An Interpretable Machine Learning Framework for Loan Default Risk Prediction in Financial Systems

Nayana H N[1], Nandini G V[1], Pruthvi D M[1], Ramya P[1]

[1]Department of Computer Science and Engineering, RV Institute of Technology and Management, Bangalore, India

[2]Faculty Guide, Department of Computer Science and Engineering, RV Institute of Technology and Management, Bangalore, India

How to Cite this Article:
P, R., M, P. D., V, N. G. & N, N. H. (2026). An Interpretable Machine Learning Framework for Loan Default Risk Prediction in Financial Systems. International Journal of Creative and Open Research in Engineering and Management, 02(05). https://doi.org/10.55041/ijcope.v2i5.265

https://doi.org/10.55041/ijcope.v2i5.265

1. Abstract

Reliable identification of loan default risk constitutes a strategic priority for financial institutions seeking to limit credit exposure while sustaining sound lending practices. This paper introduces a structured, machine learning–driven framework built on real-world financial data to address this challenge. The pipeline encompasses thorough data preparation—including missing-value removal, binary label encoding, categorical-to-numerical conversion, and class-imbalance correction—to ensure that each model receives consistently formatted, representative input.

Three well-established classifiers—Logistic Regression, Random Forest, and XGBoost—were trained and evaluated across a standard set of performance indicators: overall accuracy, area under the receiver operating characteristic curve (ROC-AUC), precision, recall, and F1-score. XGBoost delivered the strongest accuracy by leveraging its gradient-boosting architecture to capture nonlinear interactions within the data. Complementing the predictive component, an explainability layer based on SHapley Additive exPlanations (SHAP) was integrated to surface the relative contribution of each input variable to individual predictions, thereby converting an otherwise opaque ensemble into a transparent decision-support tool.

The resulting framework strikes a deliberate balance between predictive power and model transparency—a pairing that is particularly valuable in regulated credit-risk environments where decisions must be both accurate and justifiable.

Keywords: *Loan Default Prediction, XGBoost, Random Forest, Logistic Regression, SMOTE, SHAP, Credit Risk, Interpretable Machine Learning*

2. Introduction

The rapid expansion of digital lending platforms has fundamentally altered the volume and diversity of financial data that institutions must process. Assessing the creditworthiness of prospective borrowers and forecasting the probability of loan non-repayment have consequently become mission-critical capabilities. Failure to identify high-risk applicants reliably translates directly into elevated default rates, increased provisioning costs, and, in severe cases, threats to institutional solvency.

Classical credit-scoring approaches—grounded in manual underwriting guidelines or simple linear regression frameworks—struggle to model the intricate, nonlinear dependencies embedded in modern borrower data. Variables such as income, credit utilisation, loan purpose, and repayment behaviour interact in ways that defy straightforward parametric representation. Ensemble learning methods have addressed this limitation by constructing multiple weak learners that collectively approximate complex decision boundaries, consistently outperforming single-model baselines on credit classification tasks [1, 3].

A persistent criticism of high-performing ensemble models, however, is their opacity. When a model operates as a black box, regulators, compliance officers, and customers cannot scrutinise the rationale behind an adverse lending decision—a situation at odds with principles of fairness and accountability embedded in financial regulation worldwide [8]. Explainability techniques that attribute predictions to individual features address this gap by rendering complex models interpretable without sacrificing their predictive superiority.

This paper describes an end-to-end framework that integrates robust preprocessing, class-imbalance handling, multi-model training, and post-hoc interpretability to produce a system that is simultaneously accurate, reproducible, and transparent. The study makes the following specific contributions:

- A systematic preprocessing pipeline tailored to real-world lending data, incorporating SMOTE-based oversampling and feature standardisation.
- A comparative evaluation of three classifiers—Logistic Regression, Random Forest, and XGBoost—using multiple performance metrics.
- Integration of SHAP-based explainability to translate ensemble predictions into feature-level insights accessible to non-technical stakeholders.
- Empirical demonstration that accuracy and interpretability are complementary rather than competing objectives in credit risk modelling.

3. Related Work

Substantial prior research has examined machine learning approaches to credit risk. Kang et al. [1] compared a suite of supervised classifiers on lending-platform data, establishing that tree-based ensembles outperform linear models when feature interactions are strong. Sayed et al. [2] extended this line of enquiry by incorporating deep learning alongside ensemble methods and evaluating the sensitivity of results to different class-balancing strategies, concluding that gradient boosting remained competitive across most configurations.

Akinjole et al. [3] specifically investigated ensemble combinations for default-risk prediction and reported improvements in minority-class recall when stacking was employed. On the interpretability front, Bracke et al. [8] provided an early application of explainability methods—including SHAP—to default-risk models within a central-bank context, demonstrating that feature attributions align intuitively with domain knowledge about borrower risk drivers. Emmanuel et al. [6] proposed a stacked classifier augmented with a feature-selection module, achieving strong performance on an imbalanced lending dataset while maintaining interpretability through variable-importance rankings.

Collectively, these contributions motivate the present work: a framework that combines the predictive strength of XGBoost with rigorous preprocessing and SHAP-based transparency, validated on a publicly available lending dataset.

4. Methodology

The methodology follows a linear sequence of stages, each designed to address a specific challenge in constructing a reliable and interpretable default-prediction model. Figure 1 provides a high-level illustration of the overall workflow.

4.1 Dataset

Borrower and loan records were sourced from the Lending Club public repository. A working subset of approximately 50,000 observations was retained to balance computational tractability with statistical representativeness. The feature set captures both static borrower attributes and dynamic loan characteristics: loan amount, interest rate, loan term, employment length, home-ownership status, loan purpose, annual income, debt-to-income (DTI) ratio, revolving credit balance, credit utilisation rate, and FICO score range. The binary target variable distinguishes fully repaid loans (label 0) from defaulted loans (label 1).

4.2 Data Preprocessing

Any record containing missing values was removed to prevent systematic bias during training. The target column was re-coded into the binary {0, 1} format described above. Categorical variables—loan term, employment length, home ownership, and loan purpose—were mapped to integer codes via label encoding, making them compatible with all three classifiers evaluated. Although ordinal encoding implicitly imposes an ordering, the tree-based models used in this study are insensitive to monotone transformations of categorical features, and the linear baseline benefits from compact numerical representation.

4.3 Exploratory Data Analysis

Prior to modelling, the distribution of each variable and pairwise feature correlations were examined visually. Histogram plots revealed pronounced right-skewness in income and loan-amount distributions, while the correlation heatmap confirmed that interest rate and FICO score carry the strongest marginal associations with default status. These findings informed subsequent decisions about feature scaling and model selection.

4.4 Train–Test Partitioning

The preprocessed dataset was divided into a training set (80%) and a held-out test set (20%) using a stratified random split that preserves the class-ratio proportions in both partitions. All preprocessing transformers (scaler, SMOTE) were fitted exclusively on the training data to prevent data leakage.

4.5 Class-Imbalance Correction

Lending data inherently contains far more repaid loans than defaults, a class skew that can cause naïve classifiers to achieve high accuracy simply by predicting the majority class for every observation. The Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training partition to generate synthetic minority-class examples by interpolating between existing default records in feature space. This procedure produced a balanced training distribution without modifying the test set, preserving the integrity of performance evaluation.

4.6 Feature Scaling

Standardisation (zero mean, unit variance) was applied to all continuous features. While tree-based models are intrinsically scale-invariant, standardisation is necessary for Logistic Regression (gradient convergence) and was applied uniformly across all classifiers to ensure comparability. Post-scaling, each feature occupies a

comparable numerical range, preventing high-magnitude variables from dominating distance or gradient computations.

4.7 Model Development

Three classification algorithms were implemented:

- Logistic Regression served as a linear baseline. Its probabilistic output and coefficient-level interpretability make it a natural reference point for assessing the incremental benefit of nonlinear methods.
- Random Forest is a bagging ensemble that constructs a large number of decision trees on bootstrap samples of the training data and aggregates predictions by majority vote. It provides built-in feature-importance estimates and is robust to overfitting through variance averaging.
- XGBoost is a gradient-boosting framework that builds trees sequentially, with each tree correcting the residual errors of its predecessors. Regularisation terms in the objective function and efficient handling of sparse data make it well-suited to structured tabular datasets with mixed feature types.

Each model was trained with default hyperparameters as a first pass; no extensive tuning was performed in this study, leaving hyperparameter optimisation as a direction for future work.

4.8 Model Evaluation

Models were assessed on the held-out test set using five metrics: accuracy (fraction of correctly classified observations), ROC-AUC (area under the true-positive-rate–false-positive-rate curve across all decision thresholds), precision for the default class (fraction of predicted defaults that are genuine), recall for the default class (fraction of genuine defaults that are detected), and the F1-score (harmonic mean of precision and recall). This multi-metric approach prevents over-reliance on accuracy alone, which can be misleading under class imbalance.

4.9 Explainability via SHAP

SHAP values were computed for the XGBoost model to quantify each feature's contribution to individual predictions. SHAP values originate from cooperative game theory: they represent the average marginal contribution of a feature across all possible feature coalitions. Positive SHAP values indicate that a feature pushed the prediction toward default; negative values indicate the opposite. Summary plots aggregate SHAP values across all test observations to reveal global feature-importance rankings alongside the direction and magnitude of each feature's effect.

5. Results and Analysis

5.1 Class Distribution

The raw dataset exhibits a notable imbalance: fully paid loans substantially outnumber defaults, with approximately 30,000 paid-in-full records against roughly 10,000 default cases. This disparity underscores the necessity of SMOTE; without correction, a classifier predicting non-default for every sample would achieve an apparent accuracy of approximately 75% while failing entirely to identify at-risk borrowers.

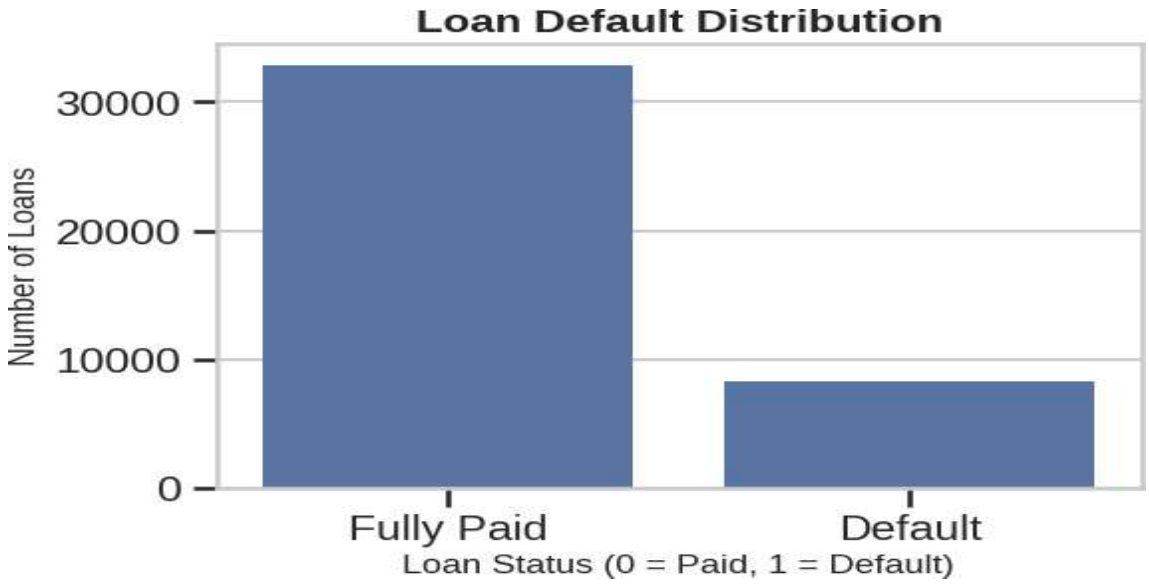


Fig 1: Distribution of Loan Status

5.2 Feature Correlations

The correlation heatmap indicates several noteworthy associations. Interest rate shows moderate positive correlation with loan status, consistent with the intuition that higher rates are charged to riskier borrowers who are more likely to default. FICO score displays a negative correlation with default, confirming its role as a protective factor. Loan amount and revolving credit balance are positively correlated with each other, likely reflecting higher overall debt levels among borrowers who access large loans. Employment length, home ownership, and DTI show weaker but still informative associations with the target variable.

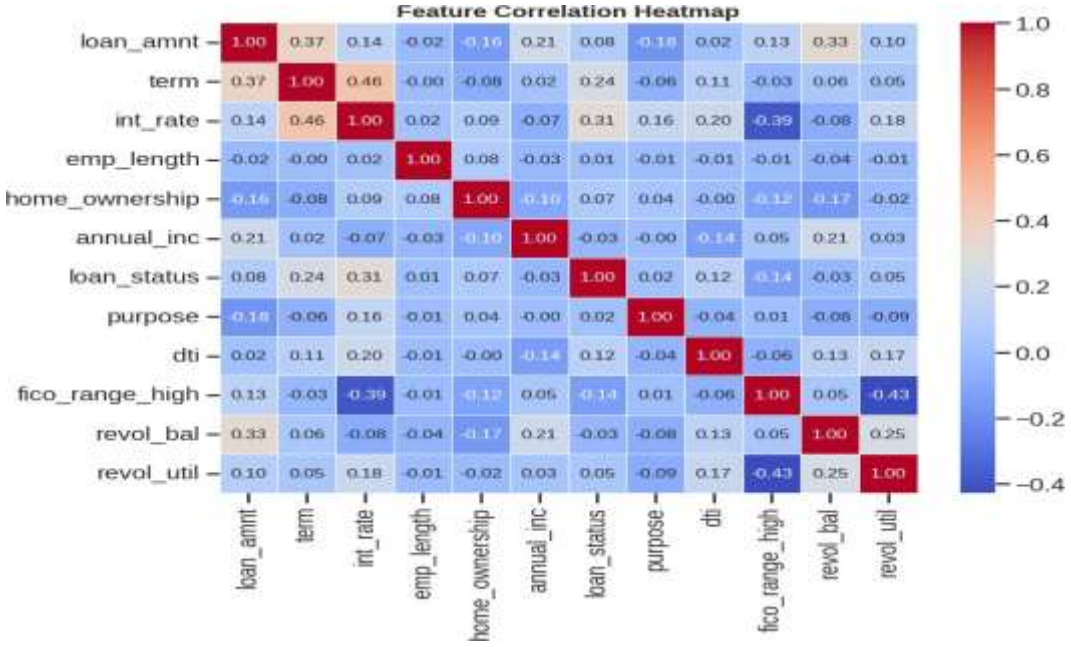


Fig 2: Correlation Heatmap of Features

5.3 Comparative Model Performance

Model	Accuracy	ROC-AUC	Precision (Default)	Recall (Default)	F1-Score (Default)
Logistic Regression	0.673	0.647	0.33	0.66	0.42
Random Forest	0.754	0.602	0.37	0.35	0.36
XGBoost	0.785	0.588	0.44	0.26	0.33

Table 1: Comparative performance of classification models across five evaluation metrics

XGBoost achieved the highest overall accuracy (0.785) and the best precision for the default class (0.44), indicating fewer false alarms relative to the other models. Logistic Regression, despite its lower accuracy (0.673), recorded the highest recall for defaults (0.66)—meaning it captured a greater share of actual defaults at the cost of more false positives. Random Forest occupied an intermediate position on accuracy but underperformed on both recall and F1-score for the minority class.

The relatively high ROC-AUC of Logistic Regression (0.647) compared to XGBoost (0.588) warrants comment. ROC-AUC measures discriminative ability across all decision thresholds and is sensitive to the score distribution rather than the final binary decision. The discrepancy suggests that the XGBoost probability calibration may benefit from post-hoc calibration, and that the optimal threshold for XGBoost may differ from the default 0.5 cutoff. Future work should explore threshold tuning to improve recall without disproportionately inflating false positives.

5.4 ROC Curve

The ROC curve for XGBoost illustrates progressive discrimination capability: the curve rises steeply from the origin and plateaus near the upper-left corner, reflecting the model's ability to achieve high true-positive rates while keeping false-positive rates comparatively low across a range of operating points. The area enclosed beneath this curve quantifies overall separability between repaid and defaulted loan classes.

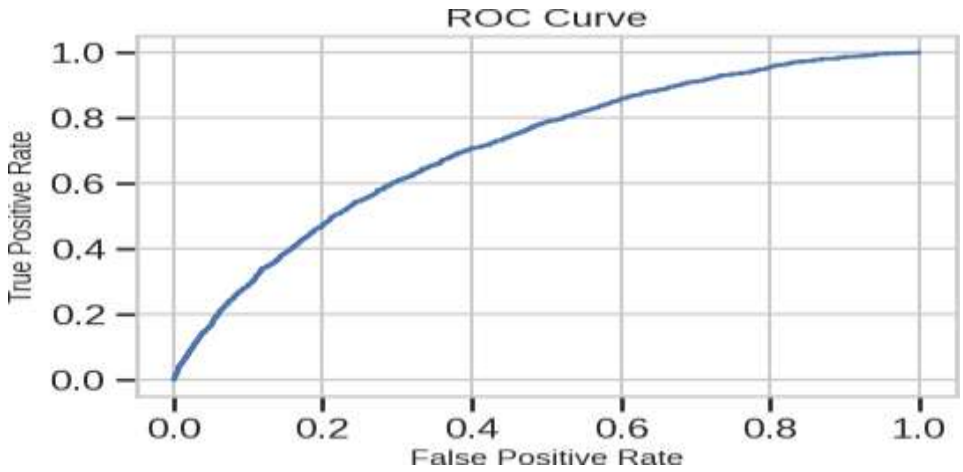
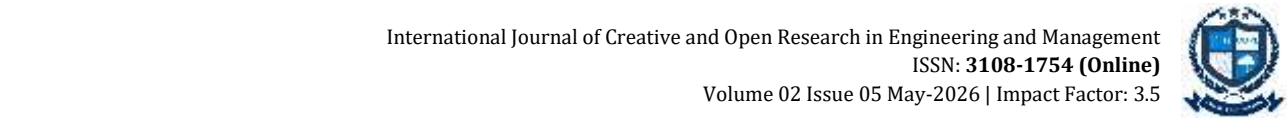


Fig 3: ROC Curve of XGBoost Model



5.5 Feature Importance

XGBoost's built-in impurity-based feature-importance scores identify interest rate (int_rate) as the single most influential predictor, contributing approximately 0.31 on a normalised scale. FICO score range (fico_range_high) ranks second at roughly 0.27. These two financial indicators collectively account for more than half of the total importance mass, reflecting their well-documented relevance in credit-risk literature. Home ownership, loan purpose, annual income, and loan amount contribute moderate importance, while employment length, DTI, revolving balance, and credit utilisation play smaller but non-trivial roles.

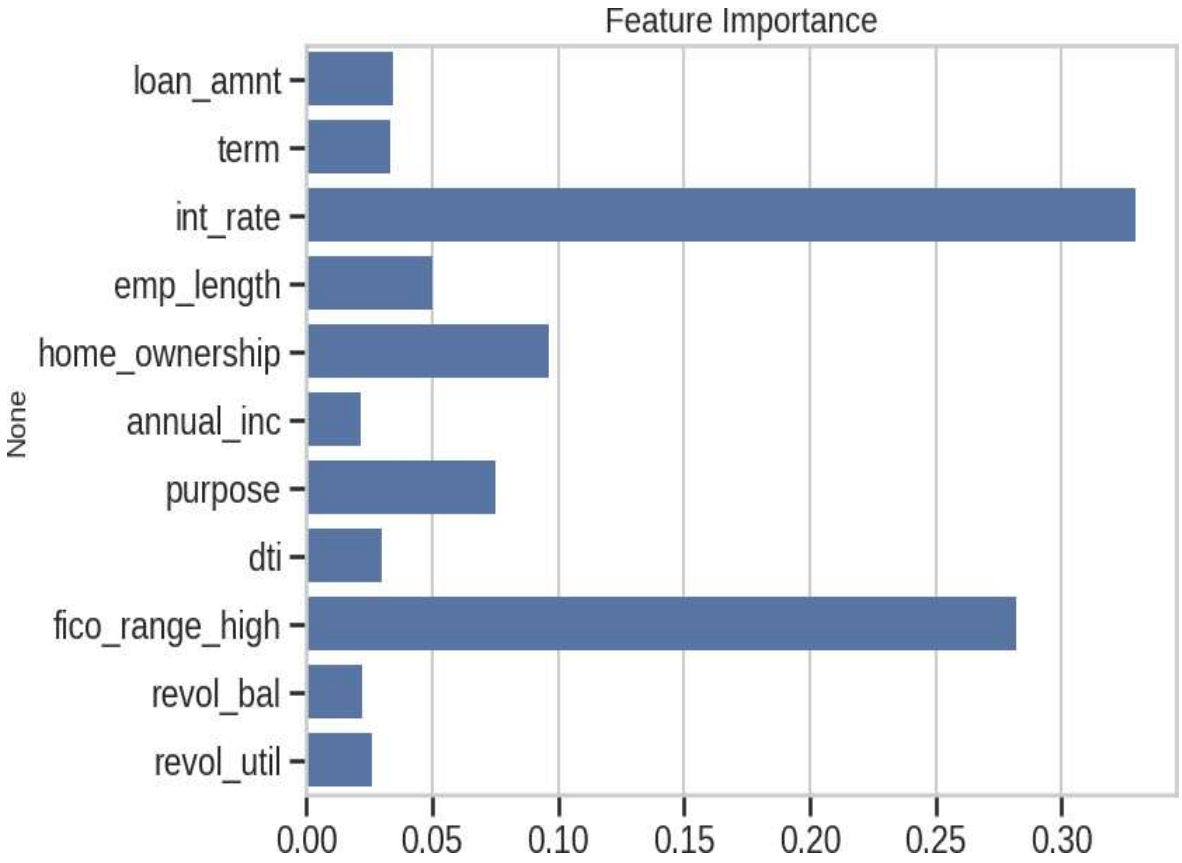


Fig 4: Feature Importance of the XGBoost Model

5.6 SHAP-Based Explainability

The SHAP summary plot complements the aggregate importance rankings by revealing the directionality and heterogeneity of each feature's effect. High interest rates consistently push predictions toward default (positive SHAP values), while higher FICO scores consistently reduce predicted default probability (negative SHAP values). These directional patterns align with established credit-risk intuition and lend credibility to the model's decision logic.

Crucially, the SHAP framework operates at the level of individual predictions, enabling practitioners to explain why a specific borrower received a particular risk assessment—a capability that distinguishes it from global-only importance measures and makes it directly applicable to adverse-action notices required under lending regulation.

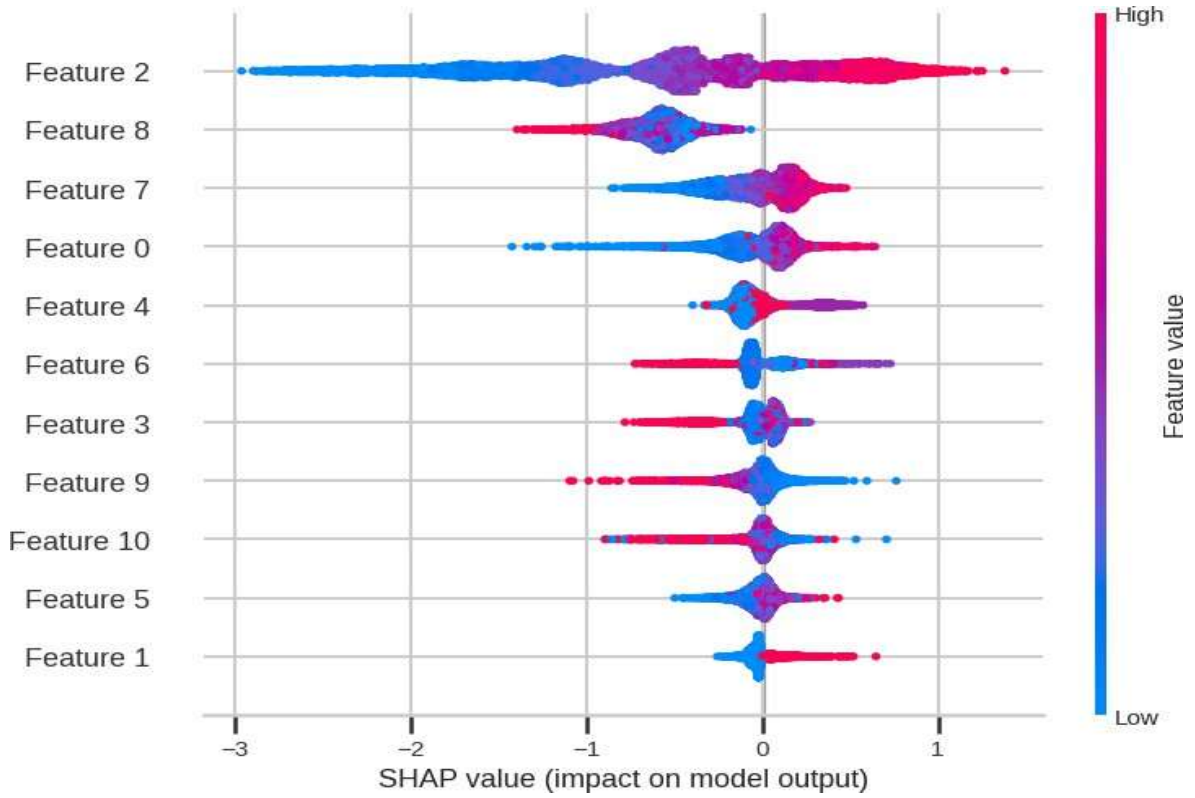


Fig 5: SHAP Summary Plot Showing Feature Contributions

6. Discussion

The findings confirm that gradient-boosting methods retain their empirical advantage over linear classifiers on structured financial data, consistent with recent benchmarks in the credit-risk literature [2, 3]. The accuracy improvement from Logistic Regression (0.673) to XGBoost (0.785) represents a substantial reduction in classification error, translating to fewer incorrectly classified loan applications at portfolio scale.

The trade-off between precision and recall merits deliberate consideration in practical deployments. A lender primarily concerned with minimising credit losses—i.e., not granting loans to borrowers who will default—would prioritise high recall for the default class, even if this means rejecting some creditworthy applicants. In this scenario, Logistic Regression's recall advantage (0.66 vs. 0.26 for XGBoost) is significant. Conversely, a lender operating in a competitive market where false rejections carry a revenue cost would weight precision more heavily, favouring XGBoost. Decision-makers should select a threshold that reflects the institution's specific loss function rather than defaulting to 0.5.

SMOTE improved minority-class representation during training and contributed to more balanced performance metrics across classes. Feature scaling, while inconsequential for tree-based models, ensured that Logistic Regression converged reliably. The integration of SHAP transformed the opaque XGBoost model into a transparent decision tool, addressing the interpretability deficit that has historically impeded regulatory acceptance of ensemble methods in credit underwriting.

Several limitations should be acknowledged. The dataset spans a single lending platform and a specific historical period, which may limit generalisability to other markets or economic conditions. Label encoding of categorical variables imposes arbitrary ordinal relationships that tree-based models handle gracefully but may distort linear model coefficients. Hyperparameters were left at default values; systematic tuning via cross-validated grid or Bayesian search could further improve performance, particularly for the Random Forest model.

7. Conclusion

This study has presented a reproducible, interpretable machine learning pipeline for loan default prediction. Beginning with raw Lending Club data and progressing through preprocessing, class-imbalance correction, model training, multi-metric evaluation, and SHAP-based explainability, the framework demonstrates that high predictive performance and model transparency are achievable simultaneously.

Among the three classifiers evaluated, XGBoost delivered the highest accuracy (0.785) and the best precision for default identification (0.44), making it the recommended model for deployment scenarios where overall error minimisation is the primary objective. Logistic Regression remains a valuable component of the analytical toolkit where high default-recall is essential or where regulatory constraints mandate fully interpretable linear models.

The SHAP integration provides an actionable explainability layer: interest rate and FICO score emerge as the dominant predictors, consistent with domain expertise, while per-prediction attributions support the generation of borrower-level explanations. This combination of accuracy and transparency positions the framework as a practical tool for responsible, data-driven lending decisions.

Future research directions include: (i) hyperparameter optimisation and ensemble stacking to close the recall gap on the default class; (ii) evaluation on multi-platform and cross-market datasets to test generalisability; (iii) exploration of calibrated probability outputs to improve ROC-AUC for XGBoost; and (iv) incorporation of time-series features capturing borrower behaviour trajectories over the loan lifecycle.

References

[1] Z. Z. Kang, S. Y. Teh, S. Y. G. Tan, and W. C. Ng, “Loan Default Prediction Using Machine Learning Algorithms,” Journal of Informatics and Web Engineering, vol. 4, no. 3, 2025.

[2] E. H. Sayed, A. Alabrah, K. H. Rahouma, M. Zohaib, and R. M. Badry, “Machine Learning and Deep Learning for Loan Prediction in Banking: Exploring Ensemble Methods and Data Balancing,” IEEE Access, vol. 12, 2024.

[3] A. Akinjole et al., “Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction,” Mathematics, vol. 12, no. 3423, 2024.

[4] N. O. Collins and I. Emmanuel, “Machine Learning for Credit Risk: Predicting Loan Defaults in Financial Institutions,” 2024.

[5] A. Chouksey et al., “Machine Learning-Based Risk Prediction Model for Loan Applications: Enhancing Decision-Making and Default Prevention,” Journal of Business and Management Studies, 2023.

[6] I. Emmanuel, Y. Sun, and Z. Wang, “A Machine Learning-Based Credit Risk Prediction Engine Using a Stacked Classifier and Feature Selection Method,” Journal of Big Data, 2024.

[7] C. Rao, Y. Liu, and M. Goh, “Credit Risk Assessment Mechanism of Personal Auto Loan Based on PSO-XGBoost Model,” Complex & Intelligent Systems, 2023.

[8] P. Bracke, A. Datta, C. Jung, and S. Sen, “Machine Learning Explainability in Finance: An Application to Default Risk Analysis,” Bank of England Working Paper, 2019.

[9] A. G. Sirishma et al., “Machine Learning for Risk Assessment: A Comparative Study of Models Predicting Loan Approval,” 2024.

[10] K. V. Babu, “Credit Risk Assessment of Consumer Loans in India Using Machine Learning Techniques,” MSc Thesis, National College of Ireland, 2024.