataset but also to the inherent complexity of such algorithms.

Understanding how a model generates predictions is fundamental for credit assessment. To address this, explainable artificial intelligence (XAI) techniques have emerged to enhance interpretability. Although early approaches to rule extraction in neural networks date back to the 1990s, XAI gained momentum from 2014 onwards with methods such as LIME, reinforced by the GDPR (2016/2018), which introduced the notion of a “right to explanation” (Guidotti *et al.*, 2018).

Overall, these approaches allow researchers and practitioners to understand the reasoning behind predictions, clarifying the features and patterns most relevant to credit risk assessment. By combining resampling strategies, ensemble methods, and XAI frameworks, financial institutions can improve predictive performance while ensuring models remain transparent, interpretable, efficient, inclusive, and aligned with industry standards and goals.

2.1.2. Risk Mitigation in Credit Scoring Models

Following the weaknesses exposed in traditional credit assessment during the 2008 subprime crisis and the subsequent implementation of Basel III, risk management became a central priority for commercial banks. These reforms laid the foundations for transforming risk evaluation by mandating stricter capital requirements and more robust supervisory practices (Orlova, 2021; Merćep *et al.*, 2021). In turn, such developments accelerated the adoption of advanced modeling techniques, most notably machine learning and deep learning, to mitigate systemic vulnerabilities and enhance credit decision-making accuracy.

Tripathi *et al.* (2021) emphasize the critical role of credit risk analysis, noting that statistical and machine learning methods are now widely applied to evaluate default probability using applicants’ historical data. Beyond model refinement, reducing classification errors remains a priority. Roy and Shaw (2023) and Liang *et al.* (2018) highlight the financial risk of Type I errors, when defaulting clients are incorrectly deemed creditworthy, while minimizing Type II errors is essential to avoid excluding creditworthy individuals, thereby promoting fairness and financial inclusion.

Innovations in data sources have also emerged as a complementary avenue for advancing credit risk modeling. Integrating alternative data, such as digital footprints, behavioral patterns, and mobile activity, provides richer insights into creditworthiness. Ala'raj *et al.* (2022) demonstrate this with a Long Short-Term Memory (LSTM) model for default risk and consumer spending prediction, while Simumba *et al.* (2021) show that incorporating mobile and public data significantly improves predictive accuracy and reduces error rates. In contexts where privacy concerns limit access to real financial data, Muñoz-Cancino *et al.* (2022) propose using synthetic data generated by Generative Adversarial Networks (GANs) and variational autoencoders. Although synthetic data may slightly reduce performance, it enables the development of scalable, privacy-compliant models suitable for highly regulated environments. This aligns with the present research, which also employs a publicly available synthetic dataset. By doing so, the study follows established academic practices while ensuring compliance with privacy constraints that restrict access to Portuguese banking data, thereby maintaining both methodological rigor and regulatory sensitivity.

2.1.3. Financial Inclusion through Machine Learning

Mhlanga (2021) provides a comprehensive overview of evolving definitions of financial inclusion and exclusion, emphasizing their relevance in addressing socioeconomic disparities. Financial inclusion encompasses not only affordability but also accessibility, availability, quality, and dignity of services, while financial exclusion refers to systemic barriers, economic, regulatory, or institutional, that prevent individuals and groups from accessing mainstream financial services. The European Commission further notes that product design and legal frameworks can unintentionally perpetuate exclusion. The urgency around financial inclusion intensified in the early 2000s due to its link to poverty reduction and inequality, spurring global initiatives such as the United Nations' call for accessible, affordable, and sustainable financial services.

Building on this foundation, Kozodoi *et al.* (2022) examine inclusion through the lens of algorithmic fairness in credit scoring, arguing that fairness criteria such as independence, separation, and sufficiency are essential to avoid discriminatory outcomes in automated credit decisions. Their work illustrates the potential of responsible AI to expand access to financial products among marginalized populations. Similarly, Moldovan (2023) demonstrates that financial models can balance profitability with fairness, reducing bias without undermining compliance or returns.

Ponsam *et al.* (2021) address the challenge of evaluating creditworthiness for individuals with limited or no credit history, a common obstacle to inclusion. They advocate the use of advanced machine learning models such as Light Gradient-Boosting Machine (LightGBM), which outperform traditional algorithms like Logistic Regression (LR) and Support Vector Machines (SVM) in both accuracy and robustness. By generating more reliable and equitable predictions, LightGBM strengthens the ability of financial institutions to make fairer lending decisions, thereby advancing financial inclusion (Mhlanga, 2021).

2.2. The Credit Assessment Process in Portugal

As previously mentioned, a clear understanding of the credit assessment process is essential for identifying gaps and proposing meaningful insights. This section examines the foundational definitions and regulatory frameworks that govern credit evaluation, followed by a contextualization of the Portuguese case. This approach clarifies the types of credit, the assessment criteria applied, and the dynamics of the evolving market.

In Portugal, credit assessment practices are regulated by national laws aligned with European standards, particularly the Consumer Credit Directive (2008/48/EC) and the Mortgage Credit Directive (2014/17/EU), while also reflecting national factors such as consumer behavior and macroeconomic conditions. The Bank of Portugal plays a central role by issuing regulatory guidance and publishing statistics on new loans, segmented by purpose (e.g., consumption, housing), type (e.g., personal, auto, revolving), maturities, interest rates, and financial inclusion indicators.

These data provide visibility into credit allocation nationwide, allowing institutions to benchmark practices and regulators to monitor systemic risk. In addition, the Bank of Portugal publishes reference rates and market averages, which are used as a safeguard against usurious lending practices and play a critical role in shaping responsible credit assessment policies.

2.2.1. Definition and Regulatory Framework

As part of the credit assessment process, credit scoring serves as a fundamental tool for evaluating whether a consumer can repay a loan in full and on time. Within this process, credit scoring is a fundamental tool for assessing borrower risk. Tripathi *et al.* (2021) define credit scoring as "*a set of decision models and their underlying techniques that aid credit* *lenders in the granting of credit.*” These models isolate the impact of applicant characteristics, such as past defaults or financial behavior, to determine whether a candidate belongs to a trustworthy or non-trustworthy group. The amount of credit extended is typically based on indicators such as annual income, declared assets, and other financial credentials. From a modeling perspective, effective feature selection is critical, since eliminating irrelevant or redundant data improves accuracy and reduces complexity.

Beyond risk estimation, credit scoring supports managerial decision-making, strengthens cash flow management, and helps calculate and mitigate credit risk. As Tripathi *et al.* (2021) emphasize, the performance of scoring models is directly linked to the profitability and long-term sustainability of credit institutions.

Credit scoring is not a single-stage procedure but a multi-phase evaluation strategy used by financial institutions to manage different aspects of risk. According to Tripathi *et al.* (2021), these stages include:

- Application Scoring: Evaluates new applicants based on social, financial, and demographic data provided during the application, helping determine authenticity and detect early risks.
- Behavioral Scoring: Applied to existing customers, it analyzes financial behavior, such as payment history and spending patterns, supporting dynamic portfolio management and individualized strategies.
- Collection Scoring: Applied to existing customers, it analyzes financial behavior, such as payment history and spending patterns, supporting dynamic portfolio management and individualized strategies.
- Fraud Detection: Estimates the likelihood of fraudulent behavior, enabling early identification and mitigation of dishonest or suspicious activity.

This layered scoring approach improves risk assessment accuracy and supports more informed, data-driven decisions throughout the credit lifecycle (Tripathi *et al.*, 2021).

In the Portuguese context, the Bank of Portugal[3] requires that creditworthiness assessment be based on documented proof of income and expenditure for the three months preceding the application. Income includes wages, pensions, or social benefits, while expenses cover essential living costs and obligations such as existing loans. Importantly, lenders are prohibited from assuming future income increases or expenditure reductions when calculating repayment capacity. Additionally, the Bank of Portugal requires institutions to comply with a

[3] https://clientebancario.bportugal.pt/en/creditworthiness-assessment

set of prudential macro-rules, which apply regardless of individual borrower assessments. These rules establish limits on the loan-to-value (LTV) ratio, the debt service-to-income (DSTI) ratio, and the maximum loan maturity.

The LTV ratio represents the proportion between the loan amount and the value of the mortgaged property. The loan amount should not exceed 90% of the property's value for loans aimed at permanent residence, 80% for loans financing other types of property, and 100% in cases involving movable property held by credit institutions or financial leasing agreements. The DSTI ratio measures the borrower's total monthly loan installments relative to their income, stipulating that the aggregate of these installments, considering both the new credit and any existing loans, should not, as a rule, exceed 50% of the borrower's monthly income net of taxes and social security contributions. Finally, the maturity of loans is also regulated, with residential mortgage contracts generally limited to an average maturity of 30 years, subject to age-related caps (40 years for borrowers aged 30 or under, 37 years for those between 30 and 35, and 35 years for borrowers older than 35). For personal credit, maturities must not exceed seven years, while loans for education, health, renewable energy, or cars are capped at 10 years, provided that such purposes are duly evidenced.

These safeguards promote responsible lending and mitigate systemic financial risk. Even borrowers with strong repayment capacity cannot legally obtain credit beyond these thresholds. In specific cases, such as low-value credit (e.g., overdrafts below ten times the minimum monthly wage or short-term credit increases), banks may apply simplified assessment methods using external data sources. However, this does not exempt them from the obligation to conduct risk analysis responsibly and transparently.

2.2.2. Consumer Credit in Portugal: Classification and Characteristics

According to the Bank of Portugal, credit contracts granted by financial institutions are classified into two primary categories: mortgage credit and consumer credit. This distinction is essential for regulatory and risk assessment purposes.

Mortgage credit refers to loans secured by real estate, typically for the acquisition, construction, or renovation of housing. These contracts are long-term and involve a mortgage guarantee, requiring stricter risk analysis due to higher loan amounts and extended repayment periods. In contrast, consumer credit and revolving credit lines usually have shorter maturities and higher interest rates, demanding stricter affordability assessments. Mortgage loans, by involving collateralized risk and longer repayment obligations, are therefore subject to different evaluation standards.

In Portugal, interest rates[4] on credit and deposit products are subject to strict regulatory oversight by the Bank of Portugal. The gross annual nominal interest rate expresses the remuneration of a deposit or the cost of borrowing over one year, without accounting for inflation or taxes. It is a nominal (not inflation-adjusted), gross (excluding taxes), and simple (not capitalized) rate. In practice, the Annual Percentage Rate of Charge (APRC) or TAEG under Portuguese law, captures the effective cost of credit by incorporating not only the nominal interest rate but also fees, taxes, and mandatory insurance. Since 2010, the Bank of Portugal has set and updated quarterly maximum APRC thresholds for each category of consumer credit, ensuring compliance with legal limits. These measures, combined with taxation on interest (e.g., 28% IRS for individuals in mainland Portugal and Madeira; 25% IRC for companies), aim to promote transparency, protect consumers, and curb abusive lending. This regulatory framework directly shapes credit risk assessment by influencing affordability and borrower solvency profiles.

While both categories are relevant for regulatory oversight and financial stability, this subsection focuses on consumer credit, given its direct importance for financial inclusion, risk assessment, and short- to medium-term personal financing.

Consumer credit covers agreements financing personal needs unrelated to real estate acquisition or investment. Under Portuguese law, it includes:

- Loans to private individuals amounting to between EUR 200 and EUR 75,000;
- Credit overruns (i.e., overdrafts), even if the amount is less than EUR 200;
- Loans for works on real estate, provided there is no mortgage or other right on immovable property, even when the amount exceeds EUR 75,000.

In practical terms, consumer credit encompasses several unsecured financial products, such as:

- Personal loans: for general-purpose spending, including education, healthcare, or travel;
- Credit cards: revolving credit lines with flexible repayment terms;
- Auto loans: loans dedicated to vehicle purchases;
- Overdraft facilities: short-term credit linked to current accounts;
- Revolving credit lines: reusable credit with predefined limits.

However, the legal framework excludes certain types of agreements from being considered consumer credit, such as:

[4] https://clientebancario.bportugal.pt/en/interest-rates

- Loans secured by a mortgage or other rights over immovable property;
- Loans whose purpose is the acquisition or maintenance of property rights over real estate;
- Credit granted by pawnshops;
- Leasing contracts without an option or obligation to purchase the asset;
- Credit granted free of interest or charges;
- Credit granted by employers to employees with interest rates below market levels.

All consumer credit types are subject to provisions on cost transparency, interest rate disclosure, and fair lending practices. Because they are generally unsecured, these contracts carry higher interest rates, which are closely monitored by the Bank of Portugal to prevent abusive pricing. Additional distinctions, such as loan duration, purpose, or type of lender (bank vs. credit institution) further refine the classification and are addressed in consumer protection materials published on the Bank of Portugal's public information portal.

2.2.3. Creditworthiness Assessment Criteria

Creditworthiness assessment is a legal obligation for all financial institutions in Portugal, ensuring that borrowers can realistically repay their debts without becoming over-indebted.

According to the Bank of Portugal, the main factors evaluated in the assessment of creditworthiness include the borrower's:

- Income stability and origin (e.g., salary, pensions, freelance work)
- Monthly fixed expenses, such as rent or other credit obligations
- Household structure, including the number of dependents
- Credit history and prior defaults
- Employment status and contract type
- Age and remaining years to retirement

A central metric in this evaluation is the Debt Service-to-Income (DSTI) ratio, which measures the share of monthly income used to service debt. The Bank of Portugal recommends a DSTI cap of 50%, including the new credit requested. This limit helps shield consumers from financial stress while supporting system stability. Institutions must also conduct interest rate stress tests, particularly for variable-rate products, to assess whether borrowers remain solvent under adverse scenarios.

Portuguese law further establishes a comprehensive set of consumer rights in credit agreements, designed to guarantee informed decision-making and prevent over-indebtedness or abusive lending.

Before a loan is granted, the institution must provide:

- A Standardised Information Sheet (SIS) for consumer credit, or ESIS for mortgage credit.
- A loan proposal with clear terms: amount, interest rate, repayment period, associated fees, and the Total Cost of Credit.
- Simulations, including worst-case scenarios for interest rates, when applicable.

Consumers must also be given:

- A reflection period before contract signing.
- The right to early repayment, in whole or in part.
- Access to support mechanisms in case of financial hardship, including renegotiation or restructuring options.

Banks are prohibited from encouraging borrowing beyond the client's capacity, and they must justify rejections or limits in a clear and objective manner. The Bank of Portugal provides ongoing supervision to ensure that these duties are respected in practice.

2.2.4. Underbanked Populations and Financial Inclusion

According to Muñoz-Cancino *et al.* (2022), underbanked populations, often called "*thin-file*" borrowers, have limited or no credit history, making their creditworthiness difficult to assess and frequently leading to financial exclusion. This view is reinforced by Mhlanga (2021), who highlights that underbanked individuals, particularly in emerging markets, lack access to formal financial services due to insufficient collateral or identification, thus requiring the use of alternative data and advanced analytics to reduce informational asymmetries. In the same line, Kozodoi *et al.* (2021) emphasize the importance of fairness in algorithmic credit scoring, noting that ML-based risk models must be carefully designed to avoid perpetuating discrimination against disadvantaged groups.

These concerns align with the broader international definition of financial inclusion. The World Bank defines it as the situation where individuals and businesses have access to useful and affordable financial products that meet their needs delivered responsibly and sustainably (World Bank, 2025). Similarly, the United Nations highlights financial inclusion as a key enabler for achieving multiple Sustainable Development Goals (UN, 2020), particularly by reducing inequalities, poverty, and fostering access to economic opportunities. At the same time, the European Commission (2010) report highlights that financial inclusion is intrinsically linked to access to credit facilities, which play a central role in enabling individuals to participate fully in economic and social life. Financial exclusion, defined as the inability to access "*financial services in the mainstream market that are appropriate to people's needs and enable them to lead a normal life in the society in which they belong*", arises when households face barriers to obtaining bank accounts or credit. The report further stresses that restrictive credit conditions, coupled with risk-averse banking practices, can reinforce social exclusion by denying services to individuals with low income or limited deposits. Such restrictive conditions and risk-averse practices can reinforce social exclusion by denying services to low-income or low-deposit individuals, creating systemic risks as conservative risk management inadvertently excludes vulnerable groups.

In Portugal, underbanked populations are documented in the Bank of Portugal's 2021 Report on Financial and Digital Inclusion, based on 2020 data. While complete exclusion (level 1) is limited, a significant share of individuals fall into level 3, meaning they hold and actively use a bank account but lack access to other financial products such as credit, insurance, or investments. This group represented 17.4% of the population in 2020, showing that access to banking services does not necessarily translate into broader participation. The report further notes that sub-inclusion is more prevalent among lower-income, less-educated, and older individuals, as well as in rural regions. Access to credit remains decisive: while mortgage and consumer credit are concentrated among households with stable employment and income, those with precarious jobs, variable earnings, or thin credit files face persistent barriers, remaining underbanked despite having bank accounts.

Initiatives such as the Basic Bank Account (Conta de Serviços Mínimos Bancários, CSMB) have reduced the unbanked population, recognized by 51% of respondents in 2020 compared to 1.4% in 2010, but primarily guarantee deposits and payment services. They do not address the challenge of expanding access to credit. The 2021 report shows that while the unbanked population is relatively small (9.1% in 2020), many individuals remain limited to basic banking services. Thus, the central issue has shifted from eradicating unbanked status to promoting fair and sustainable access to additional services, particularly credit, which is essential for full financial inclusion.

As seen, conservative risk management approaches by banks may unintentionally exclude vulnerable groups from the formal financial system. In this regard, Moldovan (2023) argues that fairness in ML offers a framework to mitigate biases by embedding fairness constraints directly into models. Rather than relying exclusively on historical data, which often encodes past discrimination, ML fairness interventions can balance predictive accuracy with equitable treatment. As shown by Oualid *et al.* (2023) and Filchenkov *et al.* (2021), ML techniques are capable of detecting patterns in sparse or fragmented datasets while maintaining robustness and fairness. This not only reduces the risk of reinforcing exclusion but also strengthens the sustainability of credit systems by aligning profitability with social responsibility.

In sum, global and Portuguese evidence highlights the urgency of inclusive credit assessment strategies, positioning Portugal as a relevant case for other European nations facing similar challenges. The integration of underbanked populations into the formal system, when conducted fairly and effectively, can foster growth, reduce inequality, and support a more resilient credit ecosystem.

Research shows that institutions increasingly adopt hybrid models, combining traditional financial metrics with behavioral and alternative data to better assess underbanked populations. At the same time, debates on responsible AI and fairness metrics have intensified, with European regulatory guidance stressing explainability, auditability, and bias mitigation in automated decisions. Together, these developments suggest that by 2025, inclusive credit assessment is moving beyond the experimental stage to become a mainstream practice with significant implications for both financial stability and social equity

CHAPTER 3

Methodology

This chapter presents the methodology adopted for the study, taking the investigation conducted thus far as the basis for the practical steps applied. It begins with the overall research approach chosen for the practical component, followed by a detailed description of the dataset, its origin, structure, and limitations. The chapter then outlines the initial stages of data exploration and preparation, as well as the analytical techniques and tools employed. Collectively, these components provide the foundation for the subsequent development and evaluation of credit risk models.

3.1. Cross-Industry Standard Process for Data Mining (CRISP-DM)

For the development of this study, the CRISP-DM methodology was selected as the guiding framework. According to Tounsi, Anoun, & Hassouni (2020), the Cross-Industry Standard Process for Data Mining (CRISP-DM) is a robust and widely adopted methodology that structures data mining initiatives through a comprehensive, cyclical sequence of phases. It comprises six well-defined stages, business understanding, data understanding, data preparation, modeling, evaluation, and deployment, each with specific objectives and interrelated tasks that ensure methodological rigor and traceability throughout the analytical process. By systematically integrating domain-specific knowledge with data-driven techniques, CRISP-DM supports the iterative refinement of models in alignment with evolving business goals and empirical findings.

CRISP-DM was selected not only for its methodological maturity and cross-domain applicability, but also for its ability to foster transparency, reproducibility, and structured decision-making. This structured approach proved particularly valuable for maintaining analytical coherence, mitigating model risk, and ensuring that insights could be effectively translated into actionable outcomes within the broader context of credit risk assessment.

For the purposes of this study, the CRISP-DM framework was implemented as follows:

- Business Understanding: This phase focused on defining credit assessment risk and contextualizing it within the Portuguese consumer credit market, with particular attention to underbanked populations. The objective was to frame the research problem not only in terms of predictive accuracy, but also in relation to regulatory requirements, fairness considerations, and the broader goal of financial inclusion. By aligning the study with the Bank of Portugal's creditworthiness assessment criteria and international definitions of financial inclusion, this stage established the foundation for connecting methodological rigor with the social and institutional realities of credit evaluation in Portugal.

- Data Understanding: A publicly available synthetic dataset from Kaggle.com was employed as an alternative, given the restrictions on accessing real banking data. A descriptive analysis was then performed to identify key characteristics of the raw data, providing insights into its structure, distributions, and limitations. These findings guided the subsequent Data Preparation phase, ensuring that the dataset was not only suitable for credit risk modeling but also, as far as possible, included variables aligned with the creditworthiness assessment criteria defined by the Bank of Portugal.
- Data Preparation: Missing values were treated, temporal and income features categorized, and outliers and asymmetries corrected to handle distortions that could bias the models. As part of this phase, a metadata summary was constructed to document data types, missing values, and distributional characteristics, supporting informed decisions in later stages of preparation and modeling.
- Modeling: A set of machine learning classifiers was evaluated under both balanced and imbalanced target scenarios using the PyCaret library. Based on comparative results, the five best-performing models were selected and re-implemented with their native libraries (scikit-learn[5] and LGBMClassifier[6]) to ensure greater flexibility and generalization.
- Evaluation: Model performance was assessed not only through accuracy and recall, but also in terms of interpretability and fairness. This involved identifying the variables that most influenced each model's outcomes and ensuring that the minority class was adequately represented, an essential consideration in credit risk assessment, where real-world data are often skewed toward non-default cases. As part of this phase, and later expanded in the Results and Discussion, an exercise was conducted to map the dataset variables against the creditworthiness assessment criteria of the Bank of Portugal. This comparison sought to establish alignment between the variables available in the synthetic dataset and those commonly used by Portuguese banks, thereby strengthening the contextual relevance of the findings.

[5] https://scikit-learn.org/stable/

[6] https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html

Although not executed in a production environment (Deployment), the study provides practical recommendations illustrating how machine learning can complement traditional approaches to promote more inclusive credit evaluation in Portugal.

3.2. Credit Data

This study draws on a synthetic dataset based on real-world credit data from LendingClub[7], a U.S.-based peer-to-peer (P2P) lending platform that connects individual investors and borrowers through an online marketplace. Publicly accessible via Kaggle[8], the dataset offers substantial granularity while preserving privacy, as it reproduces patterns from actual borrower and loan information without exposing sensitive details. It covers a broad range of borrower- and loan-level features, including personal and financial characteristics as well as detailed loan attributes, thereby enabling comprehensive assessments of creditworthiness and repayment performance over time. Although originally developed in a fintech-oriented context, the dataset is here reinterpreted through the lens of a traditional commercial banking institution. Platform-specific mechanisms, such as investor bidding behavior or funding dynamics, are deliberately abstracted to enable a more institutionally neutral analysis centered on borrower risk.

The dataset comprises 466,285 loan applications issued between 2007 and 2014, with 75 original variables. These capture borrower demographics (e.g., annual income, employment length, house ownership), credit history markers (e.g., delinquency events, public records, revolving balance), loan parameters (e.g., amount requested, interest rate, term, purpose), and repayment outcomes (e.g., loan status, total payments received, number of late payments). This multidimensional structure provides a robust empirical basis for both descriptive analytics and supervised learning models aimed at default prediction. The complete list of variables, along with their definitions and data types, is provided in Table A.1 of Appendix A. To ensure clarity and maintain consistency with the dataset, the variables definitions were taken directly from Kaggle's discussion forum[9], where they originate from the LendingClub's platform itself.

Although originating from the U.S.-based peer-to-peer lending platform, the dataset was adopted in this study primarily due to the unavailability of Portuguese banking data at the customer level, which remain strictly restricted by privacy and governance regulations. Its

[7] https://www.lendingclub.com

[8] https://www.kaggle.com/datasets/wordsforthewise/lending-club

[9] https://www.kaggle.com/datasets/wordsforthewise/lending-club/discussion/170691

public accessibility, granularity, and broad set of borrower- and loan-level variables provide a valuable foundation for methodological experimentation and learning. While the dataset cannot be considered fully representative of the Portuguese credit market, many of its core indicators such as income, loan amount, employment length, house ownership, and interest rate, are widely recognized in credit assessment practices, align with the creditworthiness assessment criteria defined by the Bank of Portugal (see Section 2.2.3), and are present in the dataset. Accordingly, this study recontextualizes the dataset in an exploratory manner, focusing on variables with institutional relevance and abstracting from platform-specific mechanisms that are not comparable to traditional banking practices. The intention is not to reproduce the Portuguese credit market in its entirety, but rather to simulate lending scenarios that enable the testing of machine learning models under criteria inspired by local regulatory principles.

3.2.1. Variables and Target Definition

To enhance clarity and interpretability, the dataset's 75 original variables were systematically grouped into conceptual categories reflecting borrower, loan, and repayment characteristics. This categorization supports analytical consistency and alignment with the evaluation dimensions typically applied by financial institutions. Developed specifically for this research, the grouping aims to facilitate understanding of the different types of information collected across the credit application and repayment cycle. The categories are as follows:

- Joint Applicant Related: Variables applicable only to joint applications, such as combined income and debt-to-income ratios.
- Inquiries and Account Openings: Indicators of recent borrower activity related to credit inquiries and newly opened accounts.
- Installment Loan Balances and Utilization: Information on outstanding installment debts and borrower utilization relative to original loan amounts.
- Revolving Credit (e.g., Credit Cards): Measures of credit card use and revolving account balances.
- Loan Metadata: Characteristics of the loan itself, such as amount, term, interest rate, and installment value.
- Location and Identifiers: Internal platform variables and anonymized geographic indicators.
- Payment and Outstanding Amounts: Historical records of principal, interest, and late payments made by the borrower.

- Delinquency and Derogatory: Records of negative credit events and derogatory marks, including collections, public records, and delinquencies.

The target variable selected for this study is *loan_status*, originally a multi-class feature with the following categories: “Fully Paid”, “Charged Off”, “Default”, “Current”, “In Grace Period”, “Late (16–30 days)”, “Late (31–120 days)”, “Does not meet the credit policy. Status: Fully Paid”, and “Does not meet the credit policy. Status: Charged Off”.

To ensure clarity and maintain consistency with the dataset characteristics, the following definitions provided by LendingClub[10] describe each loan status category as recorded in the data used for this study:

- **Fully Paid**: The borrower has fully repaid the loan without outstanding obligations.
- **Charged Off**: LendingClub classifies a loan as charged off when it no longer reasonably expects any further payments from the borrower.
- **Default**: A loan is classified as in default when it has been delinquent for more than 120 days.
- **Current**: The borrower is up to date with scheduled repayments, and the loan remains in good standing.
- **In Grace Period**: The borrower has missed a scheduled payment but remains within a short, contractually defined 15-day window where late fees or default classification have not yet been applied.
- **Late (16–30 days)**: The borrower is delinquent, with payment overdue between 16 and 30 days past the due date. This is the first tranche of loan delinquency.
- **Late (31–120 days)**: The borrower is delinquent, with payment overdue between 31 and 120 days past the due date. This is the second tranche of loan delinquency.
- The last two categories “Does not meet the credit policy. Status: Fully Paid” and “Does not meet the credit policy. Status: Charged Off” are not explicitly defined in LendingClub’s documentation. In this study, their interpretation follows the final status indicated in each label, which corresponds to the definitions provided in the list above.

Although Bank of Portugal’s official materials do not specify precise time thresholds for when a borrower enters default, its late-payment framework provides conceptual guidance for this study. According to the Bank, when a customer fails to pay an installment on its due date,

[10] https://www.lendingclub.com/help/investing-faq/what-do-the-different-note-statuses-mean

financial institutions can charge default interest equal to the annual nominal interest rate plus a maximum surcharge of 3%, calculated daily over the overdue amount. They may also impose a single recovery fee, capped at 4% of the installment (minimum €12, maximum €150; or 0.5% if the installment exceeds €50,000). In parallel, Bank of Portugal requires credit institutions to initiate Out-of-Court Arrears Settlement Procedures (OASP)[11] when borrowers miss payments preferably within 15 days if under a Pre-Arrears Action Plan (PRAP), or by the 31st day otherwise, and no later than the 60th day of arrears￼. While this "timely basis" requirement does not specify exact cut-offs analogous to those in the LendingClub dataset (e.g., "Late 16–30 days," "31–120 days," "Default"), it reveals a similar segmentation of arrears management. In this context, those labels provided by LendingClub serve as a useful exploratory proxy, offering a structured basis for modeling classification timelines and credit risk, while acknowledging the inherent limitations when applying them to the Portuguese financial landscape.

For this study, during the data preprocessing phase, the multi-class target was aggregated into two categories to enable binary classification: "Paid" and "Default". The "Default" class encompasses loans with adverse repayment outcomes or considerable delinquency, including the original statuses: "Charged Off", "Default", "Late (16-30 days)", "Late (31-120 days)", and "Does not meet the credit policy. Status: Charged Off". The "Paid" class comprises loans marked as "Current" and "In Grace Period", indicating borrowers who are currently active and not severely delinquent.

Importantly, records categorized as "Fully Paid" and "Does not meet the credit policy. Status: Fully Paid" were excluded from the analysis. This decision avoids inflating the "Paid" class with fully matured contracts, which no longer represent ongoing credit risk. Their inclusion could bias models toward the majority class and reduce sensitivity to borderline or emerging defaults, which are more relevant in real-world lending.

The concern raised in prior studies regarding the predominance of the majority class over default cases also emerged in this study. The redefinition of the target variable revealed a clear class imbalance, with "Paid" loans substantially outnumbering "Default" loans, as presented in the Figure 3.1. This imbalance has significant implications, as standard classification algorithms tend to favor the majority class, thereby reducing sensitivity to defaults. To address this challenge, a balancing strategy was applied in later stages to mitigate bias and ensure more robust predictive performance.

[11] https://clientebancario.bportugal.pt/en/arrears-management

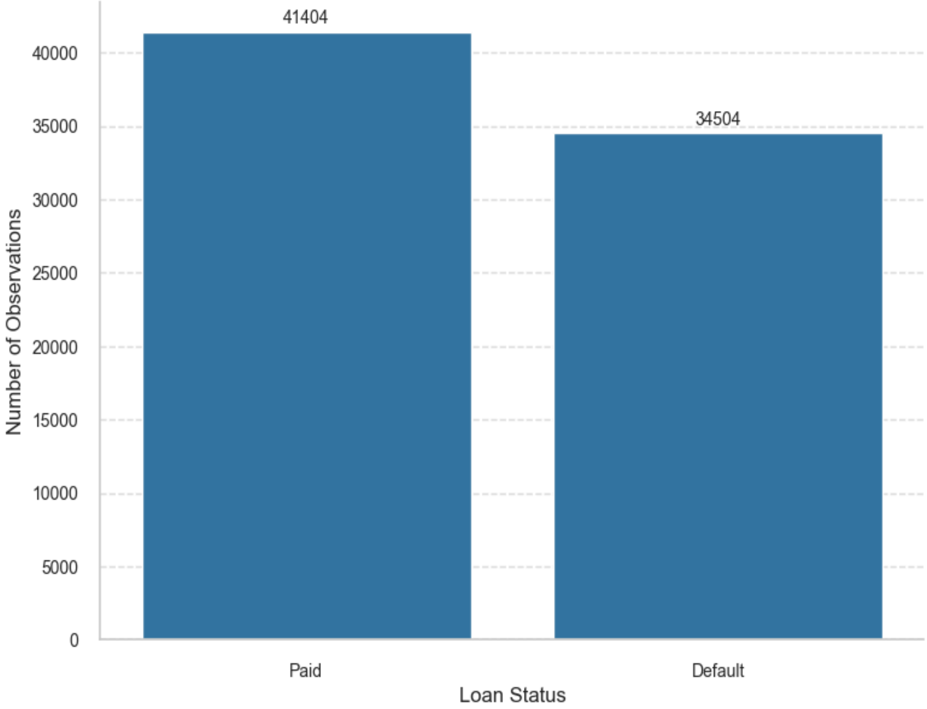

Figure 3.1. Loan Status Distribution (After Binarization)

From a technical standpoint, this binary framing enables the application of supervised learning algorithms suited for binary classification, such as Logistic Regression, Support Vector Machines (SVM), Random Forest, and Gradient Boosting Machines (GBM). These models require a well-defined binary target variable to effectively learn from labeled examples and generalize to new instances. Additionally, this simplification allows for the use of standard evaluation metrics such as accuracy, precision, recall, and AUC-ROC, which are essential for performance comparison and model selection.

From an institutional perspective, this binary outcome mirrors the logic of credit decision-making, where the central concern is to distinguish between borrowers likely to repay and those at risk of default. Misclassifying a high-risk borrower as creditworthy can lead to financial losses, while rejecting a creditworthy applicant can foster financial exclusion and reduce profitability.

3.2.2. Descriptive Statistical Analysis

A comprehensive descriptive analysis was performed to characterize the structure, quality, and distributions of the dataset, thereby guiding subsequent preprocessing strategies and informing modeling constraints. The dataset comprises a total of 75 variables, the majority of which are numerical features, including both continuous variables (float64, n = 46) and discrete variables (int64, n = 7) and a smaller share of categorical or free-text fields (object, n= 22), reflecting the multidimensional nature of credit application data. This heterogeneity spans financial figures (e.g., loan amount, revolving balance), credit behavior (e.g., inquiries, delinquencies), contract metadata, and borrower demographic proxies.phic proxies.

An initial audit of missing data revealed that 26 variables presented incomplete records. Notably, some fields had 100% missingness, including *inq_last_12m*, *total_bal_il*, *annual_inc_joint*, *dti_joint*, and *verification_status_joint*. Some variables are specific to joint loan applications, a subset not retained in this study, and were therefore marked for exclusion. Other variables exhibited partial missingness, often due to conditional logic in their generation. For instance, fields like *mths_since_last_major_derog*, *mths_since_last_record*, and *mths_since_last_delinq* had missing values tied to borrowers without derogatory marks or delinquency events. In such cases, the absence of data carries semantic information rather than noise, and imputation strategies were chosen accordingly, detailed in Chapter 3.3.

From a statistical distribution standpoint, summary metrics such as mean, median, standard deviation, and percentiles were computed for all numerical features. Several variables presented pronounced skewness and heavy-tailed distributions. For instance:

- *loan_amnt* ranged from $500 to $35,000, with a mean of $14,317 and standard deviation of $8,286.
- *annual_inc* showed substantial dispersion, with incomes exceeding $1 million in rare cases and a highly right-skewed profile, underscoring the need for robust statistical summaries (e.g., medians, IQR) or categorization.
- *revol_bal*, representing the borrower's outstanding revolving balance, also displayed extreme values far above the median, indicating potential leverage concentration among a small group of clients.

These distributions suggest the presence of economic heterogeneity across borrowers and justify the application of robust preprocessing transformations in the following stages.

In addition to missingness, a diagnostic of variable imbalance was conducted to identify categorical features dominated by a single value. This included identifier-like variables (*id*, *member_id*) and location proxies (*zip_code*), where over 95% of entries were unique or sparsely populated. Such variables offer little explanatory power for supervised models and were flagged for removal.

Complementing the general statistical profiling, a focused analysis was conducted on variables exhibiting zero inflation i.e., features with a disproportionately high number of zero values. These cases are particularly relevant in financial datasets, where the absence of certain behaviors (e.g., no collections, no inquiries, no installment loans) may itself encode risk-relevant information. For example, the variables *collections_12_mths_ex_med*, *tot_coll_amt*, and *total_rec_late_fee* had over 85% zero values, reflecting a borrower base largely devoid of recent collection activities or accumulated late fees. Rather than excluding these features outright, their sparsity was interpreted in context: for variables where the zero was informative (e.g., indicating “no collections”), they were retained, where the zero reflected structural absence, they were discarded.

The statistical asymmetry of key numerical variables reinforced the decision to avoid standard normalization and instead apply discretization strategies. Variables such as *inq_last_6mths* (number of credit inquiries in the last 6 months), *open_acc* (number of open credit lines), and *annual_inc* (annual income) were transformed into ordinal buckets based on empirical distribution and aligned with typical credit underwriting logic. This transformation served multiple purposes: (i) reducing sensitivity to outliers and extreme values; (ii) increasing model robustness and interpretability; and (iii) incorporating domain-specific thresholds often used in risk scoring (e.g., income brackets, inquiry frequency flags).

Temporal variables (*issue_d*, *last_pymnt_d*, *earliest_cr_line*, *last_credit_pull_d*) posed a particular challenge. Since these fields are recorded in different date formats, they were first converted to elapsed time in months relative to April 2025, a fixed temporal reference established for modeling and analytical alignment. For instance, the *earliest_cr_line* variable, which indicates the month of the borrower’s oldest credit line, was transformed into “months since first credit”, providing a measure of credit history maturity. After this transformation, temporal variables were further discretized into meaningful intervals (e.g., < 6 months, 6–12 months, 12–24 months, > 60 months), enabling risk stratification based on credit history recency and loan activity age.

Lastly, the data audit uncovered several quasi-constant features, i.e., variables with extremely low variance or a dominant single value. Examples include *policy_code*, *application_type*, and some status indicators, where over 95% of the values were concentrated in a single category. These features were flagged for removal given their negligible informational value and potential to introduce bias during training or inflate model complexity. In parallel, a subset of categorical variables presented extremely high cardinality, potentially limiting their direct applicability in conventional encoding strategies. This includes fields such as *url*, *desc*, and *title*, each with tens of thousands of unique entries, often reflecting unstructured text or user-defined inputs without standardization.

In summary, the descriptive statistical analysis revealed a dataset with diverse data types, relevant informational structure, but also notable challenges related to missingness, skewed distributions, zero inflation, high-cardinality categorical fields, heterogeneous or misaligned date formats and feature redundancy. These findings supported a series of data-driven decisions, such as targeted feature exclusion, discretization of continuous variables, and semantic transformation of temporal data. They also highlighted the importance of encoding business and domain-specific logic into preprocessing routines to improve model performance and interpretability. This analytical phase not only offered visibility into the structure and limitations of the dataset, but also informed a series of methodological decisions for imputation, transformation, and variable engineering procedures applied in response to these challenges, including imputation strategies, variable transformations, and final feature selection for model development.

3.3. Data Preparation

Following the descriptive data analysis, one of the first decisions in the data preparation phase was to apply a filter to the dataset based on the variable *annual_inc*, which represents the borrower's declared annual income. This choice was motivated by the identification of extreme and implausible income values in the original dataset, ranging from near-zero declarations to incomes above one million USD, figures that are statistically rare and contextually disconnected from the socioeconomic profile of the population under study.

To strengthen analytical consistency and try to avoid huge distortions, an income interval between USD 25,000 and USD 75,000 was adopted. This range does not aim to replicate Portuguese salary levels directly, but rather to establish a plausible spectrum that excludes unrealistic declarations while retaining higher-income cases that could reasonably occur in urban or specialized labor contexts. Additionally, the *annual_inc* variable was discretized into categorical brackets (e.g., 25K–50K; 50K–75K), enabling its use as an ordinal feature in subsequent modeling stages.

According to official statistics from Instituto Nacional de Estatística (INE) and PORDATA[12], the average annual gross salary for full-time workers in Portugal in 2023 was approximately €22,000, with variation across sectors and gender. Although this figure is below the adopted income interval, the inclusion of an upper threshold of was retained to

[12] https://www.pordata.pt/pt/estatisticas/salarios-e-pensoes/salarios/salario-medio-anual-ajustado-tempo-inteiro

provide methodological flexibility, accommodating profiles such as expatriates, entrepreneurs, and high-skilled professionals, while limiting disproportionate bias from outliers. Therefore, the applied filtering should not be interpreted as a direct alignment between U.S. and Portuguese salary structures but rather as a methodological adjustment intended to approximate coherence with the Portuguese context while maintaining analytical tractability of the dataset.

3.3.1.Temporal Feature Engineering

The dataset included four key temporal variables stored originally as string-type fields in the format %b-%y (e.g., Mar-15):

- *earliest_cr_line*: borrower's earliest date reported credit line was opened
- *issue_d*: loan issue date
- *last_pymnt_d*: date of last payment
- *last_credit_pull_d*: date of most recent credit report pull

To enable these features to be meaningfully incorporated into the modeling process, all were parsed into datetime objects using a consistent date parser. A fixed reference point of April 1st, 2025[13] was used to calculate the elapsed time in months between each respective date and the reference date. As a result, four new numerical variables were created:

- *earliest_cr_line_months_diff*
- *issue_d_months_diff*
- *last_pymnt_d_months_diff*
- *last_credit_pull_d_months_diff*

These variables represent, respectively, the length of the credit history, loan age, recency of last payment, and recency of last credit pull, all highly relevant metrics in credit scoring applications.

To increase interpretability and facilitate segmentation aligned with industry practices, these continuous variables were discretized into ordinal categorical variables, resulting in the following new fields:

- *earliest_cr_line_category*

[13] Methodological note: Although the dataset only extends until 2014, a fixed reference date of April 1st, 2025 was defined for the calculation of temporal variables. Without this procedure, the categorization into intervals (e.g., "< 6 months", "6–12 months") would shift each time the notebook was re-executed, undermining the stability and comparability of the results. The adoption of a future reference date is therefore a methodological decision to ensure analytical consistency, rather than an artificial extension of the dataset's coverage period.

- *issue_d_category*
- *last_pymnt_d_category*
- *last_credit_pull_d_category*

Each of these variables was grouped using the same binning logic, creating the following seven time-based categories:

1. < 6 months
2. 6–12 months
3. 12–24 months
4. 24–36 months
5. 36–48 months
6. 48–60 months
7. > 60 months

This structured binning served several purposes: it reduced the influence of outliers (e.g., extremely old or recent records), preserved ordinal temporal information, and aligned the dataset with a typical credit lifecycle segmentation. These temporal features thus offered both historical depth (via *earliest_cr_line*) and recency measures (via *last_pymnt_d* and *last_credit_pull_d*), contributing to the model's ability to distinguish between established and high-risk borrowers.

3.3.2.Discretization of Numerical Features

To reduce skewness, mitigate the influence of extreme values, and improve model interpretability, also continuous numerical features were transformed into ordered categorical variables. This process relied on domain-informed discretization thresholds, designed to emulate financial decision-making logic and enhance compatibility with models sensitive to feature distribution.

The following variables were discretized:

- *inq_last_6mths → inq_last_6mths_category*
  - Grouped according to inquiry intensity: 0, 1–2, 3–5, 6–10, 11–20, > 20, and "missing". This segmentation reflects the borrower's recent credit-seeking behavior, which is often inversely correlated with creditworthiness.
- *open_acc → open_acc_category*
  - Categorized as: 1–5, 6–10, 11–20, 21–30, 31–50, and > 50. This variable captures the breadth of the borrower's credit portfolio, with higher counts typically reflecting more complex financial profiles.

- *revol_util → revol_util_category*
    - Discretized into usage bands: 0–10%, 11–20%, 21–30%, 31–40%, 41–50%, 51–60%, 61–70%, and 71–80%. These intervals align with conventional thresholds for assessing revolving credit utilization risk.
- *total_acc → total_acc_category*
    - Grouped into: 1–10, 11–20, 21–30, 31–40, 41–50, and > 50, enabling segmentation of borrower experience based on the total number of credit lines.
- *tot_coll_amt → tot_coll_amt_category*
    - Binned into: 0, 1–500, 501–1,000, 1,001–2,000, and > 2,000, capturing the borrower's history of collections, a key indicator of previous delinquency.

Each of these transformations was executed via custom *apply()* functions from Pandas library[14], ensuring consistent bin assignment and enhancing interpretability. The resulting variables are all ordinal in nature, preserving the relative ordering of categories, which can be particularly beneficial in certain modeling contexts. This discretization strategy proved especially effective in tree-based algorithms like Decision Tree, Random Forest and Gradient Boosting Machines, which benefit from segmented feature spaces and are sensitive to outlier distributions. Additionally, the transformation contributed to noise reduction and facilitated model explainability.

Throughout the preprocessing phase, various features were transformed into categorical format, either through direct conversion (e.g., date-based binning) or domain-driven segmentation (e.g., inquiries, utilization). However, no explicit numerical encoding (such as one-hot or ordinal encoding) was applied at this stage. Instead, these variables were retained as labeled category fields, allowing encoding decisions to be tailored later in the modeling pipeline, depending on algorithmic needs. This approach preserved the semantic integrity of the features, facilitated exploratory data analysis, and allowed for encoding flexibility. For instance, models may benefit from ordinal-encoded categories, while linear models might require one-hot representations. By deferring encoding, the modeling strategy remains adaptive and model-aware.

Categorical fields were all preserved in labeled format and will be encoded in alignment with the chosen modeling techniques.

[14] https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.apply.html

3.3.3. Outliers and Skewness Treatment

Following the first preprocessing steps from the data preparation phase, a structured treatment was applied to detect and mitigate the presence of outliers and the skewness of numeric variables, factors known to adversely affect model robustness and predictive fairness. As an initial diagnostic step, boxplots were generated for all continuous numerical variables to visualize the distribution and detect potential outliers. These plots provided a clear overview of the spread, central tendency, and the presence of extreme values across the dataset as detailed in Figure 3.2.

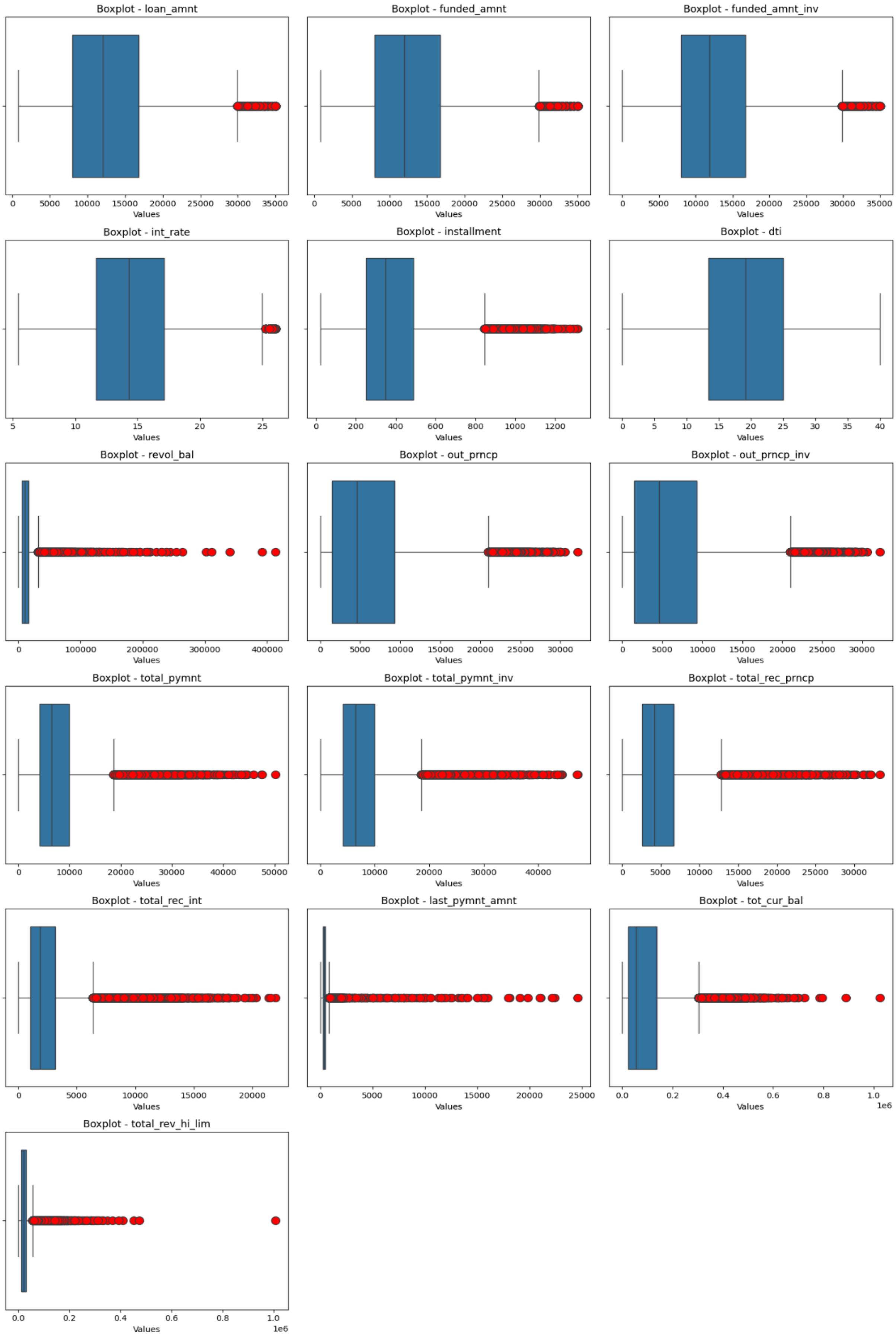
Boxplot - loan_amnt
0 5000 10000 15000 20000 25000 30000 35000
Values
Boxplot - funded_amnt
0 5000 10000 15000 20000 25000 30000 35000
Values
Boxplot - funded_amnt_inv
0 5000 10000 15000 20000 25000 30000 35000
Values
Boxplot - int_rate
5 10 15 20 25
Values
Boxplot - installment
0 200 400 600 800 1000 1200
Values
Boxplot - dti
0 5 10 15 20 25 30 35 40
Values
Boxplot - revol_bal
0 100000 200000 300000 400000
Values
Boxplot - out_prncp
0 5000 10000 15000 20000 25000 30000
Values
Boxplot - out_prncp_inv
0 5000 10000 15000 20000 25000 30000
Values
Boxplot - total_pymnt
0 10000 20000 30000 40000 50000
Values
Boxplot - total_pymnt_inv
0 10000 20000 30000 40000
Values
Boxplot - total_rec_prncp
0 5000 10000 15000 20000 25000 30000
Values
Boxplot - total_rec_int
0 5000 10000 15000 20000
Values
Boxplot - last_pymnt_amnt
0 5000 10000 15000 20000 25000
Values
Boxplot - tot_cur_bal
0.0 0.2 0.4 0.6 0.8 1.0
1e6
Values
Boxplot - total_rev_hi_lim
0.0 0.2 0.4 0.6 0.8 1.0
1e6
Values

Figure 3.2. Boxplots of Numerical Variables Before Outlier Treatment

In this initial exploration, outliers were formally identified using a decile-based approach. For each continuous variable, the 10th and 90th percentiles (P10 and P90) were calculated. Observations falling outside this range either below P10 or above P90 were flagged as outliers. This method was preferred over traditional interquartile range (IQR) techniques (based on Q1–Q3) as it allows for greater flexibility around the median, providing a more tolerant threshold suitable for inherently skewed financial variables, as commonly seen in credit risk datasets. A representative example of this analysis is presented in Figure 3.3.

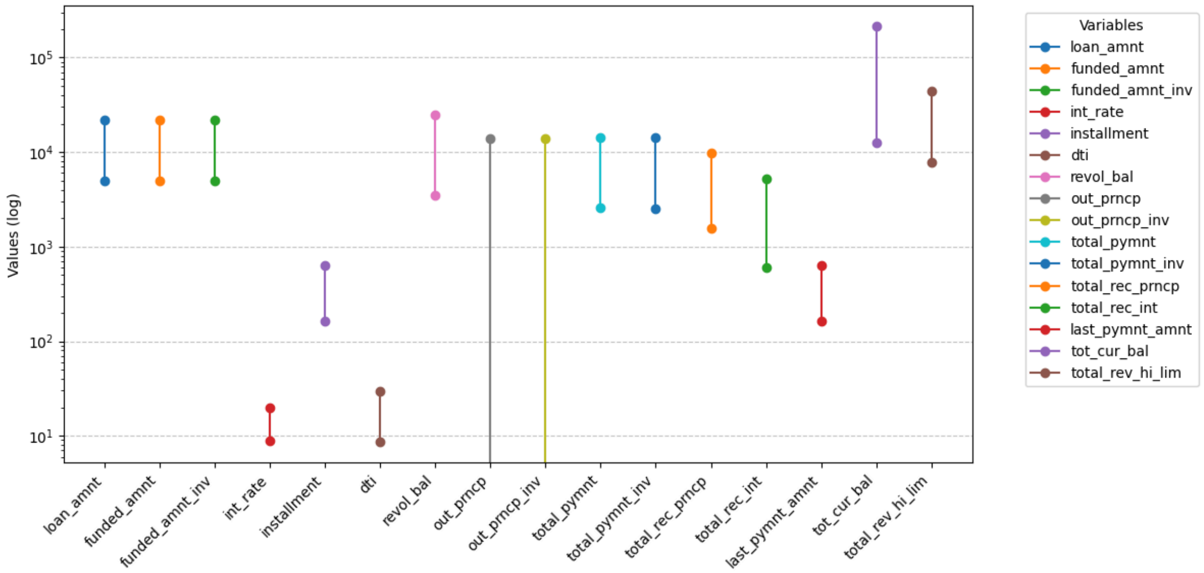

Figure 3.3. Range Between 10th and 90th Percentiles for Numerical Variables (Logarithmic Scale)

A visual summary of the numerical variable distributions and their respective decile ranges is presented in Figure 3.3. Each vertical line represents the spread between the 10th and 90th percentiles for a given numeric variable. To enhance interpretability, particularly given the wide variation in monetary values, a logarithmic scale was applied to the y-axis. This adjustment enabled the visualization of both low- and high-magnitude features within the same coordinate system, revealing the presence of long-tailed distributions and considerable intra-variable variance.

Notably, variables such as *installment*, *revol_bal*, *tot_cur_bal*, and *total_rec_int* exhibit especially wide spreads, reinforcing the necessity of targeted outlier treatment. Variables such as *out_prncp* and *out_prncp_inv*, which represent the remaining unpaid loan principal (total and investor share, respectively), also displayed significant dispersion. Their wide decile ranges reflect the natural variability in residual balances across different loan repayment stages. These features, by nature of representing outstanding debt, are especially prone to skewness and thus benefit from winsorization to reduce their disproportionate influence on model training.

This visual diagnosis complemented the statistical skewness analysis and reinforced the decision to adopt a robust, percentile-based transformation strategy.

To quantify the impact, a binary mask was created to identify rows with at least one outlier. Since this affected more than 15% of the dataset, dropping these rows would have led to significant information loss. To address this, a randomized winsorization strategy was applied: flagged outliers were replaced with random values drawn uniformly from within the P10–P90 interval of the corresponding variable. This technique preserved the original scale and variability of the data while minimizing the distorting effect of extreme values.

The outcome of this transformation is illustrated in Figure 3.4, which displays the boxplots of the numerical variables after outlier treatment. Compared to the original distributions (Figure 3.2), a noticeable reduction in extreme values and interquartile spread can be observed, especially in variables such as *revol_bal*, *tot_cur_bal*, and *out_prncp*. This confirms the effectiveness of the procedure in dampening the influence of outliers while retaining the informational structure of the data.

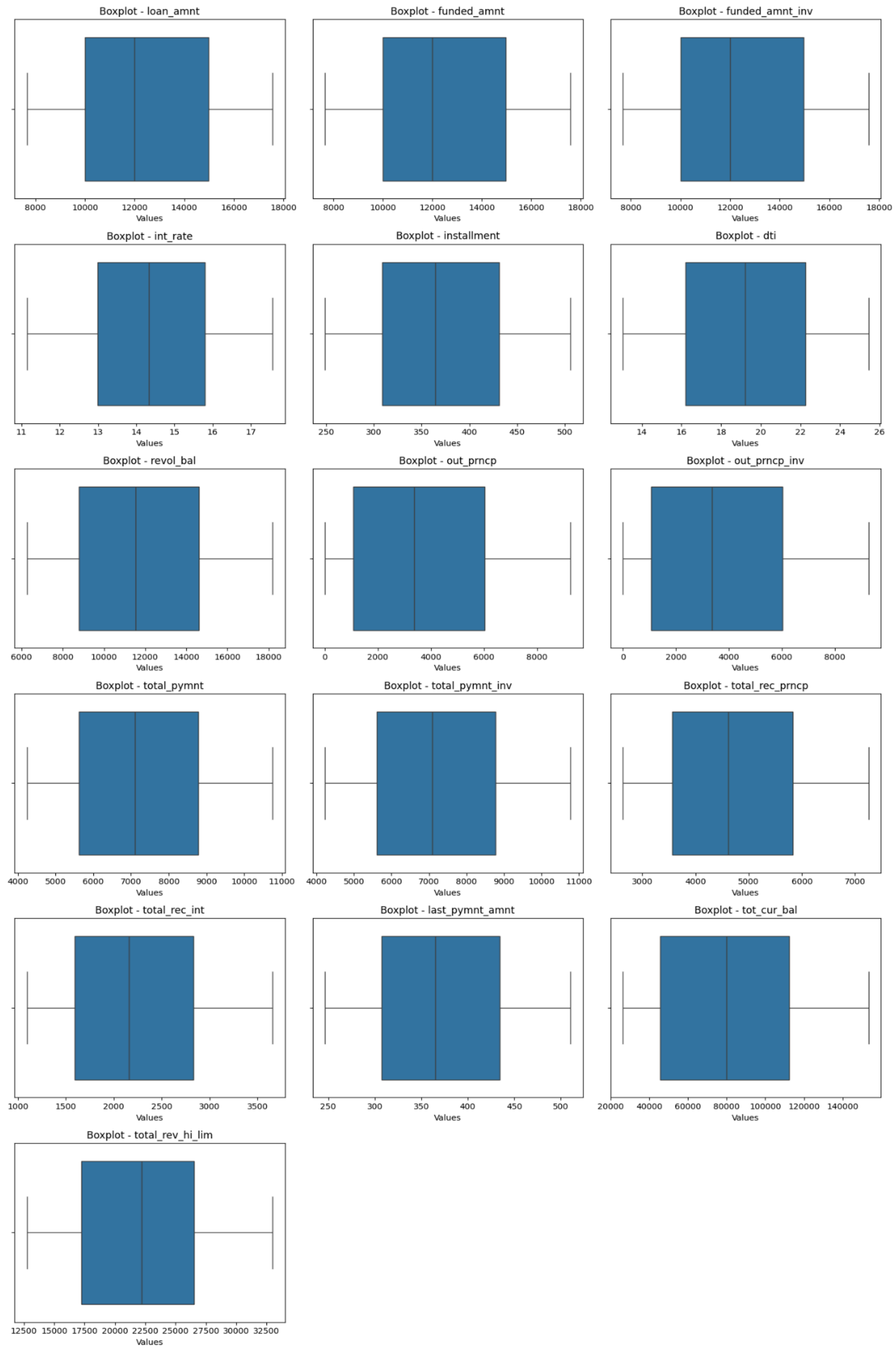
Boxplot - loan_amnt
8000 10000 12000 14000 16000 18000
Values
Boxplot - funded_amnt
8000 10000 12000 14000 16000 18000
Values
Boxplot - funded_amnt_inv
8000 10000 12000 14000 16000 18000
Values
Boxplot - int_rate
11 12 13 14 15 16 17
Values
Boxplot - installment
250 300 350 400 450 500
Values
Boxplot - dti
14 16 18 20 22 24 26
Values
Boxplot - revol_bal
6000 8000 10000 12000 14000 16000 18000
Values
Boxplot - out_prncp
0 2000 4000 6000 8000
Values
Boxplot - out_prncp_inv
0 2000 4000 6000 8000
Values
Boxplot - total_pymnt
4000 5000 6000 7000 8000 9000 10000 11000
Values
Boxplot - total_pymnt_inv
4000 5000 6000 7000 8000 9000 10000 11000
Values
Boxplot - total_rec_prncp
3000 4000 5000 6000 7000
Values
Boxplot - total_rec_int
1000 1500 2000 2500 3000 3500
Values
Boxplot - last_pymnt_amnt
250 300 350 400 450 500
Values
Boxplot - tot_cur_bal
20000 40000 60000 80000 100000 120000 140000
Values
Boxplot - total_rev_hi_lim
12500 15000 17500 20000 22500 25000 27500 30000 32500
Values

Figure 3.4. Boxplots of Numerical Variables After Outlier Treatment

In order to systematically assess the asymmetry of the dataset, skewness was computed for all continuous numerical variables prior to any normalization or transformation, but subsequent to the treatment of outliers. This procedure ensured that the evaluation reflected the intrinsic distributional characteristics of each variable, thereby facilitating the identification of features potentially requiring additional smoothing. The skewness values, calculated using the scipy.stats.skew()[15] function and reported in Table 3.1, indicate a general tendency toward moderate positive skewness, particularly in variables such as *revol_bal*, *out_prncp*, *out_prncp_inv*, *total_rec_prncp, total_rec_int, tot_cur_bal* and *total_rev_hi_lim*.

Table 3.1. Skewness Calculation Before Transformation

Variable	**Skew**	**Variable**	**Skew**
loan_amnt	0.2395	*total_pymnt*	0.4491
funded_amnt	0.2391	*total_pymnt_inv*	0.4466
funded_amnt_inv	0.2401	*total_rec_prncp*	0.4924
int_rate	0.0491	*total_rec_int*	0.6286
installment	0.2653	*last_pymnt_amnt*	0.2695
dti	0.0102	*tot_cur_bal*	0.6389
		total_rev_hi_lim	0.4817
revol_bal	0.4496		
out_prncp	0.4592		
out_prncp_inv	0.4595		

Notably in Table 3.1, a significant portion presented values clustered around 0.4–0.6. This pattern suggested that adopting the conventional threshold of 0.5 for skew correction could result in the omission of borderline cases that still deviated from symmetry. Therefore, a more conservative threshold of 0.3 was adopted.

[15] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.skew.html

Furthermore, identifying the direction of the skew (positive or negative) allowed for tailored transformation strategies. When all values were non-negative or strictly positive, logarithmic and Box-Cox[16] transformations were applied. Although the majority of variables exhibited moderate positive skewness prior to transformation, the Yeo-Johnson[17] method was predominantly applied, as it is more flexible and does not require strictly positive inputs, unlike the Box-Cox method. This transformation also preserves the original sign and structural characteristics of the data, particularly for variables with values close to zero, which indicates symmetry or near-symmetry, making it especially suitable for financial variables with skewed distributions and heterogeneous magnitudes.

This differentiated approach ensured that skewness was corrected while preserving the relative relationships among the original variables, an important concern for interpretability in credit risk applications. Although the transformations modify the scale of the variables (e.g., through logarithmic compression), they retain the ordinal structure of the data, thereby safeguarding their analytical meaning. After applying these methods, the skewness of nearly all affected variables was reduced below the 0.3 threshold, contributing to a more robust and model-friendly dataset.

To summarize the results of the skewness correction process, Table 3.2 presents a detailed comparison of the numerical variables before and after transformation. Each feature listed initially that exceeded the predefined threshold of 0.3 for skewness, it was subsequently processed using either the Square Root or Yeo-Johnson transformation. The observed reductions in skewness validate the effectiveness of the chosen methods in normalizing feature distributions while preserving interpretability and analytical consistency.

Table 3.2. Skewness Correction for Numerical Features (Before vs. After Transformation)

Variable	**Skew (Before)**	**Skew (After)**	**Transformation**
tot_cur_bal	0.6389	0.0502	Yeo-Johnson
total_rec_int	0.6286	0.0311	Yeo-Johnson
total_rec_prncp	0.4924	0.0329	Yeo-Johnson
total_rev_hi_lim	0.4817	0.0360	Yeo-Johnson

[16] https://docs.scipy.org/doc/scipy-1.16.1/reference/generated/scipy.stats.boxcox.html

[17]https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PowerTransformer.html

Variable	**Skew (Before)**	**Skew (After)**	**Transformation**
out_prncp	0.4592	0.3810	Square Root (sqrt)
out_prncp_inv	0.4595	0.3804	Square Root (sqrt)
revol_bal	0.4496	0.0505	Yeo-Johnson
total_pymnt	0.4491	0.0361	Yeo-Johnson
total_pymnt_inv	0.4466	0.0368	Yeo-Johnson

After completing the comprehensive data preparation pipeline, including data filtering, transformation, categorical binning, handling of missing values, outlier treatment, and skewness normalization, a new transformed dataset was generated containing 36 variables. This refined dataset serves as the foundation for the subsequent steps. In the following chapters, correlation analysis and feature selection techniques are applied to identify and retain the most relevant predictors. These steps were strategically adopted to reduce the dataset's dimensionality and concentrate the modeling efforts on the features most likely to influence the classification outcomes, thereby enhancing model interpretability, performance, and practical applicability in credit risk assessment.

3.4. Dimensionality Reduction: Correlation and Feature Selection

An essential step following the data transformation pipeline was the evaluation of associations both among predictor variables (to detect potential multicollinearity) and between predictor variables and the binary target variable (*loan_status*). This analysis supported informed feature selection by identifying redundant information, weakly informative predictors, and variables with moderate to strong relationships with the classification outcome.

3.4.1. Correlation Analysis

Given the distinct nature of variable types (numerical *vs.* categorical), different correlation measures were applied to capture associations both among predictors and between predictors and the binary target variable.

For numerical predictors, Pearson's correlation coefficient (ρ) was employed to quantify linear dependencies between pairs of continuous variables. This enabled the identification of multicollinearity, particularly among monetary amounts and loan-related balances. For categorical predictors, Cramér's V was used to evaluate associations between pairs of categorical features, providing a bounded measure (0 to 1) suitable for detecting redundancies within categorical information.

To assess the relationship between predictors and the binary target, two complementary approaches were adopted. Eta Squared (η^2) was calculated for numerical predictors against the encoded target variable, allowing the estimation of the proportion of variance in the target explained by each continuous feature. In parallel, Cramér's V was applied to measure the association between categorical predictors and the categorical target. This dual approach ensured that both numerical and categorical predictors were systematically evaluated in terms of their predictive relevance to the outcome of interest.

Taken together, these methods offered complementary insights. Pearson's ρ and Cramér's V revealed redundancy among predictors, while η^2 and Cramér's V highlighted the strength of each predictor's association with the target. Despite the rigorous application of these techniques, the dataset exhibited only weak to moderate predictor–target associations, with the most relevant features reaching values of approximately 0.4. This indicates that no single predictor displayed strong standalone predictive power, underscoring the importance of multivariate modeling strategies capable of capturing joint patterns and interaction effects.

However, the analysis successfully identified high collinearity among numeric variables, particularly:

- *loan_amnt*, *funded_amnt*, *funded_amnt_inv*
- *total_rec_prncp*, *total_pymnt*, *last_pymnt_amnt*
- *out_prncp* and *out_prncp_inv*

Among these, only a single representative feature was retained per group, chosen based on correlation strength with the target and domain interpretability. Conversely, categorical features with near-zero Cramér's V or η^2 values, indicating no significant association with the target, were flagged for removal.

Table 3.3. Removed Variables After Correlation and Feature Selection Analysis

Removed Variable	**Primary Reason**
out_prncp_inv	Highly correlated with *out_prncp* (Pearson ≈ 0.99) — redundant information. Also, this particular metric focuses on the outstanding

Removed Variable	Primary Reason
	principal amount for a portion funded by the investor, which is very particular for the peer-to-peer context but not in the commercial bank.
funded_amnt	Strong correlation with *loan_amnt* (Pearson ≈ 0.98) — carries overlapping signals.
funded_amnt_inv	Strong correlation with *loan_amnt* (Pearson ≈ 0.97) — adds no new explanatory power.
installment	Strong correlation with both *loan_amnt* (Pearson ≈ 0.89) and *int_rate* (≈ 0.65) — removed due to multicollinearity.
total_pymnt_inv	Redundant with *total_rec_prncp* (Pearson ≈ 0.96) — cumulative payment overlap.
total_rec_prncp	High correlation with *total_pymnt* (Pearson ≈ 0.95) — only one retained.
last_pymnt_amnt	Redundant with other cumulative repayment features (Pearson with total_pymnt ≈ 0.74).
grade	Encoded abstraction of *sub_grade* (Cramér's V ≈ 0.93) — removed in favor of a more granular feature.
term	Low association with the target (Cramér's V ≈ 0.05) — negligible predictive contribution.
verification_status	Categorical feature with near-zero association to target (Cramér's V < 0.01).
open_acc_category	Categorical binning showed η^2 (Eta Squared) < 0.01 — weak explanatory relevance.

Table 3.3 summarizes all variables excluded during the correlation exercise, specifying whether their removal was due to multicollinearity or low association with the binary target. This refinement step was essential to constructing a more parsimonious and robust feature set

by reducing redundancy and emphasizing variables with meaningful associations to the target. The diagnostic process played a key role in ensuring that the remaining features were not only individually informative but also complementary, avoiding overlap in the information they contributed to the model.

3.4.2. Feature Selection Strategy

Following the correlation-based variable screening, an additional feature selection phase was conducted to refine the dataset by identifying the most relevant predictors for supervised modeling. This complementary stage focused on evaluating the 25 variables retained from the correlation analysis, including the target one. To identify the features with the greatest predictive potential, the analysis combined three complementary approaches: univariate statistical tests, model-based importance scores, and multicollinearity diagnostics. This strategy supported dimensionality reduction, mitigated redundancy, and prioritized variables with stronger discriminative power relative to the binary target.

The process began by excluding the target variable and encoding all categorical variables into numerical format using one-hot encoding, omitting the first level to prevent the dummy variable trap. This step resulted in a fully numeric feature matrix, ensuring compatibility with the techniques employed in subsequent stages. All columns were then explicitly cast as float to guarantee uniformity across the dataset.

3.4.3. Univariate Statistical Tests (ANOVA F-test)

To evaluate the individual predictive power of each feature, an ANOVA F-test[18] was applied via the SelectKBest method using the *f_classif* scoring function. This test compares the means of each numeric feature across the two classes of the target variable ("Paid" and "Default"), identifying whether its distributions differ significantly between the groups. The corresponding F-statistics, presented in Table B.1 of the Appendix B, offered a preliminary measure of the explanatory power of each predictor with respect to default behavior. While no hard threshold was imposed at this stage, features with very low F-scores were flagged for potential removal, subject to confirmation from other methods.

3.4.4. Mutual Information Analysis

The first step involved computing Mutual Information (MI) between each predictor and the categorical target variable. MI is a non-parametric measure of the dependency between

[18] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html

variables and is well-suited for both categorical and numerical predictors. In this context, a value closer to 1 implies higher dependency, while values close to 0 indicate low or no association. In other words, MI quantified how much knowing the value of a particular feature reduced the uncertainty about the loan status.

Using the SelectKBest method with *mutual_info_classif* as the scoring function, all remaining features were evaluated. The resulting, detailed in the Table B.2 of the Appendix B, scores highlighted variables such as *out_prncp*, *int_rate*, and *loan_amnt* as those providing the highest incremental information regarding the target. In contrast, variables like *dti* and *revol_bal* registered negligible or zero MI scores, suggesting no informational contribution in predicting default or repayment.

3.4.5. SelectFromModel using Logistic Regression

To complement the univariate assessment, a model-based approach was applied using SelectFromModel[19] with Logistic Regression as the underlying estimator. This technique ranks features based on their absolute coefficients in the trained model, thus offering insight into their importance in a multivariate predictive setting.

Logistic Regression was selected due to its interpretability and appropriateness for binary classification tasks. Following the estimation of the model on the encoded feature matrix, the *get_support()* method identified variables such as *int_rate* and *dti* as meaningful contributors. This outcome not only reinforced earlier findings but also highlighted instances in which a variable appeared weak when considered in isolation (i.e., low MI) yet demonstrated relevance when assessed in combination with others. The complete list of features retained through this procedure is presented in Table B.3 of the Appendix B, serving as complementary evidence to the results discussed in the main text.

3.4.6. Feature Selection Output and Interpretation

To synthesize the results from the different selection techniques (ANOVA, Mutual Information, and Logistic Regression via SelectFromModel), a consolidated evaluation was conducted to identify the most consistently relevant predictors. The goal of this final step was to build a reduced and interpretable dataset, composed only of the features most frequently flagged as important across the three methods.

The code performed the following steps:

1. Top Variables per Method:

[19] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html

    a. The top 15 features based on ANOVA F-scores and Mutual Information were extracted.
    b. Logistic Regression-based feature selection was already binary (selected or not).
2. Unification and Mapping to Original Variables:
    a. Since many features had been previously one-hot encoded (e.g., categorical features like *home_ownership_RENT*, *term_36_months*), they were programmatically mapped back to their original categorical feature names using string parsing logic (e.g., splitting by underscores).
    b. This allowed an aggregate count of how many times each original feature was selected across all dummy columns.
3. Voting Threshold:
    a. A simple majority rule was applied: original variables with at least two selections across the three techniques were retained.
    b. This threshold ensured that only features demonstrating stable importance across different statistical and model-based perspectives were carried forward.

The application of three distinct feature selection techniques, and model-based selection via Logistic Regression, revealed both convergences and divergences in feature relevance. The ANOVA F-test highlighted *out_prncp* and *int_rate* as the most statistically discriminative variables, with exceptionally high F-values, indicating their strong ability to separate the binary classes of the target variable. Mutual Information, in turn, reinforced the relevance of *out_prncp*, assigning it a substantially high dependency score, while suggesting little to no informational gain from variables like *loan_amnt, revol_bal*, or *dti*. Interestingly, the Logistic Regression model via SelectFromModel corroborated the importance of *int_rate* and *out_prncp*, while also identifying *dti*, *total_pymnt*, and *home_ownership_RENT* as influential predictors, some of which were overlooked by purely statistical techniques.

This comparative analysis underscores the value of combining complementary methods: while statistical tests capture univariate relationships, model-based selection can uncover interaction effects and context-specific importance. Ultimately, variables repeatedly flagged across techniques were prioritized, while those showing negligible scores in both statistical and model-based evaluations were discarded. This process resulted in a refined feature set containing 10 variables that balances predictive power, interpretability, and parsimony for the subsequent modeling phase. The full cross-tabulation of results from the different feature selection approaches is presented in Table B.4 of the Appendix B, providing transparent evidence of how the final set of predictors was determined.

3.5. Modeling Techniques

To evaluate the predictive capacity of the selected features, a series of supervised classification models were implemented using the PyCaret[20] library. PyCaret is an open-source, low-code machine learning library in Python that simplifies the end-to-end process of building, comparing, and deploying machine learning models. Designed for productivity and rapid experimentation, PyCaret abstracts much of the underlying complexity by offering a unified, high-level API for data preparation, model training, hyperparameter tuning, and performance evaluation.

Its core design principle revolves around modularity and automation, allowing users to quickly compare dozens of models using consistent preprocessing pipelines and evaluation metrics. PyCaret supports various ML tasks including classification, regression, clustering, and anomaly detection and integrates seamlessly with popular frameworks like scikit-learn, XGBoost, LightGBM, and CatBoost.

For this study, the classification module of PyCaret was adopted to analyze the binary target variable (*loan_status*), with two distinct configurations:

- Unbalanced scenario, preserving the original class distribution of the target variable;
- Balanced scenario, applying undersampling techniques to mitigate class imbalance and evaluate potential performance improvements.

The choice of PyCaret was motivated by its efficiency in prototyping and benchmarking, as well as its comprehensive suite of tools for model comparison, feature importance ranking, and cross-validation. These characteristics make PyCaret not only a powerful research tool, but also it can represent a pragmatic solution for credit institutions aiming to accelerate model development cycles, enhance reproducibility, and facilitate the deployment of robust credit scoring systems with minimal code and resource requirements. In the following sections, both scenarios for class imbalance are detailed, with emphasis on model performance metrics, interpretability, and implications for credit scoring applications.

To reinforce the robustness of the results, the models were also trained using their native libraries (scikit-learn and LGBMClassifier). This replication aimed to rule out potential framework-induced biases and confirm that PyCaret’s performance gains were not artifacts of

[20] https://pycaret.org/

its automation. Standalone Python scripts were developed for the six shortlisted models: LightGBM, AdaBoost, Random Forest, Extra Trees, Gradient Boosting, and Logistic Regression, under both unbalanced and balanced target scenarios. This procedure ensured consistency in parameterization and evaluation, validated that the results were not contingent on a single environment, and demonstrated the models' ability to generalize across implementations, enabling a more detailed comparison of evaluation metrics in the subsequent chapters.

3.5.1. Unbalanced Dataset: Modeling Techniques

As established in the target definition phase, the target variable in this study is significantly imbalanced, with the "Paid" class far outnumbering the "Default" class. Such imbalance is common in credit datasets and challenges classification models by biasing algorithms toward the majority class. This section presents the initial modeling experiments conducted on the unbalanced dataset using both PyCaret's classification module and native implementations.

In both frameworks, the dataset employed was the refined output derived from the feature selection phase. It was split into training and testing sets, maintaining a 70/30 proportion and preserving the original class distribution through random sampling (*random_state* = 42). In PyCaret, the *setup()* function was used to initialize the modeling environment. This function configured the data pipeline by identifying data types, applying necessary type coercions, and setting the target variable as *loan_status*. No additional transformations were applied at this stage, given that all preprocessing, including outlier treatment, normalization, and discretization, had already been performed upstream.

In the native implementation, the training process began with loading the dataset and transforming the target variable (*loan_status*) into a binary classification task, distinguishing between Paid (1) and Default (0). The data was then divided into training and testing subsets using a 70/30 split, ensuring the target distribution was preserved. Categorical variables were explicitly converted into the category data type, enabling the LightGBM classifier to process them directly without requiring one-hot encoding.

The modeling phase began with a benchmarking stage using PyCaret's *compare_models()* function, which systematically evaluates a suite of classification algorithms using k-fold cross-validation. Accuracy was initially selected as the sorting metric to provide a broad baseline comparison of model effectiveness. In total, the *compare_models()* function of PyCaret was employed to evaluate fourteen classification algorithms under the unbalanced scenario. The set of algorithms comprised was: LightGBM, CatBoost, Extra Trees Classifier, Random Forest, Gradient Boosting Classifier, AdaBoost Classifier, Ridge Classifier, Logistic Regression, K-Nearest Neighbors, Decision Tree Classifier, Naive Bayes, Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Support Vector Machine (SVM) with a linear kernel. The resulting models were ranked by PyCaret based on their performance. The top five and Logistic Regression results are reported in Table 4.3 (Chapter 4.1). For consistency and fair comparison, the native implementations were initialized with the same hyperparameters used in PyCaret.

Light Gradient Boosting Machine (LightGBM) emerged as the top performer, followed by other ensemble-based and tree-based models. However, acknowledging the business-critical nature of the problem, where misclassifying a defaulter as a solvent borrower can lead to financial losses and reputational damage, recall was subsequently prioritized as the key evaluation metric and the one adopted for model tuning. This decision was grounded in the institutional perspective of credit risk management, where identifying as many default cases as possible is often preferred, even at the cost of a higher false positive rate. Following this prioritization, the top five models based on recall were shortlisted: LightGBM, Random Forest Classifier, AdaBoost Classifier, Gradient Boosting Classifier, and Extra Trees Classifier. Additionally, Logistic Regression, although not among the top performers, was intentionally included due to its longstanding use in traditional credit scoring practices by financial institutions.

Each of these six models was then individually fine-tuned using PyCaret's *tune_model()* function, which applies randomized grid search optimization with fold cross-validation (k=10) to identify the best-performing hyperparameter configuration based on recall. It is important to note that, by default, PyCaret automatically assigns an initial set of hyperparameters to each model upon creation through the *create_model()*[21] function. These defaults serve as a standardized baseline, allowing for fair comparison across classifiers. The *tune_model()*[22] process then systematically searches for alternative configurations around these defaults, exploring parameters such as learning rate, number of trees, maximum depth, number of leaves, minimum child weight, and subsampling strategies (as in the case of LightGBM). This structured optimization enables a balance between predictive accuracy and model complexity. Nevertheless, PyCaret also provides an explicit comparison, returning the original model if tuning does not yield improvements, which was the case for some classifiers in this study. Ultimately, very small performance gains were observed in the tuned versions of

[21] https://pycaret.gitbook.io/docs/get-started/functions/train#create_model
[22] https://pycaret.gitbook.io/docs/get-started/functions/optimize#tune_model

LightGBM, Random Forest, Gradient Boosting Classifier, and Logistic Regression, while other models retained their original configurations as the optimal choice. In this sense, in the native implementation, no hyperparameter tuning was performed, as both accuracy and recall also consistently exceeded 90% across all models during training and were further confirmed in the test set, as reported in Table 4.5 of Chapter 4.1 (p.64).

After tuning, the models were finalized with the *finalize_model()* function and tested on a hold-out validation set using *predict_model()* to assess their generalization capacity on unseen data. For each tested model, diagnostic tables are reported in the subchapters 4.2.1– 4.2.6 from Chapter 4.2 (p.66-74).

This experimental setup highlighted a fundamental challenge in real-world credit scoring: training directly on imbalanced datasets tends to bias models toward the majority class ("Paid"), undermining their ability to detect high-risk borrowers. While LightGBM achieved strong results under unbalanced conditions, achieving both an accuracy and recall of 0.9656, and an AUC of 0.9423, indicating substantial discriminatory power, the imbalance still limited its capacity to correctly classify minority-class instances. To illustrate this issue and explore potential refinements, the next section examines the application of class-balancing techniques. Although balancing did not yield superior overall results, this complementary experiment served to demonstrate the extent to which reweighting the class distribution can influence the model's behavior, particularly in mitigating minority-class underprediction.

3.5.2 Balanced Target Variable: Modeling Techniques

To complement the unbalanced dataset results, a second experiment was conducted on a balanced version of the data. The aim was not to remedy shortcomings, since the unbalanced configuration had already performed well, but to examine whether adjusting class distribution would significantly influence model behavior.

To construct the balanced dataset, the original data were first partitioned into training (70%) and test (30%) subsets using a fixed random seed (*random_state* = 42) to ensure reproducibility. The balancing procedure was applied exclusively to the training set. All 34,504 records labeled as "Default" were preserved, of which 24,083 fell within the training partition. For the majority class ("Paid"), undersampling was employed to retain approximately 20% more cases than the "Default" class, resulting in 28,889 "Paid" observations in the training data. This approach was applied in both PyCaret and the native implementations. This slight asymmetry was deliberately preserved to mitigate the risk of overfitting, which can arise when classes are artificially forced into perfect balance, particularly in real-world credit risk scenarios where such symmetry may not be representative. By maintaining a modest prevalence of the “Paid” class, the training dataset preserved realistic portfolio dynamics while ensuring adequate representation of the minority class (“Default”) for model learning. The untouched test set, in turn, retained its original distribution to provide a reliable and unbiased benchmark for evaluating model generalization.

The PyCaret environment was again initialized using the *setup()* function, with *loan_status* defined as the target variable and all preprocessing configurations kept consistent with the previous experiment to ensure comparability. No additional imbalance correction was activated within PyCaret (*fix_imbalance* = False), as class distribution had already been manually adjusted in the training set. The modeling with native libraries were initialized using the set of hyperparameters originating from PyCaret default configuration.

Model benchmarking was first performed using the *compare_models(sort='Accuracy')* function, which ranks a broad range of classification algorithms using cross-validated performance. As in the previous experiment, LightGBM emerged as the top-performing model in terms of accuracy, followed by Random Forest, AdaBoost, Gradient Boosting, and Extra Trees. Although Logistic Regression performed less effectively than the other models, it was included in the subsequent evaluation steps for the reasons previously outlined.

As in the unbalanced case, for the balanced one each of those six models was subsequently fine-tuned using PyCaret’s *tune_model()* function, where PyCaret assigns an initial set of hyperparameters when creating a model through *create_model()*, which serves as a standardized baseline for comparison with recall specified as the optimization criterion. Moreover, it allows full transparency by enabling users to inspect the hyperparameters of both versions through the *print()* function, facilitating informed interpretation of model behavior. For the unbalanced dataset, the tuned versions of Random Forest, AdaBoost, Gradient Boosting, and Extra Trees achieved better performance, while other classifiers performed best with default configurations. Nonetheless, all models exceeded 90% accuracy in both setups. For the same reason highlighted in the previous section, there was no tuning in the native-library models.

Final performance was evaluated on a held-out test set using the *predict_model()* function. LightGBM again outperformed the other models, achieving an accuracy of 0.9265, a recall of 0.9265, and an AUC of 0.9423. These results confirm its ability to detect the “Default” class with precision and consistency, showing only a slight difference compared to the unbalanced experiment. This balanced performance reinforces models' robustness across different class distributions. Notably, AdaBoost (accuracy and recall of 0.9253 and AUC of 0.9397) also delivered competitive results in both approaches, confirming its effectiveness in scenarios involving moderately imbalanced data when tuned for recall.

The chapter 4.3 presents the diagnostic plots generated for each model, which include:

- Confusion Matrix: To visualize the final distribution of true positives, false positives, true negatives, and false negatives within the test set.
- Feature Importance Plot: PyCaret's feature that identifies which predictors contributed most significantly to classification decisions during model training.

Although the class balancing strategy effectively reduced the dominance of the majority class and promoted more equitable sensitivity across categories, it did not improve predictive performance for the "Default" class in PyCaret, as the unbalanced configuration delivered slightly superior results across most evaluation metrics. In contrast, when implemented in the native libraries, five models achieved a recall of 99%, with Logistic Regression reaching 90%, still high, but the lowest among them. These findings suggest that, within the native framework, the balancing strategy may be more effective than in PyCaret's automated environment, reinforcing that such strategy remains valuable in credit scoring applications, particularly when dealing with more extreme imbalances, model fairness and interpretability are prioritized.

Furthermore, overall classification reliability, as measured by Kappa[23] and Matthews Correlation Coefficient (MCC)[24], was higher in both the unbalanced and balanced settings. Notably, MCC is especially valuable in imbalanced classification problems, as it provides a balanced evaluation even when class sizes differ, making it particularly well suited for credit scoring tasks where defaults are less frequent. These findings are detailed in subchapter 4.1.

[23] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

[24] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.matthews_corrcoef.html

CHAPTER 4

Results and Discussion

This chapter presents the results of the supervised learning experiments described in the previous section, with the objective of evaluating and interpreting the performance of various classification models in credit risk assessment. The analysis begins with a systematic comparison of models trained under unbalanced and moderately balanced target variable scenarios, emphasizing the impact of class distribution and hyperparameter optimization in the final outcome.

Evidence from the systematic literature review highlights consensus that model evaluation in credit scoring should rely on multiple performance metrics such as recall, precision, and AUC-ROC. For instance, Robisco and Martínez (2022) evaluated Random Forest and Extreme Gradient Boosting (XGBoost) models using AUC-ROC as their central metric when applied to a public "Give Me Some Credit" dataset from Kaggle. Similarly, Filchenkov *et al.* (2021) employed the weighted F1-score to balance precision and recall while accounting for class distribution, complementing it with AUC from ROC analysis. This dual reliance on weighted F1 and AUC underscores the importance of integrating multiple evaluation metrics to capture both discriminatory power and robustness in imbalanced datasets.

Following predictive performance evaluation, the chapter examines interpretability, an essential factor in regulated financial environments, through feature importance scores and classification diagnostics. These insights reveal which borrower characteristics most influenced model decision-making, enhancing transparency and aligning with industry expectations for explainable AI in credit decisions.

Results from PyCaret-based experiments are also critically compared with manually coded models using native libraries, providing a nuanced understanding of the trade-offs between modeling efficiency and customization. This comparison validates the robustness of the automated framework while offering practical guidance for institutions on adopting scalable machine learning solutions without compromising performance or governance.

Finally, the chapter contextualizes these findings within the operational realities of financial institutions, reflecting on how supervised learning can enhance credit risk evaluation by improving early default detection, reducing unjustified rejections, and promoting more inclusive, data-driven lending policies. This discussion integrates technical outcomes with business implications, offering a grounded perspective on how ML models can support fairer, more efficient, and risk-sensitive credit decisions.

4.1. Model Performance Evaluation

The evaluation covers both modeling workflows, PyCaret's low-code framework and manually implemented scripts using native libraries, under two experimental setups: with the original unbalanced target variable and with a resampled version adjusted to reduce class imbalance.

The decision to prioritize recall in this study is rooted in the institutional objective of minimizing Type II errors, specifically, the risk of failing to identify borrowers likely to default. In credit risk assessment, such misclassifications carry substantially greater financial consequences than Type I errors, where reliable borrowers are mistakenly flagged as risky. As highlighted by Chang *et al.* (2024), selecting an appropriate performance metric in these contexts is inherently complex and should align with the cost–benefit structure of lending decisions. Their study emphasizes the trade-off between true positives (profit) and false positives (cost), showing that metrics such as the F1-score and precision–recall AUC often provide more meaningful insights than accuracy alone. Although Chang *et al.* (2024) ultimately adopted the F1-score as their guiding metric, their findings reinforce the critical importance of recall in profit-sensitive domains like credit scoring. In line with this perspective, the present study designated recall as both the primary evaluation metric and the optimization target during model tuning, ensuring that the models maximized the detection of high-risk borrowers while maintaining acceptable trade-offs in precision and model complexity.

For each configuration, key performance metrics were computed, including Accuracy, Recall, Precision, F1-score, AUC-ROC, Kappa and Matthews Correlation Coefficient (MCC). These indicators provide a multidimensional view of model effectiveness. The following evaluation metrics were systematically computed and reported for each model across both validation folds and the test set, as part of the standardized performance reporting framework employed throughout the modeling process:

- Accuracy: The proportion of correctly predicted instances (both positive and negative) relative to the total number of observations. Although widely reported, accuracy can be misleading in imbalanced datasets, as a model may appear effective simply by favoring the majority class ("Paid"). In credit risk assessment, this limitation makes accuracy insufficient on its own, since it does not capture the model's ability to identify risky borrowers.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Recall (Sensitivity): Measures the proportion of actual defaulters ("Default" class) that were correctly identified by the model. In this study, recall was prioritized as the main optimization metric, given that failing to detect a defaulter (false negative) can lead to severe financial losses. By focusing on recall for the "Default" class, the evaluation aligns with the institutional objective of minimizing credit risk exposure.

$$Recall = \frac{TP}{TP + FN}$$

- Precision: Refers to the proportion of correctly identified defaulters (true positives for the "Default" class) out of all instances classified as defaulters. In this study, it reflects the model's reliability when flagging a borrower as "Default." Although not the primary optimization metric, precision remains relevant for mitigating the risk of overly conservative lending, where excessive false positives could result in the unjust rejection of creditworthy applicants.

$$Precision = \frac{TP}{TP + FP}$$

- AUC-ROC: The Receiver Operating Characteristic (ROC) curve plots the True Positive Rate (Recall) against the False Positive Rate (FPR) across varying classification thresholds. In this study, the "Default" class was treated as the reference class of interest. Accordingly, the AUC (Area Under Curve) summarizes the model's ability to correctly rank defaulters lower than non-defaulters, providing a measure of overall discriminatory power between the "Paid" and "Default" categories. An AUC close to 1 indicates excellent separation capacity, while 0.5 reflects performance equivalent to random guessing.

$$AUC = \int_0^1 TPR(FPR)\, d(FPR)$$

- F1-Score: The F1-score is the harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives. In this study, it reflects the model's capacity to correctly classify borrowers at risk of default (the minority class), while avoiding excessive misclassification of reliable borrowers. This metric is particularly useful in imbalanced classification settings, where neither precision nor recall alone fully captures performance. A higher F1-score indicates a better overall balance between identifying defaulters and maintaining lending inclusiveness.

$$F1\text{-}Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

- Kappa (κ): Cohen's Kappa quantifies the level of agreement between predicted and actual classes, adjusting for the portion of agreement that could occur by chance. Unlike raw accuracy, Kappa accounts for class imbalance, making it particularly valuable in credit risk assessment where the majority of loans are "Paid." A higher Kappa indicates that the model is providing meaningful predictions beyond chance, reflecting more reliable detection of both default and non-default outcomes.

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

  - To interpret the resulting value of κ,this study adopts the interpretation framework proposed by Landis and Koch (1977), which categorizes the strength of agreement into defined ranges. These ranges are detailed in Table 4.1.

Table 4.1. Interpretation of Cohen's Kappa Statistic

Kappa Score	Strength of Agreement
< 0	Less than chance agreement
0.01–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

- Matthews Correlation Coefficient (MCC): The MCC is a robust evaluation metric that incorporates all four components of the confusion matrix (true positives, true negatives, false positives, and false negatives) into a single value. It ranges from -1 to +1, where +1 denotes perfect classification, 0 reflects performance equivalent to random guessing, and -1 indicates complete disagreement between predictions and actual outcomes. Unlike accuracy, MCC remains reliable under class imbalance, as it provides a balanced assessment of model performance across both classes. This makes it particularly relevant in credit scoring applications, where defaults represent the minority class and models must be evaluated on their ability to handle such asymmetry.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

All performance metrics were consistently computed across the training–validation splits and independently evaluated on the held-out test set. The training outcomes are reported in Table 4.2 for the balanced target variable and in Table 4.3 for the unbalanced configuration, both derived from the PyCaret framework. Tables 4.4 and 4.5 present the corresponding results obtained using native model implementations (primarily Scikit-learn, with LightGBM implemented via LGBMClassifier). Collectively, these tables establish a robust basis for the comparative analysis of test set performance discussed in the following subchapters, providing evidence that the models did not overfit. Instead, the strong test set performance demonstrates their ability to generalize beyond the training data rather than simply memorizing it.

Table 4.2. Model Performance on Balanced Training Set (PyCaret)

Model	**Accuracy**	**Recall**	**Precision**	**F1-Score**	**AUC**	**Kappa**	**MCC**	**TT(Sec)**
LightGBM	0.9669	0.9669	0.9669	0.9658	0.9442	0.8923	0.8974	0.6260
Ada	0.9668	0.9668	0.9668	0.9657	0.9430	0.8920	0.8972	0.6140
GBC	0.9667	0.9667	0.9667	0.9656	0.9447	0.8917	0.8968	2.0620
RF	0.9665	0.9665	0.9665	0.9654	0.9370	0.8911	0.8961	1.0850
ET	0.9635	0.9660	0.9635	0.9625	0.9331	0.8822	0.8862	0.9840
LR	0.9507	0.9635	0.9507	0.9500	0.9149	0.8448	0.8459	1.2820

Table 4.3. Model Performance on Unbalanced Training Set (PyCaret)

Model	**Accuracy**	**Recall**	**Precision**	**F1-Score**	**AUC**	**Kappa**	**MCC**	**TT(Sec)**
LightGBM	0.9667	0.9667	0.9679	0.9655	0.9452	0.8915	0.8967	0.6640
Ada	0.9665	0.9665	0.9678	0.9653	0.9430	0.8909	0.8961	0.5560
GBC	0.9664	0.9664	0.9677	0.9653	0.9444	0.8906	0.8959	1.9570
RF	0.9661	0.9661	0.9673	0.9649	0.9372	0.8896	0.8947	1.0620
ET	0.9630	0.9630	0.9636	0.9619	0.9326	0.8806	0.8846	0.8590
LR	0.9500	0.9500	0.9493	0.9493	0.9130	0.8427	0.8437	1.1710

Table 4.4. Model Performance on Balanced Training Set (Native)

Model	**Accuracy**	**Recall**	**Precision**	**F1-Score**	**AUC**	**Kappa**	**MCC**	**TT(Sec)**
LightGBM	0.9671	0.9998	0.9603	0.9797	0.9676	0.8938	0.8988	0.7143

Ada	0.9266	0.8817	0.9998	0.9370	0.9433	0.8499	0.8596	1.4506
GBC	0.9268	0.8821	0.9997	0.9372	0.9469	0.8504	0.8600	5.9760
RF	0.9998	0.9996	1.0	0.9998	0.9999	0.9996	0.9996	0.6707
ET	0.9998	1.0	0.9996	0.9998	0.9999	0.9996	0.9996	0.7020
LR	0.9171	0.8818	0.9796	0.9281	0.9147	0.8309	0.8373	0.1484

Table 4.5. Model Performance on Unbalanced Training Set (Native)

Model	**Accuracy**	**Recall**	**Precision**	**F1-Score**	**AUC**	**Kappa**	**MCC**	**TT(Sec)**
LightGBM	0.9671	0.9603	0.9998	0.9797	0.9676	0.8938	0.8988	0.7184
Ada	0.9665	0.9596	0.9999	0.9793	0.9437	0.8917	0.8969	3.4079
GBC	0.9664	0.9597	0.9995	0.9792	0.9447	0.8913	0.8964	14.5517
RF	0.9998	0.9998	0.9999	0.9999	0.9999	0.9995	0.9995	1.9528
ET	0.9999	1.0	0.9998	0.9999	0.9999	0.9997	0.9997	1.7039
LR	0.9502	0.9595	0.9784	0.9689	0.9148	0.8440	0.8451	0.2730

4.2. Model Interpretation and Comparative Analysis

This section offers a comprehensive interpretation of the results across all modeling experiments, with metrics consistently reported on the held-out test set, systematically comparing algorithm performance under four distinct configurations: (i) PyCaret with unbalanced data, (ii) PyCaret with balanced data, (iii) native library implementations with unbalanced data, and (iv) native implementations with balanced data. The results are synthesized and visualized to reveal key performance trends, trade-offs, and their practical implications for credit risk classification and decision-making.

4.2.1. LightGBM

Across all experimental setups, LightGBM consistently outperformed the other classifiers, particularly excelling in the metric of recall, which was established as the key optimization target in this study due to its critical role in identifying high-risk borrowers. As detailed in Table 4.6, LightGBM achieved strong recall scores across both PyCaret and native implementations for the test set, under both balanced and unbalanced target variable configurations.

Notably, the unbalanced native model reached the highest recall of 0.9997, followed closely by the balanced native configuration at 0.9982, both surpassing the results observed in the PyCaret workflows. In parallel, the unbalanced PyCaret model delivered a recall of 0.9656, accompanied by robust values in accuracy (0.9656), F1-score (0.9644), MCC (0.8929), and Kappa (0.8873), illustrating the algorithm's strong classification power even without manual tuning in the native configuration.

These findings highlight LightGBM's ability to maintain superior predictive capacity across different experimental settings, confirming its robustness and adaptability for credit risk modeling where the early detection of defaults is essential.

Table 4.6. LightGBM Comparative Results (Test Set)

Library	Class Bal.	Accuracy	Recall	Precision	F1-Score	AUC	Kappa	MCC
PyCaret	Balanced	0.9265	0.9265	0.9344	0.9257	0.9423	0.8497	0.8586
PyCaret	Unbal.	0.9656	0.9656	0.9669	0.9644	0.9423	0.8873	0.8929
Native	Balanced	0.9249	0.9982	0.8798	0.9353	0.9440	0.8469	0.8564
Native	Unbal.	0.9655	0.9997	0.9586	0.9787	0.9434	0.8872	0.8928

The combination of high recall and consistently strong supporting metrics, including precision, AUC-ROC, and MCC, reinforces LightGBM's discriminatory capacity and reliability in correctly identifying default cases, which is essential in credit scoring applications. The results further demonstrate the model's robustness to class imbalance, as evidenced by the minimal performance differences observed between the balanced and unbalanced datasets. Although PyCaret's automated pipeline yielded slightly lower metrics in certain configurations, these differences remained marginal, indicating that automated workflows can deliver competitive results in this experimental setting without requiring extensive manual intervention. Nonetheless, it is important to note that these findings are specific to the dataset and experimental design adopted in this study and should not be generalized to all production environments, where additional factors such as data heterogeneity, system integration, and operational constraints may substantially influence performance.

These findings are further substantiated by the technical underpinnings of the algorithm itself. As described by Li *et al.* (2024), LightGBM incorporates several architectural innovations that account for its superior predictive capacity and computational efficiency. Built as a gradient-boosting framework optimized for classification, ranking, and regression tasks, LightGBM employs a leaf-wise tree growth strategy, which prioritizes leaves with the highest potential for loss reduction rather than expanding level by level as in traditional approaches. It also leverages histogram-based optimization, which reduces the search space during node splits and significantly lowers memory consumption. These design efficiencies make the algorithm particularly advantageous in large-scale or resource-constrained environments. Compared to other popular boosting frameworks such as XGBoost and CatBoost, LightGBM offers faster training times, lower memory usage, and competitive, often superior, accuracy, rendering it especially suitable for credit risk modeling where both speed and predictive precision are critical.

4.2.2. AdaBoost Classifier

Following LightGBM, AdaBoost emerged as the second-best performing model in both the PyCaret and native in the testing set, demonstrating particularly strong results across all major evaluation metrics in the test set. As shown in Table 4.7, in the native implementation, the unbalanced version again delivered the best recall within the PyCaret framework. Particularly noteworthy is the model's high recall across both balanced and unbalanced setups. This confirms AdaBoost's effectiveness in identifying defaulters, which aligns with the recall-focused optimization strategy adopted in this study.

Table 4.7. AdaBoost Classifier Comparative Results (Test Set)

Library	**Class Bal.**	**Accuracy**	**Recall**	**Precision**	**F1-Score**	**AUC**	**Kappa**	**MCC**
PyCaret	Balanced	0.9253	0.9253	0.9342	0.9243	0.9397	0.8471	0.8571
PyCaret	Unbal.	0.9654	0.9654	0.9668	0.9642	0.9402	0.8868	0.8924
Native	Balanced	0.9236	0.9997	0.8769	0.9343	0.9394	0.8440	0.8544
Native	Unbal.	0.9654	0.9997	0.9585	0.9787	0.9400	0.8868	0.8924

When deployed under the unbalanced configuration with native code, AdaBoost achieved its best performance overall, recording a Recall of 0.9997, a F1-score of 0.9787, AUC-ROC of 0.9400, Kappa of 0.8868, and a MCC of 0.8924. These values are highly comparable to the top-performing model LightGBM, reinforcing AdaBoost's robustness in handling class imbalance without the need for synthetic resampling. Similarly, even in the balanced setting, AdaBoost preserved competitive metrics (e.g., recall = 0.9997, F1 = 0.9343), with slightly lower precision and MCC due to the increased number of false positives introduced by the rebalancing strategy.

A noteworthy observation is that the performance improvements from balancing the dataset were marginal for AdaBoost. This suggests that the model may already be effectively capturing signals from the minority class due to its inherent sequential weighting mechanism. As explained by Aruleba & Sun (2024), AdaBoost builds a strong predictive model by iteratively adjusting the weights of misclassified instances, placing greater emphasis on those that are more difficult to classify while reducing the influence of those correctly classified in earlier rounds. The same authors also highlight that both Random Forest and AdaBoost benefit from ensemble learning strategies, where the aggregation of multiple decision trees helps to reduce variance and bias, ultimately enhancing predictive performance.

4.2.3. Gradient Boosting Classifier

Gradient Boosting classifiers also demonstrated competitive and stable performance across all experimental configurations. In particular, the model reached its highest scores under the native implementation with unbalanced data, achieving a recall of 0.9997, accuracy of 0.9653, F1-score of 0.9786, and MCC of 0.8923, with AUC-ROC and Kappa at 0.9415 and 0.8866, respectively. These results are outlined in Table 4.8, and confirm the model's ability to correctly identify virtually all default cases in the test set without substantial degradation in other metrics. 0.9997, F1 = 0.9786

Table 4.8. Gradient Boosting Classifier Comparative Results (Test Set)

Library	Class Bal.	Accuracy	Recall	Precision	F1-Score	AUC	Kappa	MCC
PyCaret (tuned)	Balanced	0.9263	0.9263	0.9336	0.9255	0.9434	0.8496	0.8578
PyCaret	Unbal.	0.9654	0.9654	0.9668	0.9642	0.9375	0.8868	0.8924

(tuned)								
Native	Balanced	0.9236	0.9995	0.8770	0.9343	0.9408	0.8440	0.8543
Native	Unbal.	0.9653	0.9997	0.9585	0.9786	0.9415	0.8866	0.8923

Under the PyCaret framework, Gradient Boosting performed nearly as well. The unbalanced configuration yielded a recall and accuracy of 0.9654,the highest precision of 0.9668, and an F1-score of 0.9642. Although the balanced PyCaret setup registered slightly lower values, recall of 0.9263 and F1-score of 0.9255, it still delivered consistent and high-performing results across the board.

An additional finding is that Gradient Boosting was the model, among all tested classifiers so far, for which PyCaret's tuning process produced a marginal yet consistent improvement across evaluation metrics, considering both balanced and unbalanced scenarios. In the validation folds, tuning specifically targeted at optimizing recall led to a slight but measurable gain, from 0.9666 to 0.9682 in recall and from 0.9642 to 0.9657 in F1-score, along with similar minor increases in accuracy, AUC-ROC, Kappa, and MCC. While these differences are modest, they indicate that Gradient Boosting is particularly sensitive to hyperparameter adjustments, and that PyCaret's automated tuning mechanism can extract incremental gains even in well-performing models. These results reinforce the known strengths of Gradient Boosting in credit scoring contexts, particularly its capacity to capture complex patterns and subtle interactions among features due to its sequential learning process. As highlighted in the literature, including De Lange *et al.* (2022), boosting methods like this one iteratively improve performance by correcting the errors of previous trees, effectively reducing bias and enhancing generalization when properly tuned.

However, it is also important to note that while Gradient Boosting shows remarkable results in terms of recall and F1-score, its precision values are consistently lower than those of LightGBM or AdaBoost across most configurations. This suggests that although the model is excellent at capturing true defaulters, it may include a slightly higher number of false positives, a trade-off that must be managed depending on institutional risk tolerance.

Overall, Gradient Boosting offers a highly effective compromise between sensitivity and general performance, and its relatively consistent behavior across both frameworks and class distributions reinforces its suitability for operational credit risk prediction pipelines.

4.2.4. Random Forest

As shown in Table 4.9, in the native implementation, the unbalanced version again delivered the best recall (0.9990), a value nearly equivalent to perfect sensitivity. This was paired with a precision of 0.9584, F1-score of 0.9783, AUC-ROC of 0.9351, and equally strong MCC (0.8905) and Kappa (0.8851), perfectly aligned with the unbalanced PyCaret setup. The performance parity across these two implementations highlights the minimal impact of framework-specific workflows on Random Forest, reaffirming its suitability for production environments where reproducibility is key. The balanced native model also reached impressive levels of recall (0.9933) and provided a solid F1-score (0.9321), though it slightly underperformed in precision and MCC.

Table 4.9. Random Forest Comparative Results (Test Set)

Library	Class Bal.	Accuracy	Recall	Precision	F1-Score	AUC	Kappa	MCC
PyCaret	Balanced	0.9253	0.9253	0.9342	0.9243	0.9362	0.8471	0.8571
PyCaret	Unbal.	0.9649	0.9649	0.9661	0.9636	0.9339	0.8851	0.8905
Native	Balanced	0.9214	0.9933	0.8780	0.9321	0.9355	0.8396	0.8487
Native	Unbal.	0.9648	0.9990	0.9584	0.9783	0.9351	0.8851	0.8905

These results highlight Random Forest's strong generalization capabilities across varying class distributions, consistently delivering high recall while maintaining competitive performance across other key metrics. As emphasized by De Lange *et al.* (2022), both Random Forest and Gradient Boosting emerge as essential techniques due to their ability to aggregate outputs from multiple weak learners, enhancing model robustness and reducing the risk of overfitting, an especially valuable trait in credit risk prediction where model stability and predictive confidence are critical.

The solid performance of Random Forest observed in this study also aligns with findings from Robisco and Martínez (2022), who evaluated several machine learning models, including Penalized Logistic Regression (LASSO), Classification and Regression Trees (CART), Random Forest, Extreme Gradient Boosting (XGBoost), and Deep Neural Networks (DNN), using the public "Give Me Some Credit" dataset from Kaggle.com, composed of 120,000 loan records and 22 explanatory variables (e.g., age, monthly income, and number of delinquent loans). Their evaluation considered prediction stability, transparency, latency, and interpretability. Among the models tested, XGBoost delivered the highest AUC-ROC (85.3%), followed closely by Random Forest with 84.2%, and DNN with 81.7%, confirming the ensemble's ability as a bagging ensemble method that builds multiple decision trees and aggregates their outputs. Random Forest inherently mitigates overfitting and handles noise well, which explains its stability across both setups. Its ability to perform implicit feature selection and capture nonlinear relationships also contributes to its strong predictive performance in credit scoring contexts.

4.2.5. Extra Trees Classifier

Han *et al.* (2024) refer the Extra Trees Classifier is among the most widely adopted non-linear predictive techniques, together with Gradient Boosting and Random Forest. In this study context, the Extra Trees Classifier demonstrated consistent and competitive performance across all configurations for the testing set, with slight variations depending on the framework and data balancing approach. Notably in Table 4.10, the PyCaret (tuned) configuration with unbalanced data emerged as the best performer in this group, achieving a recall of 0.9647, precision of 0.9658, F1-score of 0.9634, and MCC of 0.8898. These results reflect the model's ability to detect the minority "Default" class effectively without compromising predictive precision or overall classification stability.

Table 4.10. Extra Trees Classifier Comparative Results (Test Set)

Library	**Class Bal.**	**Accuracy**	**Recall**	**Precision**	**F1-Score**	**AUC**	**Kappa**	**MCC**
PyCaret (tuned)	Balanced	0.9254	0.9254	0.9343	0.9245	0.9435	0.8473	0.8573
PyCaret (tuned)	Unbal.	0.9647	0.9647	0.9658	0.9634	0.9268	0.8845	0.8898
Native	Balanced	0.9114	0.9714	0.8783	0.9225	0.9288	0.8196	0.8253
Native	Unbal.	0.9600	0.9926	0.9585	0.9752	0.9288	0.8705	0.8742

Interestingly, the native implementation with unbalanced data yielded the highest recall (0.9926) across all Extra Trees runs, confirming the strong model' sensitivity to identifying defaulters. However, this came at the expense of slightly reduced precision (0.9585) and a lower Matthews Correlation Coefficient (0.8742), indicating a potential trade-off between maximizing true positive rates and overall model robustness.

On the other hand, balanced datasets in both frameworks led to moderate recall gains but slightly lower overall scores on other metrics, suggesting that the Extra Trees algorithm, much like AdaBoost, may already be intrinsically robust to class imbalance due to its ensemble structure. This behavior aligns with its known strength in reducing overfitting and variance by aggregating multiple uncorrelated decision trees.

4.2.6. Logistic Regression

Logistic Regression (LR) has long been the most widely adopted method for credit scoring and is often regarded as the benchmark approach in this domain. As a parametric technique, LR offers good levels of transparency and interpretability, making it particularly suitable for applications requiring explainable decision-making. It is also computationally efficient, demanding fewer resources and smaller training datasets compared to more complex models. Parametric models like LR establish a direct and linear relationship between input features and the target variable, which facilitates a clear understanding of feature contributions. Nonetheless, their primary limitation lies in their constrained predictive capacity, especially when dealing with complex, non-linear patterns in the data, an aspect often cited as their main drawback in contemporary machine learning contexts (Han *et al.*, 2024).

Despite not ranking among the top five performing models in the PyCaret-based experiments, Logistic Regression was included in this study due to its established relevance in financial contexts. It serves as a foundational benchmark, offering a valuable point of comparison against more advanced ensemble and tree-based techniques.

Despite being a linear method, it exhibited solid performance across all configurations, especially after tuning. As illustrated in Table 4.11, under the unbalanced setup using native pipeline in the test set, the model achieved an Accuracy of 0.9491, the highest Recall of 0.9782, and F1-score of 0.9585, with a Matthews Correlation Coefficient (MCC) of 0.8400 and Kappa of 0.8389, closely matching the results obtained via PyCaret tuned implementation. These high scores suggest that even Logistic Regression can remain competitive even when benchmarked against more complex ensemble methods. Furthermore, in the balanced scenario, while the overall performance metrics were understandably lower (Accuracy and Recall at 0.8776), they remained within acceptable ranges, underscoring the model's consistency and generalization ability.

Table 4.11. Logistic Regression Comparative Results (Test Set)

Library	Class Bal.	Accuracy	Recall	Precision	F1-Score	AUC	Kappa	MCC
PyCaret (tuned)	Balanced	0.8776	0.8776	0.8776	0.8774	0.9154	0.7523	0.7526
PyCaret (tuned)	Unbal.	0.9491	0.9491	0.9483	0.9483	0.9112	0.8389	0.8400
Native	Balanced	0.8773	0.9033	0.8750	0.8889	0.9127	0.7521	0.7526
Native	Unbal.	0.9491	0.9782	0.9585	0.9683	0.9110	0.8389	0.8400

This performance aligns with expectations, as Logistic Regression has historically been a preferred choice in regulated environments due to its high explainability and transparent decision boundaries. Indeed, such characteristics are especially critical in credit risk assessment contexts, where stakeholders demand not only predictive accuracy but also clarity in how input features contribute to risk classification. As noted by recent literature, linear models such as Logistic Regression remain widely used in domains where model interpretability is non-negotiable, including finance and healthcare. Their adoption ensures that the balance between fairness, accountability, and predictive power is preserved, justifying the Logistic Regression as a dependable option for baseline credit scoring solutions.

4.3. Comparative Analysis of Model Performance and Trade-offs

This section presents a comprehensive evaluation of the classifiers tested in this study, considering both predictive performance and computational efficiency, as well as their interpretability and practical implications in the context of credit risk assessment. The analysis combines the efficiency-oriented perspective of training times with the global comparative results derived from confusion matrices and feature importance plots. By integrating these dimensions, the discussion highlights the trade-offs that financial institutions face when adopting machine learning models for credit scoring.

4.3.1 Predictive Performance and Efficiency Trade-offs

To assess the predictive performance, training time, the measured TT(Sec) metric from PyCaret, contributes to a critical role in selecting models for real-world applications. This study reveals notable trade-offs between accuracy and computational efficiency across the evaluated classifiers. LightGBM emerged as the most well-rounded model, combining outstanding predictive performance with consistent training and test times of approximately 0.71 and 0.61 seconds, which remained stable across both the balanced and unbalanced scenarios. This robustness positions it as the most practical choice in both performance and deployment terms. Similarly, AdaBoost demonstrated high accuracy and recall with the fastest training time among top performers (0.55-0.60s), slightly edging out LightGBM in efficiency. For this last model, AdaBoost, it is worth noting that the fastest training time (TT, in seconds) was achieved when using the PyCaret framework, whereas the native implementation required approximately 1–3 seconds, while for LightGBM in native this time was 0.73-0.83 seconds.

Conversely, Gradient Boosting, despite its strong classification capabilities, required over ≈2 seconds to train in the PyCaret, while in the native it took 5 seconds for the balanced class and 14 seconds for the unbalanced one, highlighting a limitation in computationally constrained environments or scenarios that demand frequent retraining. Random Forest and Extra Trees Classifier, although slightly behind in performance compared to LightGBM and AdaBoost, offered a more balanced trade-off: both achieved solid predictive metrics while maintaining reasonable training times (RF ≈0.60–1.0s and ET ≈0.90–1.0s), making them viable candidates consistency are also desired.

Logistic Regression remained the fastest model across all experiments (with training times as low as 0.1s), but this trend was not replicated in PyCaret, where training times reached 1.1s for the balanced dataset and 1.2s for the unbalanced one. Its comparatively lower predictive power compared to the other tested models, positioned it behind ensemble methods in high-stakes credit scoring tasks. These results underscore the importance of evaluating models not only in terms of predictive performance but also with respect to computational efficiency, particularly in operational contexts where retraining frequency, infrastructure constraints, and response time are critical factors.

4.3.2 Global Comparative Results and Interpretability

The conducted comparative analysis across all modeling configurations further clarifies the strengths and limitations of each model. This evaluation incorporates the confusion matrix derived from the test set, enabling an assessment of each model's capability to generalize unseen data. In addition, the analysis leverages the feature importance plots generated through the *plot_model()* function from PyCaret, providing a straightforward and accessible means of interpreting the relative contribution of individual predictors.

LightGBM, the top-performing model across all scenarios, demonstrated exceptional classification behavior in both training and test sets, particularly in the unbalanced native configuration. For instance, the confusion matrix for the test set, presented in Figure 4.1, revealed only 9 false negatives (out of more than 40,000 paid loans), while correctly identifying 8,550 defaults and misclassifying 1,719 default cases as paid. This trade-off emphasizes the model's high recall for the default class. The feature importance plot, shown in Figure 4.2, supports this performance, highlighting outstanding principal invested (*out_prncp*), total payments (*total_rec_int, total_pymnt*), and interest rate (*int_rate*) as the most influential features.

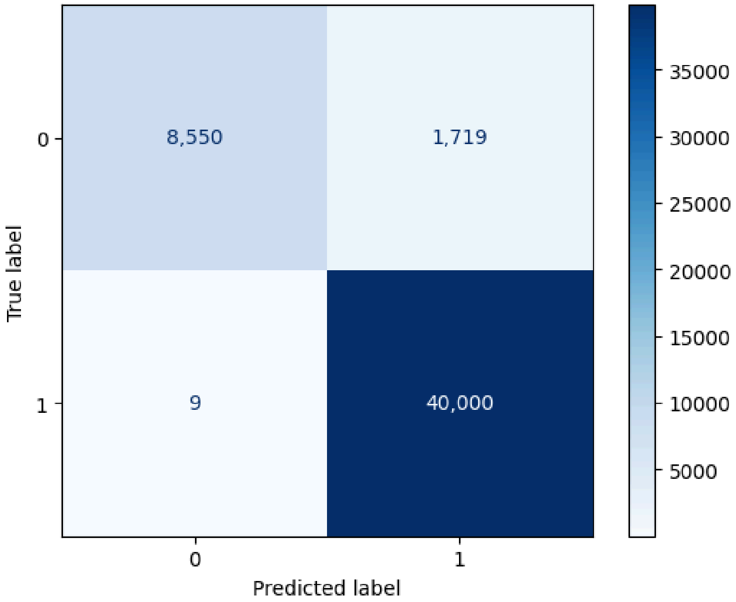

Figure 4.1. Confusion Matrix — LightGBM (Unbalanced Test)

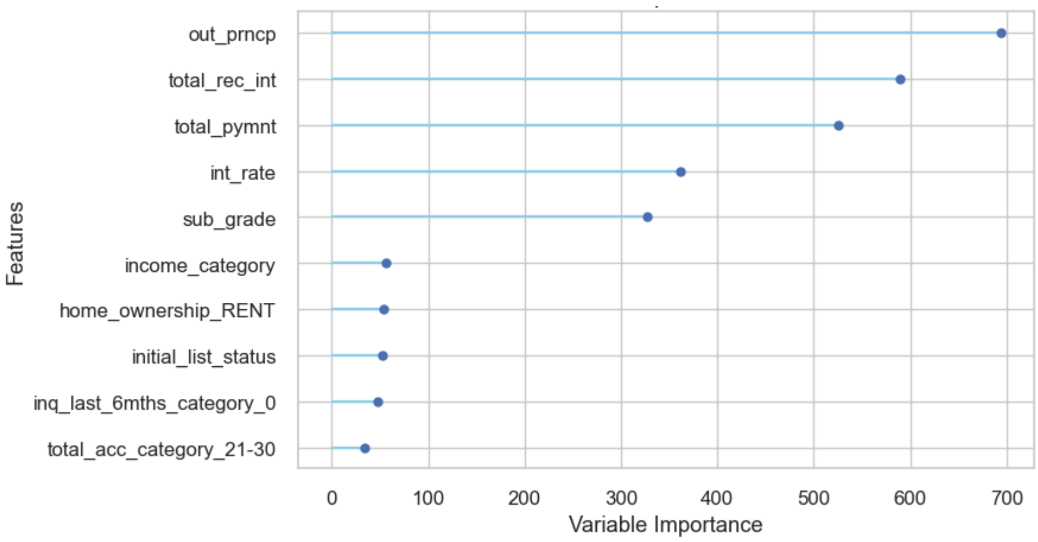

Figure 4.2. Feature Importance Plot — LightGBM (Unbalanced Training)

AdaBoost achieved near-perfect recall, correctly identifying 8,543 defaults while misclassifying only 8 paid loans as defaults. However, this strong performance came at the expense of 1,726 false positives, where defaulted cases were incorrectly predicted as paid, as illustrated in Figure 4.3. This trade-off reflects a conservative orientation consistent with the prudential credit policies of financial institutions.

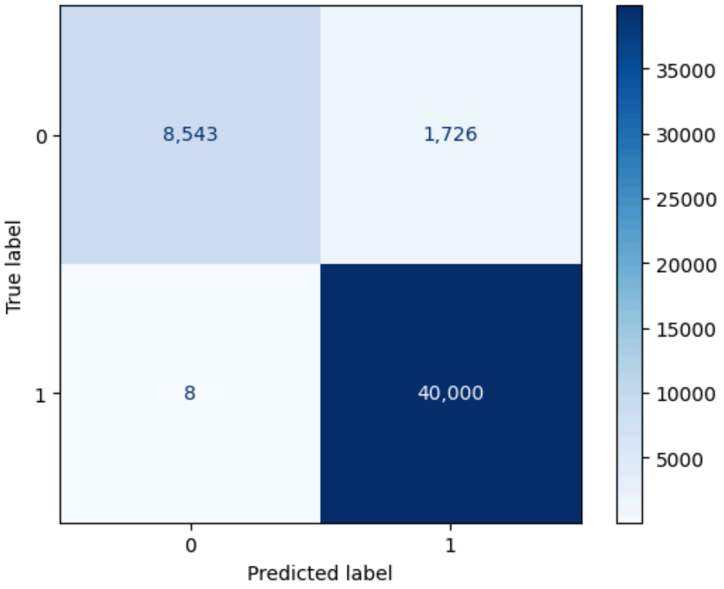

Figure 4.3. Confusion Matrix — AdaBoost (Unbalanced Test)

Its reliance on *sub_grade*, *total_rec_int*, *total_pymnt*, *out_prncp*, and *int_rate* as top predictors reflects both structural and behavioral components of risk assessment, and shown in the feature plot in the Figure 4.4.

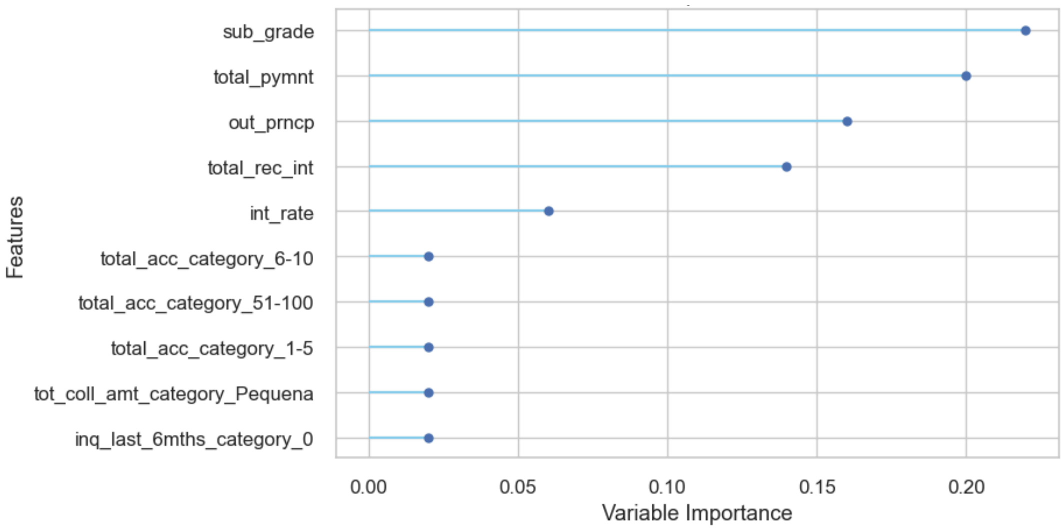

Figure 4.4. Feature Importance Plot — AdaBoost (Unbalanced Training)

Gradient Boosting also excelled in predictive power, with only 10 false negatives under the unbalanced native configuration, as outlined in the Figure 4.5. However, its longer training time positions it as less attractive for frequent retraining environments.

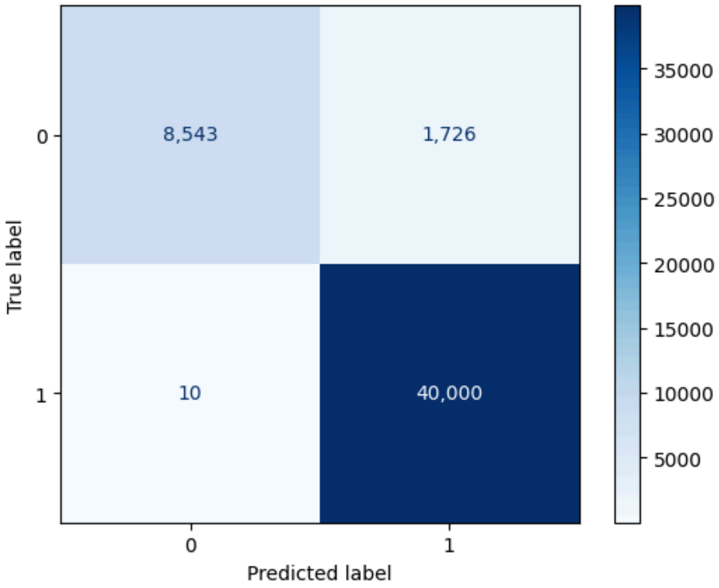

Figure 4.5. Confusion Matrix — Gradient Boosting (Unbalanced Test**)**

Table 4.6 outlines the feature importance plot for the Gradient Booting model, where the variable *out_prncp* has an importance score very close to 1.0, meaning the model relied almost entirely on this single variable to make predictions.

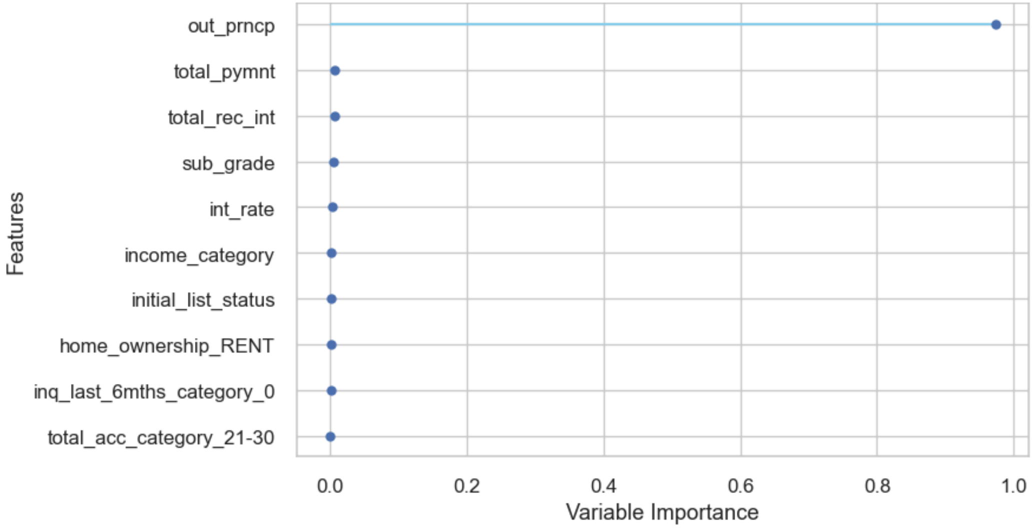

Figure 4.6. Feature Importance Plot — Gradient Boosting (Unbalanced Training)

For Random Forest, particularly in the unbalanced native configuration, where it combined consistent recall with a reasonable training time. Its confusion matrix, presented in Figure 4.7 shows 36 false negatives (paid loans misclassified as defaults), the highest value in this scenario, while correctly identifying 8,543 defaults. This balance underscores the model's suitability for risk-averse applications, where minimizing undetected defaults remains a critical priority.

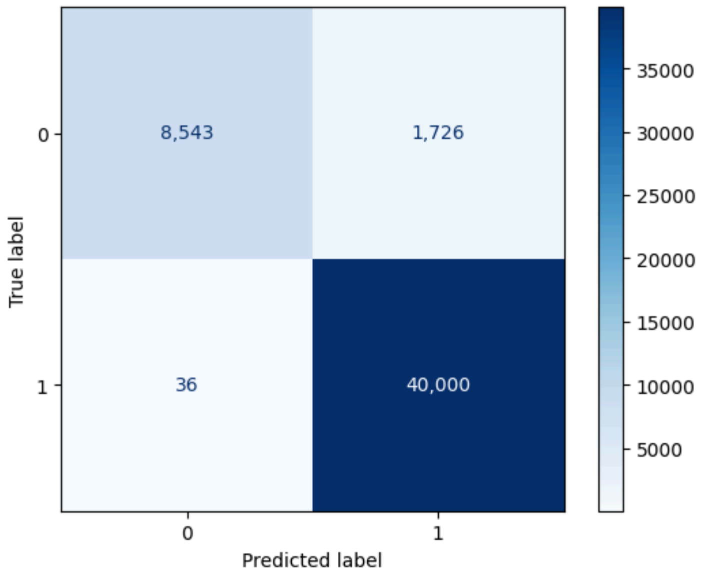

Figure 4.7. Confusion Matrix — Random Forest (Unbalanced Test

Feature importance plot presented in the Figure 4.8 reveals some alignment with LightGBM, with *int_rate*, *out_prncp*, and *total_pymnt* as central variables, alongside behavioral indicators such as credit inquiries.

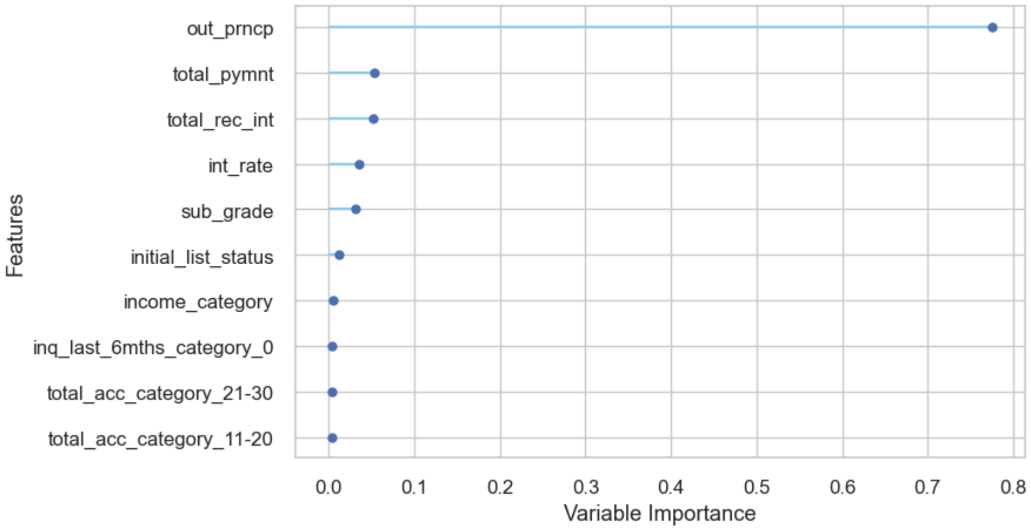

Figure 4.8. Feature Importance Plot — Random Forest (Unbalanced Training)

The Extra Trees model, as illustrated in its confusion matrix (Figure 4.9) from the PyCaret (tuned) configuration with unbalanced data, revealed a conservative classification behavior. Out of more than 40,000 paid loans, 184 were misclassified as defaults (false negatives), showing the prioritization of the recall, achieving approximately 99.5% for the majority class. However, this robustness came at the expense of 1,714 false positives, where defaulted loans were incorrectly classified as paid. This trade-off highlights the model's orientation towards minimizing undetected defaults, a valuable feature for risk-averse financial institutions.

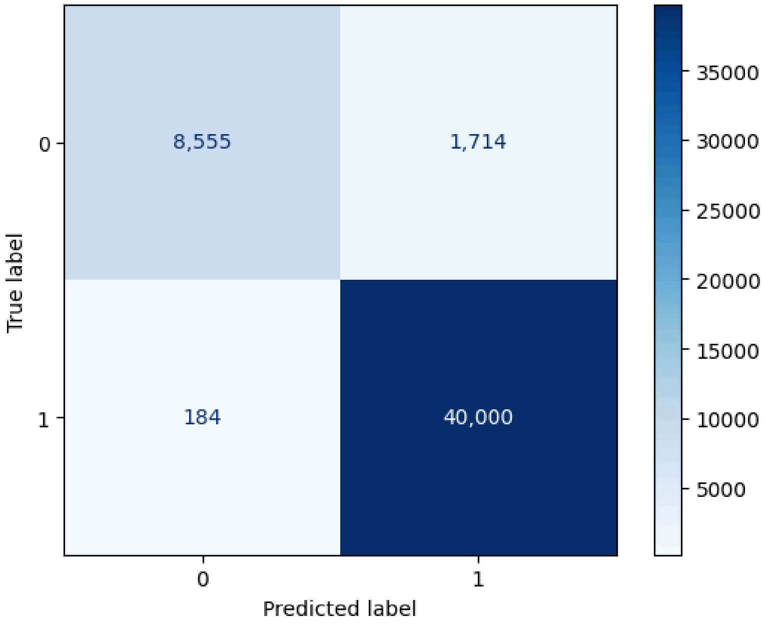

Figure 4.9. Confusion Matrix — Extra Trees (Unbalanced Test)

Nevertheless, this conservative stance may inadvertently penalize underbanked populations, whose incomplete or irregular credit histories heighten the risk of being misclassified as high-risk borrowers. This concern is illustrated in the feature importance plot (Figure 4.10), where, in the case of Gradient Boosting, the model relied almost exclusively on a single variable (*out_prncp*) to drive its predictions.

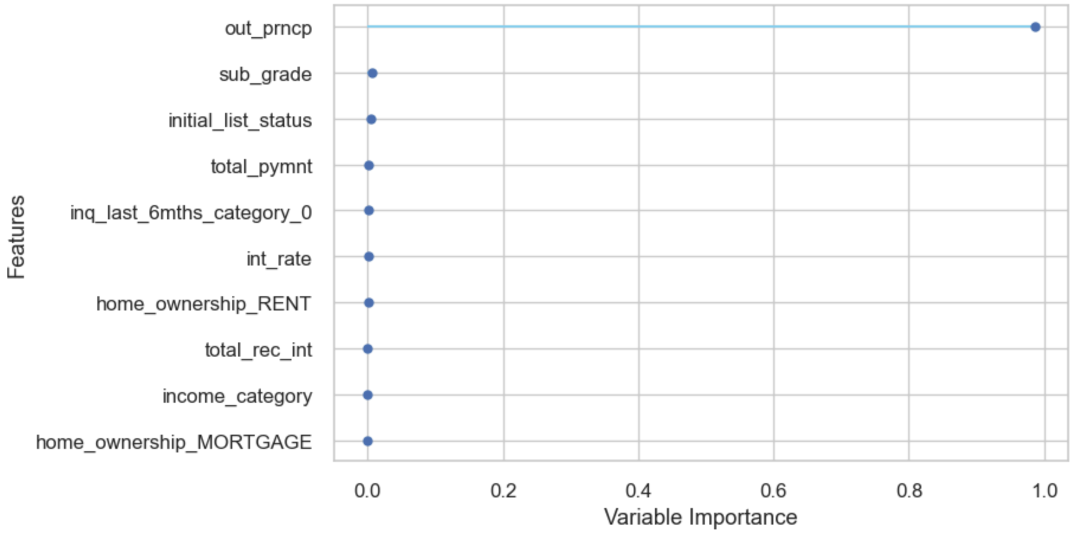

Figure 4.10. Feature Importance Plot — Extra Trees (Unbalanced Training)

Finally, the confusion matrix for Logistic Regression (Figure 4.11) highlights a distinctive performance profile. Among all classifiers, it was the model that misclassified the lowest number of defaults as paid loans (1,686), performing slightly better than ensemble methods such as LightGBM and AdaBoost, which showed higher counts above 1,700. This reflects a conservative orientation consistent with prudential banking policies, as it reduces the risk of undetected defaults. However, this came at the cost of a comparatively larger number of false negatives (867 paid loans wrongly classified as defaults), which can unfairly exclude creditworthy borrowers from access to financial services. This trade-off illustrates Logistic Regression's strengths in prudential risk management but also its limitations regarding financial inclusion, which may be affecting underbanked populations in particular. While it correctly identified the majority of paid loans (39,026) and captured 8,583 defaults, its higher exclusion rate of good borrowers underscores its limited suitability in contexts where equitable access to credit is a priority alongside risk mitigation.

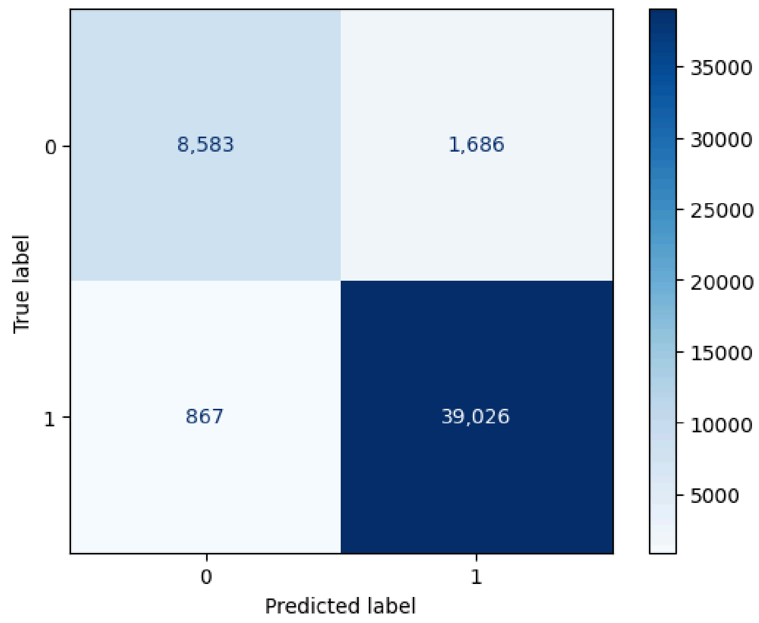

Figure 4.11. Confusion Matrix — Logistic Regression (Unbalanced Test)

Variables such as income category, housing status, and credit inquiries directly correspond to the assessment dimensions formally required by financial institutions, namely repayment capacity (*total_pymnt*), indebtedness (*out_prncp*, *int_rate*), and borrower profile (*income_category*, *home_ownership*, *inq_last_6mths*). Its reliance on traditionally accepted variables (as detailed in the feature importance plot in Figure 4.12), and inquiry counts, reinforces its role as a defensible baseline in regulated contexts, particularly under European frameworks such as the GDPR and the EBA loan origination guidelines.

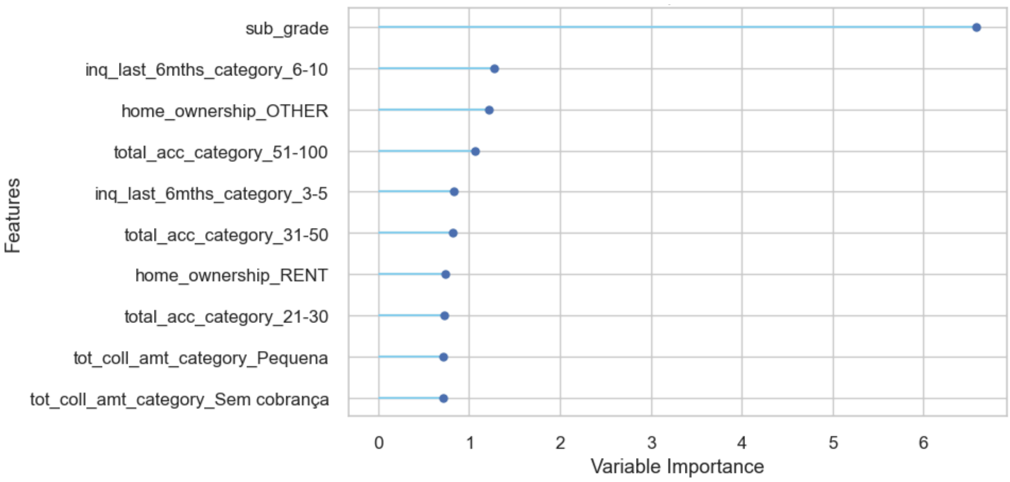

Figure 4.12. Feature Importance Plot — Logistic Regression (Unbalanced Training)

4.3.3 Synthesis and Implications

The comparative analysis highlights clear trade-offs between accuracy, interpretability, computational efficiency, and inclusiveness across all classifiers. LightGBM and AdaBoost consistently achieved the strongest predictive performance, particularly in minimizing false negatives, thereby ensuring a more reliable identification of borrowers at risk of default. Both models also exhibited greater diversity in the features driving their predictive power, as evidenced in the feature importance plots. This characteristic reinforces their robustness in credit risk modeling, since they combine strong predictive performance with reasonable training time and acceptable interpretability.

By contrast, Gradient Boosting, Random Forest, and Extra Trees delivered solid performance metrics, generally above 90% across evaluation measures, yet their computational requirements were comparatively higher, particularly in the case of Gradient Boosting. Moreover, these models demonstrated a narrower reliance on a limited set of dominant features, which, while effective in certain contexts, risks amplifying exclusionary effects for unbanked or underbanked populations. When borrower information is sparse or incomplete, over-reliance on a single variable may undermine the fairness and generalizability of credit assessments, emphasizing the need for models that integrate a broader spectrum of predictors.

Importantly, the feature importance analysis revealed that the models consistently prioritized variables directly aligned with the Bank of Portugal's supervisory framework, including repayment capacity (*total_pymnt*), indebtedness (*out_prncp*, *int_rate*), and borrower profile characteristics (*income_category*, *home_ownership*, *inq_last_mths*). This alignment suggests that machine learning can complement and reinforce established evaluation practices, while also expanding their reach by incorporating behavioral and structural dimensions of credit risk. In doing so, ML-driven models hold the potential to improve financial inclusion, particularly for underbanked segments, while remaining consistent with regulatory expectations.

These findings underscore the strategic importance of balancing predictive accuracy with interpretability and inclusiveness, demonstrating that technological innovation in credit scoring should be pursued in tandem with governance and compliance principles. This perspective provides a foundation for final conclusions and the practical recommendations developed in Chapter 5.

CHAPTER 5

Conclusion

5.1. Summarize Key Findings

Based on the in-depth evaluation of all models, a clear conclusion emerges regarding their relative suitability for credit risk management, with implications that may also inform future discussions within the Portuguese banking sector. Ensemble methods such as LightGBM, AdaBoost, Gradient Boosting, Random Forest, and Extra Trees consistently achieved outstanding results in recall, accuracy, and AUC, although each exhibited distinct trade-offs between false positives and false negatives, training time, and interpretability.

To address these limitations, interpretability tools such as SHAP and LIME have been widely explored. However, this study opted for the potential of the PyCaret framework, which, despite being relatively new, provides straightforward functions to enhance interpretability. In particular, the feature importance plots facilitated the identification of the most impactful variables driving model performance, serving as a cornerstone for understanding the factors that shaped the final results.

The use of PyCaret also proved advantageous due to its practicality and efficiency. It delivered the shortest training time (TT in seconds) across all models, even those that typically require longer runtimes when trained with native libraries. The framework enabled the simultaneous training and ranking of multiple models, significantly streamlining the comparative analysis of algorithms. In addition, its built-in tuning functions allowed optimization to be directly aligned with the chosen metric, recall, prioritized to minimize Type II errors in accordance with this study's objectives.

Among the models, LightGBM stood out as the best-performing overall, combining excellent recall, acceptable training times, and a feature importance profile distributed across multiple conventional creditworthiness indicators rather than concentrating on a single variable. AdaBoost similarly showed near-perfect recall and effectively integrated financial and behavioral features, making it a highly competitive option in high-risk settings where default minimization is paramount. Gradient Boosting, Random Forest, and Extra Trees also delivered very strong results but remained constrained by higher computational demands and reduced interpretability, which limit their suitability where transparency and inclusiveness are critical.

By contrast, Logistic Regression retains substantial value. While it was the most "exclusive" model toward non-defaulter borrowers, misclassifying more "Paid" loans as defaults, it continues to deliver predictive utility and remains highly aligned with regulatory expectations due to its transparency and explainability. For this reason, ensemble models may be best positioned for pilot deployment in innovation labs or parallel testing environments, while Logistic Regression still offers sufficient predictive value for immediate integration into production contexts, particularly for reliably identifying defaulters.

Beyond technical performance, the findings also carry broader implications for consumer credit and financial inclusion. As highlighted in Chapter 1, traditional credit scoring often struggles to assess individuals with sparse or non-traditional credit histories, frequently excluding otherwise creditworthy applicants. Other illustrative cases include newly arrived immigrants or recent graduates seeking financing for vehicles or employment-related expenses, who may lack extensive credit histories despite demonstrating stable income and repayment capacity.

Here, the potential of machine learning, particularly ensemble models like LightGBM and AdaBoost, becomes evident. Their ability to capture complex, non-linear interactions among behavioral and transactional variables, such as recent payment patterns, income stability, and housing status, provides a more accurate reflection of short-term creditworthiness than reliance on static historical variables alone. This capacity may enable more equitable assessments, reducing unjustified rejections while maintaining rigorous risk control.

Integrating such models into modern credit evaluation pipelines also supports more nuanced client segmentation strategies. Institutions could differentiate product offerings for high-risk new entrants and low-risk salaried workers with short credit histories, tailoring financing solutions to promote sustainable onboarding rather than outright denials. This can be particularly relevant in Portugal, where financial institutions are increasingly challenged to balance prudence with inclusivity in addressing the needs of underbanked populations.

5.2. Reinforce Study Relevance

This study directly addresses the growing demand for robust, data-driven approaches to credit risk assessment within the banking sector, a domain undergoing rapid transformation due to evolving financial regulations, digitalization, and increasingly diverse consumer profiles. By applying a structured and comparative methodology to evaluate both traditional and ensemble machine learning models, this work provides timely insights into how credit institutions can better align predictive analytics with operational requirements and regulatory expectations.

In a context where access to credit remains simultaneously a necessity and a challenge for many segments of the population, particularly low-income households and young workers, the ability to reliably assess creditworthiness through behavioral and transactional data becomes a decisive factor in promoting financial inclusion. In the Portuguese case, these challenges are especially pronounced among individuals with lower incomes, reduced educational attainment, older age groups, and those residing in less urbanized regions. For these populations, accurate and fair credit evaluation mechanisms are not only instrumental for enabling access to essential financial services but also for fostering broader socioeconomic participation.

5.3. Acknowledge Limitations

Despite the robust methodological framework and comparative analysis applied in this study, several limitations must be acknowledged. The most significant relates to the reliance on publicly available synthetic data from the LendingClub platform, a U.S.-based peer-to-peer lending company. While this dataset provides substantial granularity and is valuable for benchmarking and methodological exploration, it does not fully capture the specific characteristics of Portuguese credit applicants or the evaluation practices employed by commercial banks in consumer credit assessment. Important variables such as nationality, residency status, marital status, or household composition, factors often central to real-world credit decisions, were not available. This absence restricts the model's ability to reflect the full complexity of borrower profiles and apply it within the Portuguese banking system.

A second limitation concerns the absence of sensitive attribute analysis, including gender, age, or ethnicity, which are frequently central to fairness assessments in credit scoring. While this study prioritized model performance, training efficiency, and explainability in line with institutional objectives, ethical dimensions such as bias detection and mitigation remain unaddressed. This represents an important avenue for future research, particularly as regulatory frameworks increasingly emphasize fairness and transparency in algorithmic decision-making.

Finally, although experiments were conducted using both PyCaret and native scikit-learn frameworks, the models were not deployed in a production environment nor integrated into internal bank decision systems. As a result, definitive claims regarding scalability, user interaction, or operational risk cannot be made. These factors are critical for real-world implementation and must be carefully considered in future studies seeking to transition from experimental validation to practical adoption.

5.4. Recommendations for Future Works

Building upon the frameworks and findings of this study, future research can meaningfully advance the field by leveraging real-world credit datasets, ideally sourced directly from Portuguese financial institutions. Access to anonymized transactional data enriched with relevant sociodemographic indicators would enable a deeper understanding of model behavior across different borrower profiles, as well as fairness-aware modeling and segmentation across sensitive groups. Furthermore, datasets that track borrower behavior over time would allow for the evaluation of model stability, seasonality effects, and long-term risk evolution. This, in turn, could foster the integration of temporal modeling techniques, such as time-series forecasting, to assess default probabilities across multiple horizons rather than at a single static point in time. Collaborations with banks through sandbox or testbed environments would also provide the opportunity to simulate production conditions, enabling models to be evaluated under real-world operational constraints such as latency, throughput, and policy overrides.

In addition, the integration of alternative data sources presents a transformative opportunity for the credit scoring landscape, particularly in enhancing access to credit for underbanked populations. According to Orlova, (2021), digital footprint data, such as online behavior, mobile phone usage, geolocation patterns, or in-app activity, can serve as powerful predictors of creditworthiness in the absence of traditional credit history. In the Portuguese context, this could be especially relevant for immigrant populations, who may be newly arrived in the country and thus lack formal records in credit bureaus. For example, data reflecting how frequently a user explores credit-related content in a banking app or engages with financial education tools may serve as proxies of intent and financial discipline (Orlova, 2021). Incorporating such alternative data responsibly, and with appropriate ethical safeguards, would not only improve predictive accuracy but also contribute to financial inclusion initiatives.

Future research should therefore explore both the operational viability and fairness implications of combining alternative and traditional data streams in credit scoring. Such exploration is especially critical in contexts where innovation must coexist with regulatory scrutiny, ensuring that technological advances support not only predictive performance but also transparency, accountability, and inclusion.

References

Ala'raj, M., Abbod, M. F., Majdalawieh, M., & Jum'a, L. (2022). A deep learning model for behavioural credit scoring in banks. Neural Computing and Applications, 34(16), 13689–13705. https://doi.org/10.1007/s00521-021-06695-z

Alonso Robisco, A., & Carbó Martínez, J. M. (2022). Measuring the model risk-adjusted performance of machine learning algorithms in credit default prediction. Financial Innovation, 8, 54. https://doi.org/10.1186/s40854-022-00366-1

Aruleba, I., & Sun, Y. (2024). Effective credit risk prediction using ensemble classifiers with model explanation. IEEE Access, 12, 12345–12357. https://doi.org/10.1109/ACCESS.2024.3445308

Bulut, C., & Arslan, E. (2024). Comparison of the impact of dimensionality reduction and data splitting on classification performance in credit risk assessment. Artificial Intelligence Review, 57, 10904. https://doi.org/10.1007/s10462-024-10904-1

Banco de Portugal (2020). Portuguese digital financial inclusion varies significantly across age groups. Banco de Portugal. https://www.bportugal.pt/en/comunicado/portuguese-digital-financial-inclusion-varies-significantly-across-age-groups

Banco de Portugal. (n.d.-a). What it is and types of credit. Retrieved July 17, 2025, from https://clientebancario.bportugal.pt/en/what-it-and-types-credit

Banco de Portugal. (n.d.-b). Creditworthiness assessment. Retrieved July 17, 2025, from https://clientebancario.bportugal.pt/en/creditworthiness-assessment

Banco de Portugal. (n.d.-c). How to enter into a consumer credit agreement. Retrieved July 17, 2025, from https://clientebancario.bportugal.pt/en/how-enter-consumer-credit-agreement

Banco de Portugal. (n.d.-d). Interest rates in consumer credit. Retrieved July 17, 2025, from https://clientebancario.bportugal.pt/en/interest-rates-consumer-credit

Banco de Portugal. (n.d.-e). Evolution of new loans. Retrieved July 30, 2025, from https://clientebancario.bportugal.pt/en/evolution-new-loans

Banco de Portugal. (n.d.). Portuguese digital financial inclusion varies significantly across age groups. Banco de Portugal. Retrieved July 30, 2025, from https://www.bportugal.pt/en/comunicado/portuguese-digital-financial-inclusion-varies-significantly-across-age-groups

De Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for credit assessment in banks. Journal of Risk and Financial Management, 15(12), 556. https://doi.org/10.3390/jrfm15120556

Feng, X., Xiao, Z., Zhong, B., Dong, Y., & Qiu, J. (2019). Dynamic weighted ensemble classification for credit scoring using Markov chain. Applied Intelligence, 49(6), 2106–2121. https://doi.org/10.1007/s10489-018-1253-8

Filchenkov, A., Khanzhina, N., Tsai, A., & Smetannikov, I. (2021). Regularization of autoencoders for bank client profiling based on financial transactions. Risks, 9(3), 54. https://doi.org/10.3390/risks9030054

Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2014). Modelling credit risk with scarce default data: On the suitability of cooperative bootstrapped strategies for small low-default portfolios. Journal of the Operational Research Society, 65(6), 806–821. https://doi.org/10.1057/jors.2013.119

Han, S., Jung, H., Yoo, P. D., Cali, A., & Provetti, A. (2024). NOTE: Non-parametric oversampling technique for explainable credit scoring. Scientific Reports, 14, 78055. https://doi.org/10.1038/s41598-024-78055-5

Kozodoi, N., Jacob, J., & Lessmann, S. (2022). Fairness in credit scoring: Assessment, implementation and profit implications. European Journal of Operational Research, 297(3), 1083–1098. https://doi.org/10.1016/j.ejor.2021.06.023

Li, C., Wang, H., Jiang, S., & Gu, B. (2024). The effect of AI-enabled credit scoring on financial inclusion: Evidence from an underserved population of over one million. MIS Quarterly, 48(1), 18340. https://doi.org/10.25300/MISQ/2024/18340

Li, Y., & Wen, G. (2023). Research and practice of financial credit risk management based on federated learning. Engineering Letters, 31(1).

Liang, D., Tsai, C. F., Dai, A. J., & Eberle, W. (2018). A novel classifier ensemble approach for financial distress prediction. Knowledge and Information Systems, 54, 211–231. https://doi.org/10.1007/s10115-017-1061-1

Merćep, A., Mrčela, L., Birov, M., & Kostanjčar, Z. (2021). Deep neural networks for behavioral credit rating. Entropy, 23(1), 27. https://doi.org/10.3390/e23010027

Mhlanga, D. (2021). Financial inclusion in emerging economies: The application of machine learning and artificial intelligence in credit risk assessment. International Journal of Financial Studies, 9(3), 39. https://doi.org/10.3390/ijfs9030039

Moldovan, D. (2023). Algorithmic decision-making methods for fair credit scoring. IEEE Access, 11, 60524–60536. https://doi.org/10.1109/ACCESS.2023.3286018

Moral-Garcia, S., & Abellan, J. (2023). Improving the results in credit scoring by increasing diversity in ensembles of classifiers. IEEE Access, 11, 59625–59636. https://doi.org/10.1109/ACCESS.2023.3284137

Muñoz-Cancino, R., Bravo, C., Ríos, S. A., & Graña, M. (2022). Assessment of creditworthiness models privacy-preserving training with synthetic data. In Hybrid

Artificial Intelligent Systems (HAIS 2022) (pp. 347–358). Springer. https://doi.org/10.1007/978-3-031-15471-3_32

Orlova, E. V. (2021). Methodology and models for individuals' creditworthiness management using digital footprint data and machine learning methods. Mathematics, 9(15), 1820. https://doi.org/10.3390/math9151820

Orlova, E. V. (2021). Data-driven design to credit risk management using digital footprint intelligence. Proceedings of the 2021 3rd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA 2021). IEEE. https://doi.org/10.1109/SUMMA53307.2021.9632188

Oualid, A., Maleh, Y., & Moumoun, L. (2023). Federated learning techniques applied to credit risk management: A systematic literature review. EDPACS, 57(2), 14–23. https://doi.org/10.1080/07366981.2023.2241647

Ponsam, J. G., Bella Gracia, S. V. J., Geetha, G., Karpaselvi, S., & Nimala, K. (2021). Credit risk analysis using LightGBM and a comparative study of popular algorithms. Proceedings of the 2021 4th International Conference on Computing and Communications Technologies (ICCCT 2021). IEEE. https://doi.org/10.1109/ICCCT53315.2021.9711896

Rajesh, D. M. V., Lakshmanarao, A., & Gupta, C. (2023). An efficient machine learning classification model for credit approval. Proceedings of the 3rd International Conference on Artificial Intelligence and Smart Energy (ICAIS 2023). IEEE. https://doi.org/10.1109/ICAIS56108.2023.10073706

Robinson, N., & Sindhwani, N. (2024). Loan default prediction using machine learning. Proceedings of the 11th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO 2024). IEEE. https://doi.org/10.1109/ICRITO61523.2024.10522232

Roy, P. K., & Shaw, K. (2023). An integrated fuzzy credit rating model using fuzzy-BWM and new fuzzy-TOPSIS-Sort-C. Complex and Intelligent Systems, 9, 115–132. https://doi.org/10.1007/s40747-022-00823-5

Sadok, H., Sakka, F., & El Maknouzi, M. E. H. (2022). Artificial intelligence and bank credit analysis: A review. Cogent Economics & Finance, 10(1), 2023262. https://doi.org/10.1080/23322039.2021.2023262

Simumba, N., Okami, S., Kodaka, A., & Kohtake, N. (2021). Spatiotemporal integration of mobile, satellite, and public geospatial data for enhanced credit scoring. Symmetry, 13(4), 575. https://doi.org/10.3390/sym13040575

Soomro, A., Zakariyah, H., Aftab, S. M. A., Muflehi, M., Shah, A., & Meraj, S. (2024). Loan default prediction using machine learning algorithms: A systematic literature review

2020–2023. Pakistan Journal of Life and Social Sciences, 22(2), 469. https://doi.org/10.57239/PJLSS-2024-22.2.00469

Tripathi, D., Edla, D. R., Bablani, A., Shukla, A. K., & Reddy, B. R. (2021). Experimental analysis of machine learning methods for credit score classification. Progress in Artificial Intelligence, 10(3), 423–437. https://doi.org/10.1007/s13748-021-00238-2

United Nations, Department of Economic and Social Affairs. (n.d.). Poverty eradication. Retrieved September 17, 2025, from https://social.desa.un.org/issues/poverty-eradication

World Bank. (2025, January 27). Financial Inclusion. World Bank. https://www.worldbank.org/en/topic/financialinclusion/overview

Xu, Z., Cheng, X., Wang, K., & Yang, S. (2020). Analysis of the environmental trend of network finance and its influence on traditional commercial banks. Journal of Computational and Applied Mathematics, 376, 112907. https://doi.org/10.1016/j.cam.2020.112907

Yao, J., Wang, Z., Wang, L., Liu, M., Jiang, H., & Chen, Y. (2022). Novel hybrid ensemble credit scoring model with stacking-based noise detection and weight assignment. Expert Systems with Applications, 198, 116913. https://doi.org/10.1016/j.eswa.2022.116913

Yemmanuru, P. K., Yeboah, J., & Nti, I. K. (2024). Customer credit risk: Application and evaluation of machine learning and deep learning models. Proceedings of the IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI 2024). IEEE. https://doi.org/10.1109/ICMI60790.2024.1058589

APPENDIX A

Tables

Table A.1. Full listing of all variables derived from the original dataset

Variable	Definition	Data Type
acc_now_delinq	The number of accounts on which the borrower is now delinquent.	float64
addr_state	The state provided by the borrower in the loan application.	object
all_util	Balance to credit limit on all trades.	float64
annual_inc	The self-reported annual income provided by the borrower during registration.	float64
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration.	float64
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers.	object
collection_recovery_fee	post charge off collection fee.	float64

Variable	Definition	Data Type
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections.	float64
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years.	float64
desc	Loan description provided by the borrower.	object
dti	A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested loan, divided by the borrower’s self-reported monthly income.	float64
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested loan, divided by the co-borrowers' combined self-reported monthly income.	float64
earliest_cr_line	The month the borrower's earliest reported credit line was opened	object
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.	object
emp_title	The job title supplied by the Borrower when applying for the loan.	object
funded_amnt	The total amount committed to that loan at that point in time.	int64
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.	float64
grade	LendingClub assigned loan grade.	object

Variable	Definition	Data Type
home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. The values are: RENT, OWN, MORTGAGE, OTHER.	object
id	A unique assigned ID for the loan listing.	int64
il_util	Ratio of total current balance to high credit/credit limit on all install acct.	float64
initial_list_status	The initial listing status of the loan. Possible values are – W (“Whole”), F (“Fractional”).	object
inq_fi	Number of personal finance inquiries.	float64
inq_last_12m	Number of credit inquiries in the past 12 months.	float64
inq_last_6mths	The number of inquiries in the past 6 months (excluding auto and mortgage inquiries).	float64
installment	The monthly payment owed by the borrower if the loan originates.	float64
int_rate	Interest Rate on the loan.	float64
issue_d	The month in which the loan was funded.	object
last_credit_pull_d	The most recent month pulled credit for this loan.	object

Variable	Definition	Data Type
last_pymnt_amnt	Last total payment amount received.	float64
last_pymnt_d	Last month payment was received.	object
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.	int64
loan_status	Current status of the loan.	object
max_bal_bc	Maximum current balance owed on all revolving accounts.	float64
member_id	A unique assigned Id for the borrower member.	int64
mths_since_last_delinq	The number of months since the borrower's last delinquency.	float64
mths_since_last_major_derog	Months since the most recent 90-day or worse rating at time of application for the secondary applicant.	float64
mths_since_last_record	The number of months since the last public record.	float64
mths_since_rcnt_il	Months since most recent installment accounts opened.	float64
next_pymnt_d	Next scheduled payment date.	object

Variable	Definition	Data Type
open_acc	The number of open credit lines in the borrower's credit file.	float64
open_acc_6m	Number of open trades in the last 6 months.	float64
open_il_12m	Number of installment accounts opened in the past 12 months.	float64
open_il_24m	Number of installment accounts opened in the past 24 months.	float64
open_il_6m	Number of installment accounts opened in the past 6 months.	float64
open_rv_12m	Number of revolving trades opened in the past 12 months.	float64
open_rv_24m	Number of revolving trades opened in the past 24 months.	float64
out_prncp	Remaining outstanding principal for total amount funded.	float64
out_prncp_inv	Remaining outstanding principal for a portion of total amount funded by investors.	float64
policy_code	publicly available policy_code=1, new products not publicly available policy_code=2.	int64
pub_rec	Number of derogatory public records.	float64

Variable	Definition	Data Type
purpose	A category provided by the borrower for the loan request.	object
pymnt_plan	Indicates if a payment plan has been put in place for the loan.	object
recoveries	post charge off gross recovery.	float64
revol_bal	Total credit revolving balance.	int64
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.	float64
sub_grade	LendingClub assigned a loan subgrade.	object
term	The number of payments on the loan. Values are in months and can be either 36 or 60.	object
title	The loan title provided by the borrower.	object
tot_coll_amt	Total collection amounts ever owed.	float64
tot_cur_bal	Total current balance of all accounts.	float64
total_acc	The total number of credit lines currently in the borrower's credit file.	float64

Variable	Definition	Data Type
total_bal_il	Total current balance of all installment accounts.	float64
total_cu_tl	Number of finance trades.	float64
total_pymnt	Payments received to date for the total amount funded.	float64
total_pymnt_inv	Payments received to date for a portion of the total amount funded by investors.	float64
total_rec_int	Interest received to date.	float64
total_rec_late_fee	Late fees received to date.	float64
total_rec_prncp	Principal received to date.	float64
total_rev_hi_lim	It is the sum of the maximum credit limits of all revolving credit lines (such as credit cards) held by the customer.	float64
Unnamed: 0		int64
url	URL for the LendingClub page with listing data.	object
verification_status	Indicates if income was verified by LendingClub, not verified, or if the income source was verified.	object

Variable	Definition	Data Type
verification_status_joint	Indicates if the co-borrowers' joint income was verified by LendingClub, not verified, or if the income source was verified.	float64
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.	object

APPENDIX B

Tables

Table B.1. Feature Selection Results Using ANOVA-F

Variable	ANOVA-F
out_prncp	136780.236898
initial_list_status_w	2855.841445
int_rate	1771.431500
inq_last_6mths_category_3-5	1048.629246
total_pymnt	781.072185
inq_last_6mths_category_1-2	620.134637
home_ownership_RENT	427.161080
tot_coll_amt_category_Sem	418.767194
sub_grade_A5	364.571909
sub_grade_B1	363.856368
inq_last_6mths_category_6-10	310.452066
income_category_50K-75K	292.161961

Variable	ANOVA-F
sub_grade_A4	275.889703
sub_grade_A2	262.585541
sub_grade_A3	235.275873
sub_grade_B2	230.629505
sub_grade_E4	213.654631
tot_coll_amt_category_Média	203.191316
emp_length_10+	201.727156
sub_grade_F5	189.081190
sub_grade_F2	185.394026
sub_grade_F3	177.217220
sub_grade_E2	174.353446
sub_grade_E5	156.948345
sub_grade_F1	156.689153

Table B.2. Feature Selection Results Using Mutual Information (MI)

Variable	Mutual Information (MI)
out_prncp	0.371590
int_rate	0.051970
tot_cur_bal	0.045496
total_rev_hi_lim	0.044520
total_pymnt	0.039545
total_rec_int	0.036488
earliest_cr_line_category_>	0.031327
tot_coll_amt_category_Sem	0.023686
income_category_50K-75K	0.023534
initial_list_status_w	0.022448
home_ownership_RENT	0.016043
inq_last_6mths_category_1-2	0.012468
total_acc_category_11-20	0.010171
emp_length_10+	0.007716
total_acc_category_21-30	0.006756

Variable	**Mutual Information (MI)**
out_prncp	0.371590
int_rate	0.051970
tot_cur_bal	0.045496
total_rev_hi_lim	0.044520
total_pymnt	0.039545
total_rec_int	0.036488
earliest_cr_line_category_>	0.031327
loan_amnt	0.004856
total_acc_category_31-50	0.004223
inq_last_6mths_category_3-5	0.003386
sub_grade_F3	0.002571
sub_grade_G3	0.002265
revol_util_category_>	0.002226
revol_util_category_31-40%	0.001994
dti	0.001809
sub_grade_B3	0.001775

Variable	Mutual Information (MI)
out_prncp	0.371590
int_rate	0.051970
tot_cur_bal	0.045496
total_rev_hi_lim	0.044520
total_pymnt	0.039545
total_rec_int	0.036488
earliest_cr_line_category_>	0.031327
sub_grade_C1	0.001664

Table B.3. Feature Selection Results Using SelectFromModel (Logistic Regression)

Variable	**Selected_SFM (Logistic Regression)**
int_rate	TRUE
dti	TRUE
out_prncp	TRUE
total_pymnt	TRUE
home_ownership_RENT	TRUE
inq_last_6mths_category_1-2	TRUE
inq_last_6mths_category_3-5	TRUE

Table B.4. Cross-tabulation of feature selection results across Statistical Tests (ANOVA-F), Mutual Information, and Logistic Regression

Variable	ANOVA-F	MI	LR
inq_last_6mths_category	✔	✔	✔
sub_grade	✔	✔	✘
home_ownership	✔	✔	✔
int_rate	✔	✔	✔
out_prncp	✔	✔	✔
total_pymnt	✔	✔	✔
income_category	✔	✔	✘
initial_list_status	✔	✔	✘
tot_coll_amt_category	✔	✔	✘
total_acc_category	✘	✔	✘
dti	✘	✔	✔
earliest_cr_line_category	✘	✔	✘

Variable	ANOVA-F	MI	LR
emp_length	✔	✔	✘
tot_cur_bal	✘	✔	✘
total_rec_int	✘	✔	✘
total_rev_hi_lim	✘	✔	✘