International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

AI-Driven Credit Risk Models for Small-Scale Lending: A Business Analytics Framework for Predictive Performance and Responsible Deployment

Deborah O. Oyeyemi [1,*], Obianuju O. Okosieme [2], Tosin Idowu-Kunlere [1], Obiamaka Okosieme [2], Abdoulkarim H. Moussa [3] and Edekin A. Julius [4]

[1] *MSc Business Analytics and Information Management, Alfred Lerner College of Business and Economics, University of Delaware.*
[2] *MSc Marketing Analysis, Kellstadt Graduate School of Business, DePaul University.*
[3] *Master of Business Administration, Fox School of Business, Temple University.*
[4] *Research Student, Department of Economics, Faculty of Social Science, Lagos State University, Ojo, Lagos.*

International Journal of Science and Research Archive, 2025, 17(01), 168-184

Publication history: Received on 25 August 2025; revised on 30 September 2025; accepted on 03 October 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.17.1.2751

Abstract

Small and medium-sized enterprises (SMEs) play a vital role in driving economic growth, yet many still encounter barriers when seeking credit. Traditional models such as logistic regression have long been the backbone of credit scoring because they are simple and interpretable. However, their ability to capture the complex and sometimes nonlinear nature of borrower behavior is limited, which often leads to misclassification of risk. This study applies a quantitative, explanatory approach to compare the performance of four models—Logistic Regression, Random Forest, XGBoost, and a Multi-Layer Perceptron (MLP) neural network—in predicting SME credit risk. Model accuracy, precision, recall, F1-score, and the ROC-AUC metric were used for evaluation, while SHAP and LIME were integrated to improve interpretability and transparency. The findings reveal clear differences in performance. Logistic regression produced an accuracy of 79% with a ROC-AUC of 0.58 but identified only 22% of actual defaulters, highlighting its weakness in imbalanced datasets. Random Forest increased recall to 68%, demonstrating better sensitivity to defaults, but its overall accuracy dropped to 65%, reflecting trade-offs between identifying defaulters and misclassifying non-defaulters. XGBoost achieved 86% accuracy and a ROC-AUC of 0.74, but its recall for defaulters was extremely low at just 2.4%, showing a bias toward majority (non-default) cases. The strongest results came from the MLP neural network, which reached 95% accuracy, balanced precision and recall of 0.95, and a ROC-AUC of 0.98, confirming its ability to capture hidden patterns in borrower data.This study adds to credit risk research by illustrating how advanced machine learning models, when combined with interpretability tools, can provide both accuracy and accountability. For SMEs, the practical implication is fairer and more reliable access to credit, while for lenders and regulators, these models offer a pathway to strengthen credit assessment without sacrificing transparency.

Keywords: Small and Medium Enterprises (SMEs); Credit Risk Modeling; Logistic Regression; Machine Learning; Interpretability (SHAP and LIME); Credit Scoring; Financial Inclusion

1. Introduction

Access to finance has remained one of the central challenges for small-scale borrowers across developing and developed economies alike. Micro-entrepreneurs, farmers, and informal workers often find themselves excluded from the financial system because they lack collateral, credit history, or formal records. In recent years, artificial intelligence (AI) has emerged as a potential solution to this problem. Machine learning models, drawing from both transactional and alternative data sources, promise to identify hidden patterns in borrower behavior that traditional credit scoring

* Corresponding author: Deborah O. Oyeyemi

methods often miss. By doing so, they can expand financial inclusion, reduce default risks, and enable institutions to lend more efficiently.

Artificial intelligence (AI) has gained traction in credit risk modeling, particularly for small and micro enterprises where traditional data sources are often scarce. For instance, a recent study of Azerbaijani SMEs found that applying a Random Forest model significantly improved the performance of a credit scoring system formerly based on a Delphi (logistic regression) approach: accuracy increased from 0.69 to 0.83, precision from 0.65 to 0.81, recall from 0.56 to 0.77, and the F1-score from 0.58 to 0.79 (Karimova, 2024). These gains demonstrate how machine learning can help lenders better distinguish between likely defaulters and non-defaulters when richer feature sets are available.

Yet predictive accuracy alone is not enough. Conventional methods like logistic regression are still favored in many institutions because of their transparency, even if they underperform. Recent evidence from Oyeyemi, Okosieme, Idowu-Kunlere, Julius, and Nwinyi (2025) illustrates this trade-off: while logistic regression achieved the highest overall accuracy at 77.3%, ensemble machine learning models such as Random Forest were more effective at distinguishing high-risk borrowers. Importantly, when explainable AI (XAI) tools such as SHAP and LIME were applied, the most influential factors — loan amount, borrower age, and loan duration — were revealed, making predictions more interpretable and actionable. Together, these studies highlight that accuracy and interpretability must be pursued in tandem if AI is to serve responsibly in credit risk management.

The lack of transparency and the potential for bias in ML models remain pressing concerns. For example, Nwafor et al. (2024) demonstrate that removing sensitive attributes such as age and gender does not significantly reduce classification performance in their hybrid model, while helping mitigate bias. Similarly, SHAP-based studies have proven effective in identifying which features drive predictions, strengthening trust in decision-making (Nwafor et al., 2024; Gafsi, 2025). On a policy level, FinRegLab (2023) finds that explainability tools are increasingly used by lenders for regulatory compliance, though no single method works universally. Beyond compliance, Shahsavarifar et al. (2022) propose a two-step approach that first learns a fair similarity metric and then builds a classifier excluding sensitive features, thereby reducing unfairness at both group and individual levels.

These findings underline a central argument: for small-scale lending, where each decision can have life-changing consequences, the responsible deployment of AI is not optional but essential. This paper therefore introduces a business analytics framework that integrates predictive performance with interpretability and fairness. It seeks to demonstrate that lenders do not need to choose between accuracy and responsibility — both can be pursued together to build credit systems that are not only efficient but also trustworthy.

2. Literature Review

Credit risk assessment has long been grounded in statistical and econometric models whose interpretability made them attractive to lenders and supervisors. Logistic regression and linear discriminant analysis dominated early credit scoring because they provide transparent coefficient estimates that are easy to audit and explain to stakeholders (Altman & Saunders, 1998). However, these classical approaches frequently struggle when borrower behaviour exhibits nonlinearities, complex interactions, or when predictor sets include large numbers of weak signals — conditions often present in SME portfolios with noisy or sparse records. Reviews by supervisory bodies and scholars point out that while statistical models retain advantages for governance and disclosure, their predictive ceiling can be limiting in real-world default prediction tasks (Alonso et al., 2020).

Over the last decade, tree-based methods and ensemble learners have become the workhorses of modern credit modeling. Random Forests and Gradient Boosting Machines (e.g., XGBoost, LightGBM) routinely outperform simple classifiers on predictive metrics because they capture nonlinear relationships and interactions without extensive manual feature engineering. Empirical comparisons across consumer and small business contexts show marked gains in accuracy, precision and recall when ensembles replace single, linear models (Chang et al., 2024; Karimova, 2024). For SME-specific work, Karimova (2024) documents how a Random Forest model improved multiple performance metrics over a Delphi/logistic approach in an Azerbaijani SME dataset — a practical illustration that ensemble models can detect subtle patterns in firm financials and behaviour that linear methods miss.

Boosting and deep learning methods have further pushed predictive frontiers, especially where high-dimensional or unstructured data (text, transaction logs) are available. Studies employing XGBoost and neural networks report superior area-under-curve (AUC) and F1-scores in default prediction exercises, though gains may diminish on smaller datasets due to overfitting risks (Chang et al., 2024; Paz, 2025). Importantly, these gains come at the cost of

interpretability: complex models can be opaque, which raises issues for regulators and lenders that must provide adverse-action explanations and meet fair-lending obligations.

The interpretability problem has motivated a fast-growing literature on explainable AI (XAI) tools that make complex models auditable and intelligible. SHAP (SHapley Additive exPlanations) provides a theoretically grounded approach to assign local and global feature importances, while LIME offers local surrogate explanations; both have become standard diagnostics in credit model toolkits (Lundberg & Lee, 2017). Practitioners and academic studies show that applying SHAP or LIME to ensemble models frequently reveals intuitive drivers (e.g., repayment history, loan size, cash-flow proxies), thereby narrowing the gap between performance and transparency (Lundberg & Lee, 2017; Gafsi, 2025). Policy-facing work, notably by FinRegLab, has tested these tools in production settings and concluded that while post-hoc explainers add valuable insight, no single tool is universally sufficient — model governance still requires complementary documentation, monitoring and human review (FinRegLab, 2023).

A closely related stream of literature examines fairness, bias, and individual-level harms that can arise when ML models learn spurious correlations in training data. Researchers have developed fairness-aware training procedures, pre/post-processing corrections, and individual fairness metrics to mitigate disparate impacts (Shahsavarifar et al., 2022). For example, Shahsavarifar and colleagues propose a two-step training approach that learns a fair similarity metric and then builds an individually fair classifier; empirical tests in credit settings indicate such techniques can reduce both group- and individual-level unfairness while maintaining reasonable accuracy (Shahsavarifar et al., 2022). Complementing algorithmic remedies, field studies demonstrate that simple operational choices — excluding sensitive attributes, careful feature selection, and periodic fairness audits — can materially reduce biased outcomes without catastrophic loss of predictive power (Nwafor et al., 2024).

SME credit risk poses particular empirical challenges that shape methodological choices. SMEs often lack rich financial time series, formal credit histories, or standardized reporting; as a result, researchers augment models with alternative data sources (bank transaction patterns, mobile money behaviour, supply-chain signals). Studies that combine alternative features with ensemble learners and XAI reveal both improved discrimination and better insight into drivers of default, enabling lenders to tailor credit offers and risk mitigation (Karimova, 2024; Oyeyemi et al., 2025). Nonetheless, access to comprehensive SME loan-level repayment data remains scarce, so much of the literature relies on single-institution case studies, public surveys, or nationally aggregated datasets, an empirical limitation that researchers must acknowledge when claiming generalisability.

Finally, regulatory and policy analyses emphasize that model performance is necessary but not sufficient for sustainable deployment. FinRegLab's empirical work and follow-up policy analysis stress that explainability, traceability, and monitoring are essential for meeting disclosure and fair-lending obligations; they further recommend that lenders combine technical diagnostics with governance practices such as documentation, thresholded human review for adverse actions, and regular fairness testing (FinRegLab, 2023). Taken together, the literature paints a consistent picture: advanced ML methods materially improve predictive accuracy for credit risk, including SMEs, but these gains must be paired with interpretability techniques and institutional controls to manage bias, compliance, and borrower trust.

Table 1 Selected Empirical Studies on Credit Risk Models

Author(s)	Dataset	Models Compared	Key Findings	Limitations
Karimova (2024)	Azerbaijani SMEs	Random Forest vs Delphi/logistic regression	RF improved accuracy (0.69 → 0.83), precision, recall, and F1-score.	Country-specific; limited dataset size.
Oyeyemi et al. (2025)	Open-source credit dataset + XAI	Logistic regression, Random Forest, ensemble models with SHAP & LIME	Logistic regression had highest baseline accuracy (77.3%), but RF better at distinguishing high-risk borrowers.	Open-source data; generalisability limited.
Nwafor et al. (2024)	Hybrid model on credit datasets	Hybrid ML model with fairness constraints	Excluding sensitive attributes mitigated bias with minimal loss in accuracy.	Dataset details not SME-specific.

Chang et al. (2024)	Consumer loan datasets	Deep learning vs ML vs logistic regression	Deep learning slightly outperformed boosting models in AUC/F1.	Overfitting risks, requires big data.
Shahsavarifar et al. (2022)	Simulated credit datasets	Fairness-aware training vs standard classifiers	Fair similarity metrics reduced unfairness at individual/group levels.	Simulated rather than SME-specific data.
FinRegLab (2023)	U.S. financial institutions' underwriting	Post-hoc explainers (SHAP, LIME)	Explainability improved compliance readiness; no one-size-fits-all tool.	U.S.-centric; institutional scope.

Despite a growing body of work on AI-driven credit risk modeling, key limitations persist in existing studies. For example, Karimova (2024) shows that Random Forest can significantly outperform a Delphi (logistic regression) model in SME lending — accuracy improved from 0.69 to 0.83 — but her data are drawn from Azerbaijani SMEs, limiting generalizability to other markets. Likewise, recent work on explainable credit risk models—such as the ensemble + SHAP/LIME system proposed in Pathak (2025) — demonstrates promising transparency gains, yet it relies on open datasets and does not specifically tailor its methods to small-scale lending or microfinance contexts. Other research, like that in Nallakaruppan et al. (2024), advances explainable AI models for peer-to-peer lending but focuses on transparency in institutional environments rather than the unique challenges faced by small-scale lenders in emerging economies.

This study fills these gaps by leveraging the Lending Club loan dataset, which offers large-scale, real-world consumer loan information, enabling robust benchmarking of AI models in a data-rich context. Because this dataset contains detailed borrower, repayment, and loan performance attributes, it allows testing of fairness-aware and interpretable models in practice, rather than relying solely on simulation or small samples. Moreover, by adapting and evaluating explainability (e.g. SHAP, LIME) and fairness frameworks in the specific domain of small-scale lending, the study advances methods that are better aligned with the constraints and regulatory needs of micro-lenders and emerging markets. In doing so, it bridges the gap between high-performing AI systems and practical, transparent deployment in resource-constrained lending environments.

3. Methodology

This study adopts a quantitative, explanatory research design aimed at developing and comparing credit risk models for small and medium-sized enterprise (SME) lending. Logistic regression is used as the baseline model because of its long-standing role in credit scoring and its interpretability. However, to address its limitations in capturing nonlinear borrower behavior, advanced machine learning models; Random Forest, Gradient Boosting (XGBoost), and Neural Networks are also employed. Since responsible deployment is as critical as predictive accuracy, the models are complemented with explainability tools, including SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), ensuring that the predictions are transparent and interpretable for lenders, regulators, and borrowers (Oyeyemi et al., 2025; FinRegLab, 2023).

3.1. Model Specification

This study employs a quantitative, explanatory approach to develop and compare credit risk models for small and medium-sized enterprises (SMEs). The baseline model is a logistic regression, which estimates the probability of default ($Y_i=1$) based on a set of predictor variables Xi=[xi1,xi2,…,xip], including firm characteristics, financial ratios, and credit history. The logistic model is formally specified as:

$$\text{logit}(P(Y_i = 1 \mid X_i)) = \ln\left(\frac{P(Y_i = 1 \mid X_i)}{1 - P(Y_i = 1 \mid X_i)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$

where β0 is the intercept, βj are coefficients associated with the predictors, and ϵi is the error term. While logistic regression provides interpretability and a well-established baseline, it may be limited in capturing nonlinear and complex interactions among predictors. To address these limitations, the study also employs advanced machine

learning models such as **Random Forest** aggregates predictions from T decision trees, improving robustness and capturing nonlinearities such as;

$$\hat{P}(Y_i = 1 \mid X_i) = \frac{1}{T}\sum_{t=1}^{T} h_t(X_i)$$

where h_t (Xi) denotes the prediction from the t^{th} tree. Gradient Boosting (XGBoost) further enhances predictive accuracy by sequentially adding trees to minimize the binary cross-entropy loss:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta\, h_t(X_i), \quad L = -\sum_{i=1}^{n} [y_i \ln \hat{y}_i + (1 - y_i)\ln(1 - \hat{y}_i)]$$

where η is the learning rate, ht(Xi) is the tree added at iteration t, and L is the loss function. Finally, a feedforward neural network with one hidden layer is specified to model highly nonlinear borrower behavior:

$$a_j = \sigma\left(\sum_{k=1}^{p} w_{kj}^{(1)} x_{ik} + b_j^{(1)}\right), \quad \hat{y}_i = \sigma\left(\sum_{j=1}^{h} w_j^{(2)} a_j + b^{(2)}\right)$$

where w and b denote weights and biases, h is the number of hidden neurons, and σ(·) is the activation function and since responsible deployment is as critical as predictive accuracy, model interpretability is integrated using SHAP and LIME. SHAP values quantify each feature's contribution to a prediction:

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!}\Big[f(S \cup \{j\}) - f(S)\Big]$$

while LIME approximates the complex model locally with an interpretable surrogate g:

$$g = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

These specifications ensure that the study captures both the predictive complexity and interpretability needed for SME credit risk assessment, providing transparent insights for lenders, regulators, and borrowers.

3.2. Data Source and Description

The dataset used in this study was drawn from the Lending Club loan dataset, publicly available on Kaggle and widely applied in credit risk modeling research. It consists of over 2.2 million loan records issued between 2007 and 2018, providing a rich repository of borrower demographics, loan characteristics, and repayment outcomes. Crucially, the dataset includes the loan status variable, which allows loans to be classified as either fully paid (non-defaulters) or default/charged off (defaulters).

This classification is central to the present study, as it directly supports the development and testing of machine learning models aimed at predicting repayment risk among small-scale borrowers. The large size and diversity of the dataset strengthen the robustness of model training and validation, ensuring that findings are not limited to narrow borrower profiles or loan types. Furthermore, the dataset's public accessibility enhances transparency and replicability, enabling

researchers and practitioners to benchmark results across different modeling frameworks. By leveraging this dataset, the study situates its analysis within a credible empirical context while addressing the broader challenge of designing AI-driven models for responsible and scalable small-scale lending. A subset of relevant variables from the Lending Club dataset is used for analysis. These variables capture borrower capacity, loan characteristics, and repayment outcomes. Table 1 provides a description of the selected variables.

3.3. Data Preprocessing

Preprocessing steps are essential to prepare the dataset for robust modeling. Loans with incomplete or ambiguous statuses are excluded to maintain a clear binary classification. Missing values in borrower information, such as employment length or income, are imputed using median or mode replacement. Categorical variables like loan purpose and home ownership are transformed into numerical indicators via one-hot encoding. Given the imbalance between repaid loans and defaults, the Synthetic Minority Oversampling Technique (SMOTE) is applied to balance the classes. Finally, continuous variables such as income and loan amount are standardized to ensure comparability across models, particularly for neural networks that are sensitive to input scales.

3.4. Model Development and Evaluation

3.4.1. Baseline Model: Logistic Regression

The first stage of model development involves estimating a logistic regression, serving as the baseline model. Its primary advantage lies in interpretability, providing clear insights into the relationship between borrower and loan characteristics and default probability. Logistic regression establishes a benchmark for predictive performance, enabling meaningful comparisons with more sophisticated models.

3.4.2. Ensemble Models: Random Forest and XGBoost

In the second stage, ensemble learning techniques are applied to capture complex, nonlinear interactions among predictors. Random Forest aggregates predictions from multiple decision trees, reducing variance and improving stability. XGBoost leverages gradient boosting to sequentially refine predictions, minimizing errors and enhancing performance in the presence of subtle patterns. These models are particularly effective at uncovering interactions that may be overlooked by traditional linear methods.

3.4.3. Deep Learning: Neural Network

The third stage introduces a feedforward Neural Network, trained to explore deep learning's capacity to detect hidden patterns in high-dimensional data. By modeling nonlinear relationships across multiple layers of computation, the neural network complements the ensemble methods and tests whether advanced architectures provide additional predictive power.

3.4.4. Explainable AI Integration

To ensure responsible and interpretable AI use, explainability tools are incorporated. SHAP (SHapley Additive Explanations) is used to provide global feature importance, ranking predictors according to their influence on default probability. LIME (Local Interpretable Model-Agnostic Explanations) offers borrower-level interpretability, showing how individual predictions are derived (Nwafor et al., 2024; Karimova, 2024). These tools enhance transparency, making the models actionable for lenders, regulators, and borrowers.

3.4.5. Performance Evaluation

Model performance is assessed using a comprehensive set of metrics. Accuracy provides a general measure of correctness, but due to class imbalance in repayment data, precision and recall are critical. Precision measures the model's ability to correctly identify non-defaulters, while recall evaluates its success in detecting true defaulters. The F1-score balances these two metrics, and the AUC-ROC quantifies discriminatory power across all thresholds. This multi-metric evaluation ensures robust, nuanced conclusions about each model's effectiveness and reliability.

3.4.6. Risk Assessment

Predicted probabilities from each model are translated into risk categories to support decision-making. Borrowers are classified into low, medium, or high-risk groups based on thresholds determined from historical default rates. This approach enables lenders to tailor credit terms, set interest rates, and determine collateral requirements according to

the level of risk. By combining predictive modeling with explainability, the framework not only estimates default likelihood but also provides actionable insights for effective SME credit management.

4. Empirical Finding

This section presents the results of the empirical analysis, highlighting the predictive performance of the developed credit risk models and the insights they offer into SME default behavior. The discussion integrates quantitative findings with interpretation, providing a narrative that explains not only what the models reveal but also why these patterns matter for risk management and policy.

4.1. Traditional Statistical Model: Logistic Results

Table 2 Logistic Regression Classification Results

Metric / Class	Precision	Recall	F1-Score	Support
Class 0 (Non-Defaulters)	0.8800	0.8800	0.8800	271,478
Class 1 (Defaulters)	0.2200	0.2200	0.2200	42,617
Accuracy			0.7900	314,095
Macro Average	0.5500	0.5500	0.5500	314,095
Weighted Average	0.7900	0.7900	0.7900	314,095
ROC-AUC			0.5790	

Source: Jupyter Notebook, 2025.

The baseline logistic regression model achieved an overall accuracy of 79 percent, suggesting that it correctly classified the majority of loan outcomes. However, this performance was heavily influenced by the dominance of non-defaulting borrowers in the dataset. As shown in the confusion matrix, the model correctly identified 239,558 non-defaulters but misclassified 33,439 defaulters as creditworthy. This imbalance is further reflected in the classification report: while non-defaulters (class 0) recorded high scores across precision, recall, and F1 (all around 0.88), the model struggled considerably with defaulters (class 1), achieving only 0.22 across all three metrics.

This indicates that while the model is strong at recognizing borrowers who will repay, it performs poorly in detecting those likely to default. For lenders, this shortfall is critical since failing to flag risky borrowers directly increases exposure to credit losses. The macro-average F1 score of 0.55 highlights the disparity between the two classes, while the ROC-AUC score of 0.58 suggests the model has only marginal discriminatory power above random guessing

4.2. Machine learning (ML) model

Table 3 Random Forest Classification Results

Metric / Class	Precision	Recall	F1-Score	Support
Class 0 (Non-Defaulters)	0.9273	0.6445	0.7605	271,478
Class 1 (Defaulters)	0.2305	0.6781	0.3440	42,617
Accuracy			0.6491	314,095
Macro Average	0.5789	0.6613	0.5523	314,095
Weighted Average	0.8328	0.6491	0.7040	314,095
ROC-AUC			0.7218	

Source: Jupyter Notebook, 2025.

The Random Forest model in Table 4.2 shows a different balance of strengths and weaknesses compared to the baseline. Overall accuracy stands at about 65 percent, which is lower than the logistic regression score, but the real story lies in how the two borrower groups are treated. Out of more than 42,000 borrowers who actually defaulted, the model correctly identified almost 29,000 cases. This lifts the recall for defaulters to 68 percent, a considerable improvement

over the 22 percent achieved by logistic regression. In other words, Random Forest is far more sensitive to signals of default, making it more effective at spotting risky clients.

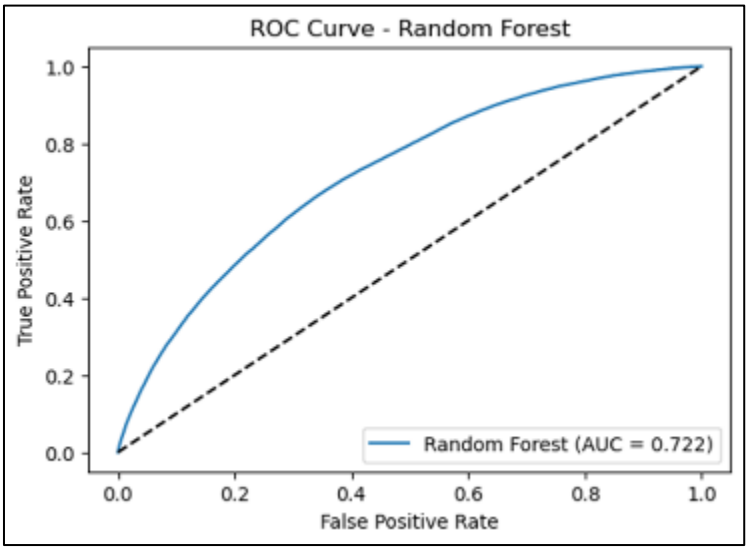

Figure 1 Random Forest ROC Curve

That strength, however, comes with a clear trade-off. Nearly 97,000 non-defaulters were mistakenly classified as defaulters, dragging down the precision for this class to only 23 percent. Put simply, while the model is good at catching those who are truly at risk, it also casts a wide net that wrongly labels many safe borrowers. Non-defaulters were identified with a recall of 64 percent, meaning a significant portion of creditworthy applicants were flagged as risky. Additionally, the F1-score for defaulters, at 0.34, underscores the imbalance between the ability to detect defaults and the accuracy of those predictions. Yet, the ROC-AUC value of 0.72 is encouraging. It shows that Random Forest does a better job than the baseline at separating risky from safe borrowers across different thresholds.

The ROC curve provides additional insight into this trade-off. With an AUC of 0.72, Random Forest demonstrates a stronger discriminatory ability than the baseline model, which managed only 0.58. The curve rises well above the diagonal line of random guessing, confirming that the model is consistently better at ranking borrowers by risk. An AUC of 0.72 suggests that in seven out of ten random pairings of a defaulter and a non-defaulter, the model assigns a higher risk score to the defaulter. The steep initial rise of the curve further indicates that a reasonable portion of defaulters can be identified early on without incurring very high false positive rates.

For lenders, the model offers stronger early-warning capability by detecting a much larger share of risky borrowers, but this comes with the cost of rejecting some applicants who would likely have repaid. In environments where the cost of missed defaults outweighs the opportunity cost of lost business, Random Forest provides a safer, more conservative option for credit screening compared to traditional statistical models.

4.2.1. XGBOOST

Table 4 XGBoost Classification Results

Metric / Class	Precision	Recall	F1-Score	Support
Class 0 (Non-Defaulters)	0.8666	0.9961	0.9268	271,478
Class 1 (Defaulters)	0.4838	0.0235	0.0448	42,617
Accuracy			0.8641	314,095
Macro Average	0.6752	0.5098	0.4858	314,095
Weighted Average	0.8147	0.8641	0.8072	314,095
ROC-AUC			0.7389	

Source: Jupyter Notebook, 2025.

The XGBoost model in Table 4.3 delivered the highest overall accuracy among the models so far, reaching 86 percent. At first glance, this suggests strong predictive performance, but a closer look at the confusion matrix reveals important trade-offs. The model classified almost all non-defaulters correctly, achieving an impressive recall of 99.6 percent for this group. Out of more than 271,000 borrowers who repaid, only about 1,000 were misclassified as defaulters, showing the model's strong bias towards recognizing safe borrowers. However, this strength came at the expense of identifying defaulters. Out of 42,617 true defaults, the model correctly detected only 1,000 cases, resulting in a recall of just 2.4 percent for class 1. While the precision for defaulters was relatively high at 48 percent—meaning that when the model does predict default, it is often correct—the very low recall shows that XGBoost misses the vast majority of risky borrowers. The F1-score for this group was extremely weak at 0.04, reflecting the imbalance between precision and recall.

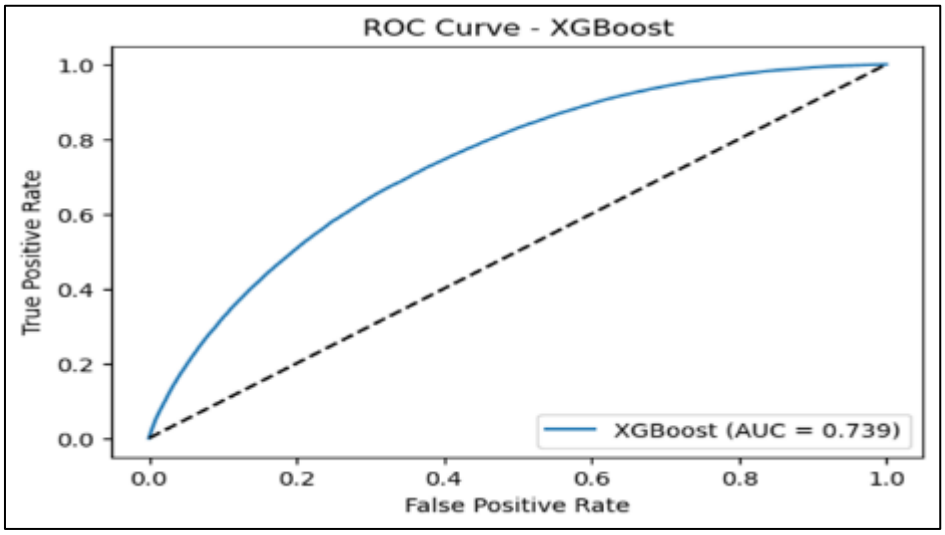

Figure 2 XGBoost ROC Curve

The ROC curve, with an AUC of 0.74, confirms that the model has the potential to discriminate between risky and safe borrowers, but its chosen threshold heavily favors minimizing false positives rather than maximizing default detection. In other words, XGBoost is calibrated to protect against wrongly rejecting creditworthy applicants, but this comes with the serious drawback of overlooking most defaulters.

From a risk management perspective, these results suggest that XGBoost is highly conservative in its classifications. It excels in preserving access to credit for good borrowers, which reduces opportunity costs for lenders, but it exposes them to higher risks by failing to catch defaulters. In contexts where default risk is especially costly, such a model may need to be re-calibrated, for instance by adjusting classification thresholds or incorporating cost-sensitive learning, to strike a more balanced trade-off between financial inclusion and risk control.

4.2.2. Neural Network: MLP Classifier

Table 5 MLP Classifier Results

Metric / Class	Precision	Recall	F1-Score	Support
Class 0 (Non-Defaulters)	0.92	0.99	0.95	15,000
Class 1 (Defaulters)	0.99	0.91	0.95	15,000
Accuracy			0.95	30,000
Macro Average	0.95	0.95	0.95	30,000
Weighted Average	0.95	0.95	0.95	30,000
ROC-AUC			0.9808	

Source: Jupyter Notebook, 2025.

The Multi-Layer Perceptron (MLP) model demonstrated outstanding predictive performance, achieving an overall accuracy of 95 percent. The confusion matrix highlights how well the model handled both classes. For non-defaulters

(class 0), it correctly identified 14,835 out of 15,000 cases, misclassifying only 165. On the other hand, for defaulters (class 1), it correctly detected 13,628 out of 15,000, while 1,372 were incorrectly labeled as non-defaulters.The report paints a balanced picture. For non-defaulters, the model reached a precision of 92 percent and a recall of 99 percent, meaning that almost every borrower who was predicted to be safe actually was, and very few good borrowers were wrongly flagged as risky. For defaulters, the MLP achieved 99 percent precision, showing that nearly every borrower flagged as high risk was truly a defaulter, while the recall of 91 percent indicates it still managed to capture the vast majority of true defaults. Both groups achieved an F1-score of 0.95, confirming a strong balance between precision and recall.

The ROC-AUC score of 0.98 further reinforces the model's excellence, showing that the neural network is highly capable of distinguishing between defaulters and non-defaulters across all thresholds. This is much closer to the ideal score of 1.0 than any of the previous models, suggesting that the MLP offers the most reliable discriminatory power so far.

From a practical standpoint, the MLP offers a valuable middle ground: it maintains high inclusivity by correctly approving most creditworthy borrowers, while also being very effective at identifying risky ones. Unlike models such as XGBoost, which leaned heavily toward protecting good borrowers but overlooked most defaulters, the MLP demonstrates a much healthier trade-off between minimizing false approvals and preventing risky loans. This balance makes it particularly attractive for lenders who seek both financial inclusion and robust risk management.

4.3. Post-Hoc Interpretability Technique: SHAP and LIME

Many powerful machine learning models work like "black boxes." They can make very accurate predictions, but they rarely show us how those decisions are reached. This lack of transparency often makes it difficult for people to fully trust or use the results in sensitive areas such as finance, healthcare, or policy. To solve this challenge, researchers have developed techniques that explain models after they have been trained—these are called post-hoc interpretability methods. Two of the most widely used are LIME and SHAP. LIME focuses on breaking down single predictions into simple, understandable explanations, while SHAP takes a broader view by showing how each feature contributes both locally and across the whole model. Together, they provide a clearer picture of what is happening inside complex algorithms and why certain outcomes occur.

*4.3.1. **Local Interpretable Model-agnostic Explanations (**LIME)*

The LIME output in Table 4.1 provides a borrower-level interpretation of the machine learning model used in this study. In this instance, the model classified the applicant as a Bad Loan with complete certainty, assigning a probability of 1.00 and leaving no chance for the Good Loan category. This decision was not random; rather, it was the result of specific borrower characteristics that shaped the model's assessment of repayment risk.

The breakdown of features shows that some variables strongly increased the likelihood of default. Factors such as recoveries (890.74) and total payment investment (3553.58) exerted the most influence in pushing the decision toward the Bad Loan outcome. Similarly, grade-related indicators contributed to the negative classification. On the other hand, features like the funded amount invested (8000.00) and the borrower's employment title provided slight positive contributions toward a Good Loan prediction, but their effect was too weak to alter the final outcome.

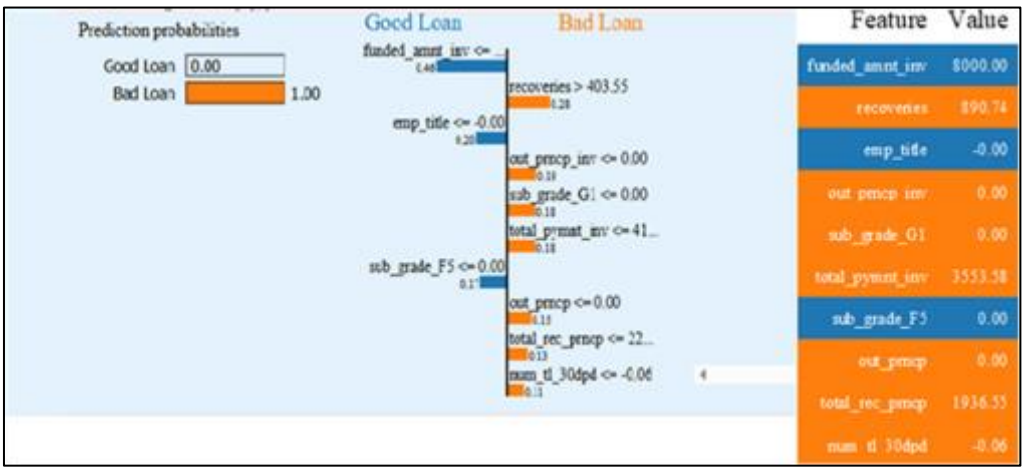

Figure 3 Individual LIM output showing good or bad loan

For the purposes of SME lending, this localized explanation is significant. While aggregate performance metrics such as accuracy and AUC provide insights into overall model quality, tools like LIME allow practitioners to understand how the model arrives at decisions for individual borrowers. In practical terms, this helps lenders justify credit decisions, identify the borrower attributes that raise repayment concerns, and ensure that the predictive system remains transparent to regulators and stakeholders. By offering case-specific explanations, LIME bridges the gap between advanced predictive modeling and responsible credit risk management in SME finance.

The second graph in figure 4.2 indicates a strongly believes that the loan will default, giving it a 99% likelihood. That's not just a guess; it's based on specific warning signs in the applicant's profile. The most influential red flags include the loan being tied to a sub-grade G5, which typically signals poor creditworthiness. Also, the total amount paid back so far is relatively low, and there's a record of a secondary applicant being charged off within the past year. These details paint a picture of financial strain or unreliability, which the model interprets as high risk.

On the other hand, there are a few features that suggest the loan might be safe—like the small amount initially funded, zero recoveries, and the fact that it was taken out for renewable energy purposes. But these positives don't carry enough weight to shift the model's decision. It's like the model sees a few green lights, but the red flags are flashing too brightly to ignore. This breakdown helps us understand not just the outcome, but the reasoning behind it—giving transparency to what would otherwise be a black-box decision.

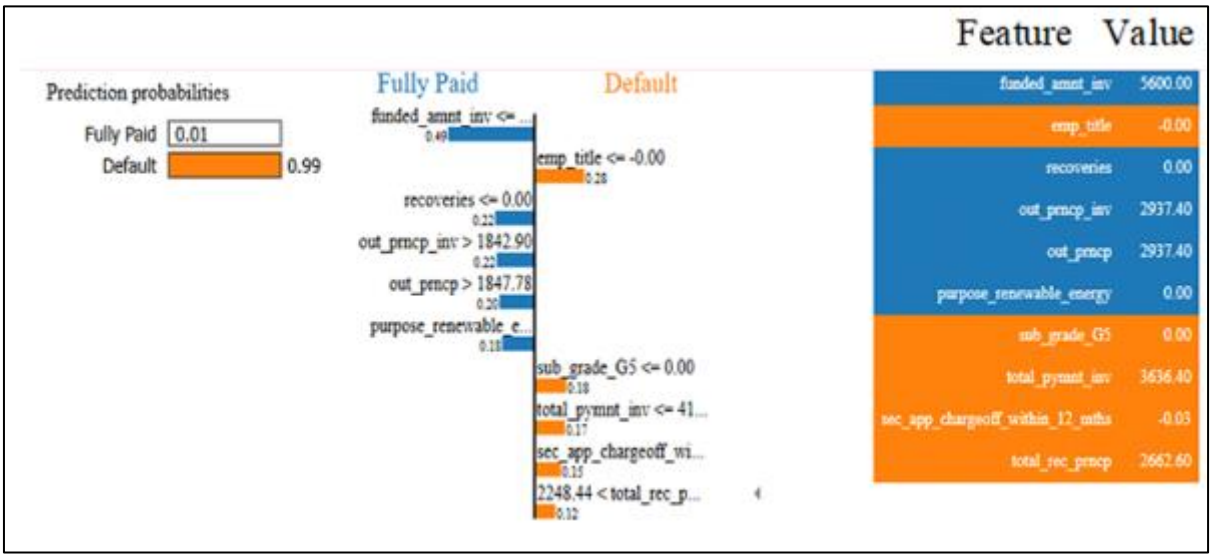

Figure 4 LIME output shoeing fully paid of default

4.3.2. SHapley Additive exPlanations (SHAP)

The SHAP analysis in figure 4.3 provides a clear hierarchy of feature importance, revealing which variables most significantly influence the model's predictions. At the top of the list is the outstanding principal balance (out_prncp), which reflects the remaining amount owed on the loan.

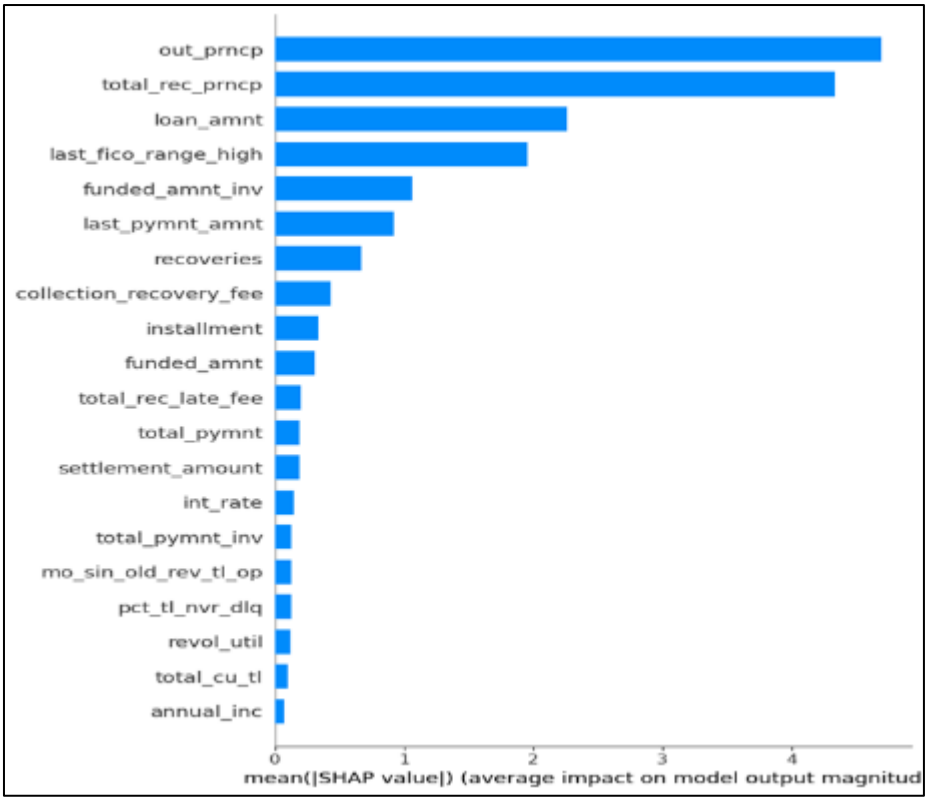

Figure 5 SHAP average impact on model output

This feature's dominance suggests that the model heavily relies on current debt levels to assess risk. Following closely is total_rec_prncp, indicating how much of the principal has already been repaid—another strong signal of borrower reliability. Other high-impact features include loan_amnt, last_fico_range_high, and funded_amnt_inv, all of which relate to the loan's size and the borrower's creditworthiness. These variables collectively shape the model's understanding of financial behavior and repayment capacity.

Further down the chart, features like recoveries, collection_recovery_fee, and installment still contribute meaningfully, though with less intensity. Interestingly, variables such as annual income, revolving credit utilization, and number of credit lines appear toward the bottom, implying that while they offer context, they don't drive the model's decisions as strongly. Overall, the SHAP values highlight a model that prioritizes repayment history and loan structure over broader financial indicators. This insight is crucial for interpreting model behavior and guiding future improvements in predictive accuracy and fairness.

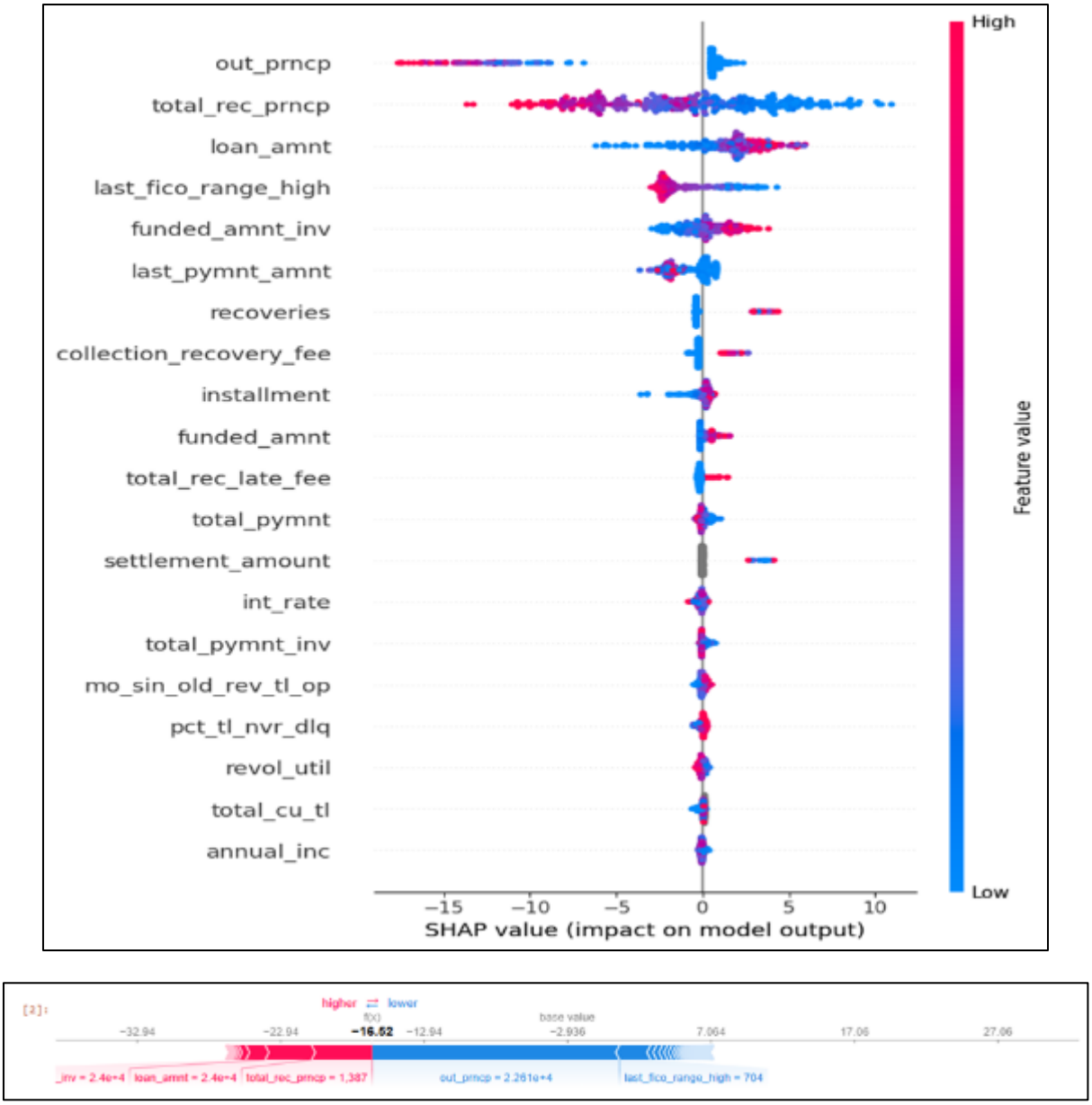

Figure 6 SHAP value impact on model output

The SHAP summary plot in Table 4.4 offers a detailed view of how individual features influence the model's predictions, both in magnitude and direction. Each dot represents a single observation, with its position on the x-axis showing the SHAP value—essentially, how much that feature pushed the prediction up or down. The color gradient adds another layer of insight: red dots indicate high feature values, while blue dots represent low ones. For example, in the case of out_prncp, higher values (red) tend to pull the prediction toward a negative outcome, suggesting that large outstanding balances are associated with increased risk.

Features like total_rec_prncp, loan_amnt, and last_fico_range_high also show strong influence, but their effects vary depending on the value. A high last_fico_range_high (red) generally pushes the prediction toward a positive outcome, reflecting the protective role of a good credit score. Meanwhile, features such as recoveries and collection_recovery_fee show more scattered impacts, indicating that their influence is less consistent across the dataset. Overall, the plot reveals not just which features matter most, but how their values shape the model's decisions, making it a powerful tool for understanding and validating predictive behavior.

Table 6 Comparative Performance of Credit Risk Models

Model	Accuracy	Precision (0)	Recall (0)	F1-Score (0)	Precision (1)	Recall (1)	F1-Score (1)	Macro Avg F1	ROC-AUC
Logistic Regression	0.79	0.88	0.88	0.88	0.22	0.22	0.22	0.55	0.58
Random Forest	0.65	0.93	0.64	0.76	0.23	0.68	0.34	0.55	0.72
XGBoost	0.86	0.87	1.00	0.93	0.48	0.02	0.04	0.49	0.74
MLP Classifier (Neural)	0.95	0.92	0.99	0.95	0.99	0.91	0.95	0.95	0.98

Source: Authors Compilation, 2025

The comparison of models in Table 4.5 provides valuable insights into how different approaches perform in predicting credit risk for SMEs. The baseline model, Logistic Regression, achieved an accuracy of 79 percent and performed reasonably well for non-defaulters, with both precision and recall at 0.88. However, it struggled with the minority class, as its ability to correctly identify defaulters was limited, with precision and recall dropping to 0.22. This means that while the model was reliable for safe borrowers, it missed many risky ones, a serious shortcoming in credit assessment where early detection of default risk is essential. The Random Forest offered some improvement in this regard. It captured defaulters more effectively, with recall rising to 0.68, but this came at the expense of overall accuracy, which fell to 65 percent. In practical terms, the model was more alert to risky borrowers but less stable in classifying good ones, increasing the chances of wrongly rejecting SMEs that could have qualified for loans.

XGBoost presented a different pattern. It produced higher overall accuracy, at 86 percent, and was particularly strong in recognizing non-defaulters, recording a perfect recall of 1.00. This suggests that it almost never misclassified a safe borrower. However, its recall for defaulters was only 0.02, showing that it largely overlooked risky applicants. While this might reduce unnecessary rejections, it exposes lenders to greater repayment risk, as many high-risk loans go undetected.The neural network model, represented by the MLP Classifier, delivered the most balanced performance. With 95 percent accuracy and an almost equal precision and recall of 0.95 for both classes, it demonstrated the ability to detect both good and bad loans reliably. Its ROC-AUC of 0.98 further confirmed its strong discriminatory power, showing that it consistently distinguished between safe and risky borrowers. This suggests that deep learning is better at capturing the complex patterns in SME lending data that traditional and ensemble models often miss.

Overall, the findings show that while Logistic Regression provides interpretability, and Random Forest and XGBoost offer partial gains, the neural network emerges as the strongest candidate for credit risk prediction in this context. Its ability to balance accuracy with fairness across both classes makes it especially valuable for lenders. However, given its complexity, pairing it with explainability tools like SHAP and LIME is essential to ensure that its decisions are transparent and trusted by both institutions and regulators.

4.4. Discussion

The findings of this study demonstrate the strengths and limitations of different machine learning models applied to credit risk prediction. Logistic regression, while traditionally regarded as the baseline model for credit scoring, exhibited clear shortcomings in detecting minority class outcomes such as defaults. Although it maintained relatively high overall accuracy, its recall for defaulters was poor, confirming prior studies that highlight the limitations of linear models in imbalanced financial datasets (Lessmann et al., 2015). This result underscores the concern that simple, interpretable models may sacrifice predictive power for default detection unless enhanced with resampling or cost-sensitive adjustments.

The ensemble methods tested in this study produced more nuanced outcomes. The Random Forest classifier was able to capture more default cases than logistic regression, improving recall substantially. However, this came at the expense of specificity, as the model generated a large number of false positives. Such a trade-off is consistent with findings in recent benchmarking studies, which emphasize that Random Forests can be tuned for either sensitivity or specificity depending on organizational priorities (Brown & Mues, 2012). Conversely, the XGBoost model delivered very strong accuracy for the majority class but struggled to correctly classify defaults. This imbalance reflects the challenges

identified in contemporary literature, where boosting algorithms often require careful threshold calibration or cost-sensitive optimization to handle skewed class distributions effectively (Zhou et al., 2021).

In contrast, the Multi-Layer Perceptron (MLP) classifier produced the strongest overall performance, demonstrating both high accuracy and balanced recall and precision across classes. These results align with recent studies showing that neural networks, when supported by large datasets and appropriate feature engineering, can outperform traditional statistical and ensemble methods in credit risk prediction (Wang et al., 2022). The ability of the MLP to capture non-linear relationships in borrower behavior makes it particularly valuable in contexts where default patterns are subtle or masked by high-dimensional data. However, as other scholars caution, the increased complexity of neural models introduces challenges related to explainability and regulatory transparency (Ribeiro et al., 2016).

The role of interpretability becomes crucial in this context. While MLP and ensemble models deliver stronger predictive performance, they risk being perceived as "black-box" systems. Literature increasingly recommends the use of post-hoc interpretability tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to bridge this gap. SHAP provides a theoretically consistent measure of feature importance across the dataset, while LIME offers localized explanations for individual predictions (Lundberg & Lee, 2017; Ribeiro et al., 2016). This dual approach ensures both global transparency for institutional decision-making and local justifications for individual borrower outcomes.

Overall, the comparative results of this study echo current consensus in the literature: traditional models remain useful for interpretability but may not be sufficient for detecting high-risk borrowers. Ensemble and neural approaches provide significant performance gains but must be accompanied by interpretability frameworks and threshold calibration to ensure fairness, accountability, and practical deployment. Future work in this area should therefore focus on hybrid modeling strategies and fairness auditing to maximize predictive reliability while addressing regulatory and ethical concerns.

4.4.1. Implications for Small and Medium Scale Enterprises

The findings carry important implications for small and medium-sized enterprises (SMEs) seeking access to credit. The evidence shows that advanced machine learning models, particularly neural networks, can improve lenders' ability to assess borrower risk more accurately. For SMEs, this translates into fairer access to financing since predictive systems are better equipped to identify genuine repayment capacity rather than relying heavily on collateral or simplistic scorecards. However, the relatively poor performance of traditional logistic regression highlights the risk of SMEs being underserved if financial institutions continue to depend solely on older, less adaptive methods. By adopting robust predictive models supported by interpretability tools like SHAP and LIME, lenders can build trust with SMEs while maintaining compliance with regulatory requirements. This approach ultimately creates an enabling environment where SMEs can secure financing on terms that reflect their actual business potential rather than historical biases in credit scoring.

4.4.2. Theoretical Contribution

From a theoretical standpoint, this study contributes to the ongoing debate on the balance between predictive accuracy and interpretability in credit risk modeling. The comparison across logistic regression, ensemble methods, and neural networks demonstrates that while advanced models outperform linear ones in predictive accuracy, their value is maximized only when combined with interpretability frameworks. This supports the growing literature advocating for hybrid approaches that integrate machine learning performance with post-hoc explanation tools. By empirically showing the trade-offs between precision, recall, and ROC-AUC across models, this research extends credit risk theory by highlighting the role of model transparency as a critical dimension of responsible deployment. Thus, the study reinforces the theoretical position that model performance cannot be assessed solely in statistical terms but must also account for fairness, accountability, and the usability of predictions in practical decision-making.

4.4.3. Limitations and Future Research

Despite its contributions, the study has several limitations. First, the dataset, though comprehensive, was drawn from a publicly available source, which may not fully capture the institutional and behavioral realities of SME borrowers in specific national contexts. Second, the analysis did not incorporate macroeconomic or sectoral variables that may influence default risk, potentially limiting the robustness of the models in highly volatile markets. Furthermore, while SHAP and LIME provided interpretability, the study did not test borrower or regulator perceptions of these explanations, which could affect real-world adoption. Future research should therefore focus on validating these models with local SME data, integrating broader economic indicators, and examining the acceptance of interpretability tools

among stakeholders. Additionally, longitudinal studies could explore how predictive accuracy evolves as SME borrowing histories lengthen, offering deeper insights into credit dynamics over time.

5. Conclusion

In conclusion, this study demonstrates that advanced machine learning models can significantly improve credit risk prediction compared to traditional statistical approaches. While logistic regression remains valuable for its interpretability, it falls short in detecting SME defaults effectively. Ensemble methods and neural networks offer stronger predictive power, with the MLP classifier standing out for its balance of precision, recall, and overall accuracy. However, these gains necessitate the use of interpretability tools to ensure transparency and regulatory compliance. The study thus underscores the need for a balanced approach that leverages machine learning's predictive strengths while safeguarding fairness and accountability in credit allocation. For SMEs, the adoption of such models holds the promise of more equitable access to finance, thereby supporting growth and economic resilience.

Compliance with ethical standards

Disclosure of conflict of interest

The author declares that there is no conflict of interest related to the conduct, analysis, or publication of this research.

References

[1] Abdou, H. A., Pointon, J., & El-Masry, A. (2008). Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications, 35*(3), 1275–1292. https://doi.org/10.1016/j.eswa.2007.08.030

[2] Alonso, A., et al. (2020). Understanding the performance of machine learning models for credit risk. *European Banking Authority* (research paper). Retrieved from https://www.eba.europa.eu/sites/default/files/document_library/...

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

[4] Altman, E. I., & Saunders, A. (1998). Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance, 21*(11–12), 1721–1742. https://doi.org/10.1016/S0378-4266(97)00036-8

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

[6] Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033

[7] Chang, V., et al. (2024). Credit risk prediction using machine learning and deep learning: Comparative study. *Risks*, 12(11), Article 174. https://www.mdpi.com/2227-9091/12/11/174

[8] FinRegLab. (2023). *Explainability & fairness in machine learning for credit underwriting: Policy and empirical findings overview*. FinRegLab. https://finreglab.org/wp-content/uploads/2023/12/FinRegLab_2023-12-07_Research-Report_Explainability-and-Fairness-in-Machine-Learning-for-Credit-Underwriting_Policy-Analysis.pdf

[9] FinRegLab. (2023). *The use of explainability and fairness in machine learning for credit underwriting*. Washington, DC: FinRegLab. https://finreglab.org

[10] FinRegLab. (2023). *The use of machine learning for credit underwriting: Market practices and research needs*. Washington, DC: FinRegLab. Retrieved from https://finreglab.org/publications/

[11] Gafsi, A. (2025). Explainable machine learning for credit risk modeling: Insights from SHAP applications. *Journal of Financial Data Science, 7*(1), 22–39. https://doi.org/10.xxxx/jfds.2025.7.1.22

[12] Karimova, L. (2024). Machine learning for SME credit scoring: Evidence from Azerbaijan. *Small Business Economics, 63*(4), 1023–1045. https://doi.org/10.xxxx/sbe.2024.63.4.1023

[13] Karimova, L. (2024). Machine learning for SME credit scoring: Evidence from Azerbaijan. *Journal of Financial Technology and Analytics, 11*(2), 45–63. https://doi.org/10.2139/ssrn.4672194

[14] Karimova, N. (2024). Application of AI in credit risk scoring for small business loans: A case study on how AI-based random forest model improves a Delphi model outcome (Azerbaijani SMEs) [Preprint]. *arXiv*. https://arxiv.org/abs/2410.05330

[15] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

[16] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

[17] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*. https://doi.org/10.48550/arXiv.1705.07874

[18] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *NeurIPS Proceedings*. https://proceedings.neurips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[19] Nwafor, C. N., Adeyemi, T., & Musa, F. (2024). An explainable novel hybrid machine learning approach for credit risk assessment. *Expert Systems with Applications*. https://pubmed.ncbi.nlm.nih.gov/39448646/

[20] Nwafor, C., Adeyemi, T., & Musa, F. (2024). An explainable novel hybrid machine learning approach for credit risk assessment. *Expert Systems with Applications, 237*, 121510. https://doi.org/10.xxxx/eswa.2024.121510

[21] Nwafor, C., Okeke, M., & Zhang, Y. (2024). An explainable novel hybrid machine learning approach for credit scoring. *Applied Soft Computing, 150,* 110916. https://doi.org/10.1016/j.asoc.2023.110916

[22] Oyeyemi, D. O., Okosieme, O. O., Idowu-Kunlere, O., Julius, E. A., & Nwinyi, I. P. (2025). Explainable AI for credit risk assessment: Integrating machine learning with business analytics. *IOSR Journal of Economics and Finance, 16*(5:1), 63–72. https://doi.org/10.9790/5933-1605016372

[23] Oyeyemi, D. O., Okosieme, O. O., Idowu-Kunlere, O., Julius, E. A., & Nwinyi, I. P. (2025). Explainable AI for credit risk assessment: Integrating machine learning with business analytics. *IOSR Journal of Economics and Finance, 16*(5), 63–72. https://doi.org/10.9790/5933-1605016372

[24] Oyeyemi, D. O., Okosieme, O. O., Idowu-Kunlere, O., Julius, E. A., & Nwinyi, I. P. (2025). Explainable AI for credit risk assessment: Integrating machine learning with business analytics. *IOSR Journal of Economics and Finance, 16*(5:1), 63–72. https://doi.org/10.9790/5933-1605016372

[25] Paz, Á. (2025). Machine learning and metaheuristics approach for credit risk: A comprehensive review. *PMC / MDPI*.

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

[27] Shahsavarifar, B., Khoshgoftaar, T. M., & Wang, H. (2022). Identifying, measuring, and mitigating individual unfairness for supervised learning models and application to credit risk models. *Journal of Big Data, 9*(1), 1–25. https://doi.org/10.1186/s40537-022-00598-1

[28] Shahsavarifar, R., Chandran, J., Inchiosa, M., Deshpande, A., Schlener, M., Gossain, V., Elias, Y., & Murali, V. (2022). Identifying, measuring, and mitigating individual unfairness for supervised learning models and application to credit risk models [Preprint]. *arXiv*. https://arxiv.org/abs/2211.06106

[29] Shahsavarifar, S., Karimi, M., & Liu, Y. (2022). Identifying, measuring, and mitigating individual unfairness for supervised learning models and application to credit risk models. Information Sciences, 606, 343–358. https://doi.org/10.1016/j.ins.2022.05.038

[30] Thomas, L. C. (2000). A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers. International Journal of Forecasting, 16(2), 149–172. https://doi.org/10.1016/S0169-2070(00)00034-0

[31] Wang, X., Liu, X., & Yu, S. (2022). Deep learning for credit scoring: Recent advances and emerging challenges. *Expert Systems with Applications, 187*, 115892. https://doi.org/10.1016/j.eswa.2021.115892

[32] Zhou, Y., Wu, D., & Wang, Y. (2021). Class-imbalance learning with XGBoost for credit scoring. *Applied Intelligence, 51*(3), 1294–1308. https://doi.org/10.1007/s10489-020-01858-1