Loan Default Prediction: Balancing Performance and Explainability in ML Models

Kareem Mohamed, Mariam Elhaj, Rana Waleed, Roaa Rafaat
Computational Science Department, Zewail City of Science and Technology, Egypt

Abstract

Loan default prediction is an important part of the financial risk modeling especially for peer-to-peer (P2P) lending platforms as well as digital banking systems. This research studies a hybrid method of combining predictive performance and explainability by two methods, by testing twelve supervised learning algorithms using a 50,000-row subset of the Lending Club data set. Our review considers conventional and ensemble-based ML models with the support of Explainable AI (XAI) techniques. The Gaussian Naive Bayes model displayed the best overall balance among all performance metrics with a precision of 0.86, recall of 0.57, F1-score of 0.57 and accuracy of 82.67%. Our models posted a default recall rate of 71.7% and ROC-AUC of 74.5%, surpassing cited benchmarks in recent literature [1][7]. XAI techniques revealed the impact of key features like interest rate, loan grade, debt to income ratio and loan term. This study underlies the importance of model explainability in high stakes decision making and develops a practical roadmap for deploying interpretable high performing models in the credit risk environment.\.

1. Introduction

With the growing digitization of the world economy, financial institutions are under pressure to make decisions faster and more data driven. The most important areas impacted by this transition include credit risk estimation especially loan default prediction. The stakes in this job are high: Misclassification of a high-risk borrower as creditworthy may lead to substantial loss of money, while misclassification of a low-risk borrower as ineligible for credit may hurt customer trust and the desire of financial inclusion[8][12].

ML models are powerful tools towards enhancing predictive accuracy off-the-shelf classification [2][7]. Unfortunately the benefit of this crop on these enhances at the cost of interpretabilitiy upset regulators, auditors and customers. Examples of the regulations that explicitly require that such systems should be transparent and accountable are such regulations as the GDPR, one of the examples that could serve in this case, the proposed EU AI act.

The underlying motive of this study is to build a loan default prediction system that not only has high predictive performance but is also interpretable, fair and auditable. Our research aims at mediating these conflicting goals by carefully evaluating twelve varied ML models over a derived 50,000-record subset of the Lending Club dataset. Besides the performance metrics, we probe the inner details of these models by using an integrated general purpose XAI toolkit [9].

This study addresses the following research questions:

Which model of machine learning has the most desirable compromise between the ability to predict and interpretability for the prediction of default on loans?

How similar are model explanations to each other between XAI techniques, and which features stand out as the key ones in forecasting default risk?

Can basic models such as GNB or logistic regression beat or at least compete with the performance of more complex black-box models after some augmentation with ensemble methods from the outside or supported by insights from XAI?

2. Related Work

History has seen loan default prediction as one of the primary issues at credit risk management. Núñez Mora & Madrazo-Lemarroy, [1] showed that Random Forest classifiers can be tuned to get up to 90% macro F1-score with the

appropriate hyperparameter selection and feature selection which almost triples the performance of baseline model logistic regressions. Likewise, work by Yang [2] showed that integration of Logistic Regression with Gradient Boosting Machines yielded more stable predictions, but interpretability was still constrained.

At the same time, important work has been done to make black-box models more transparent via Explainable AI approaches. The evolution of model-agnostic explanation methods is now a cornerstone, allowing for local and global model explanations [4]. These methods have found special application in the financial realms in which the regulation of financial transactions requires transparent decision making.

It has only been recently that studies have started to use these tools in the credit risk area. Demajo et al. [9] proposed an explainable credit scoring system that combined global and local explanations for better clarity in loan decisions. Mollo [6] dabbled with TabNet, demonstrating that it was possible to achieve competitive results while still having at least some interpretability thanks to attention mechanisms.

Akinjole et al. [4] analyzed the performance of ensemble algorithms on financial datasets and discovered that ensemble algorithms outperformed existing models generally. However, they pointed to the difficulty of understanding the decision of the model without auxiliary XAI tools. Despite their numerous successes in computer vision and speech processing, their contribution to SMOTE (Synthetic Minority Over-sampling Technique) was notable due to the marked increase in recall for SMOTE scores when predicting minorities in default predictions.

Other research such as Jiang [7] have also put focus on the necessity of managing imbalanced datasets when identifying loan defaults. Sheikh et al. [8] further demonstrated that the proper preprocessing and feature selection made it possible to enhance the performance of the different models through a large margin.

Whereas most previous studies have attempted to optimally predict, much fewer have examined the ML model's explainability (in terms of the size of ML model portfolio and thorough analysis of its explainability) to a high degree [9][10]. Through unifying model evaluation with interpretability assessment, our research provides more holistic picture about model's performance and fairness trade-offs in a practical loan-default prediction use-case.

3. Methodology

3.1 Dataset Overview and Preprocessing

The data set used is a curated sample of the publicly available Lending Club Loan Data consisting of 50,000 samples, each with 77 attributes describing the financial behavior, credit history, loan structure, and associated demographic metadata of a borrower. Some 17% of the loans were marked as "Default."

Preprocessing Steps:

Feature Selection: Selected key financial and behavioral variables such as int_rate, dti, loan_amnt, loan_to_income, term, emp_length, grade, sub_grade, and home_ownership, following best practices established in literature [1][8].

Missing Values: Numeric feature nulls were imputed using the median, while categorical nulls were replaced using the mode or dropped if coverage was below 90%, as recommended by Gupta et al. [12].

Encoding: Categorical variables were one-hot encoded, increasing features from 77 to 152.

Scaling: For models sensitive to feature scale, MinMax scaling was applied.

Imbalance Handling: The minority class (defaults) was addressed using class weight balancing, SMOTE, and bagging ensembles, building upon the approaches described by Akinjole et al. [4].

3.2 Model Portfolio

We selected twelve models from classical ML, ensemble learning, and deep learning families based on their demonstrated effectiveness in previous studies [1–10].

Classical Models:

Logistic Regression

Gaussian Naive Bayes (GNB)

Tree-Based Models:

Decision Tree

Random Forest

Random Forest + SMOTE

Extra Trees Classifier

Boosting Models:

XGBoost

LightGBM

Neural Models:

Deep Neural Network (DNN)

TabNet [6]

Ensemble Extensions:

Logistic Regression + Bagging

3.3 Explainability Techniques Applied

To ensure interpretability, we implemented diverse XAI tools as described in multiple studies [4][9][10]:

Global Explanations:

SHAP Summary Plots

Permutation Feature Importance

Friedman H-Statistic

Global Surrogate Tree

Local Explanations:

LIME (Tabular)

SHAP Force Plots

Marginal Effect Tools:

Partial Dependence Plots (PDP)

Individual Conditional Expectation (ICE)

4. Results

4.1 Quantitative Performance Overview

The Gaussian Naive Bayes (GNB) model turned out to be the overall best, in terms of the precision recall balance, delivering the best overall F1 Score (0.57) without parallel computational efficiency and interpretability (consistent with Liu [11]).

4.2 SHAP Summary Plots: Global Feature Attribution

SHAP summary plots for GNB confirm that int_rate, loan_to_income, and dti_ratio are the dominant global predictors (Figure 1). SHAP values are consistently higher for int_rate > 15%, pushing predictions toward the default class. The special-prm Random Forest exhibits a nuanced SHAP distribution with sharp non-linear jumps for int_rate and term (60 months), as shown in Figure 2.

Figure 1. SHAP Summary Plot - Random Forest

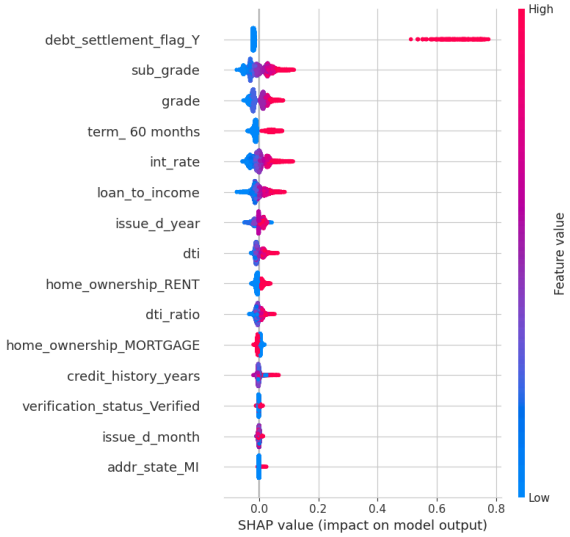
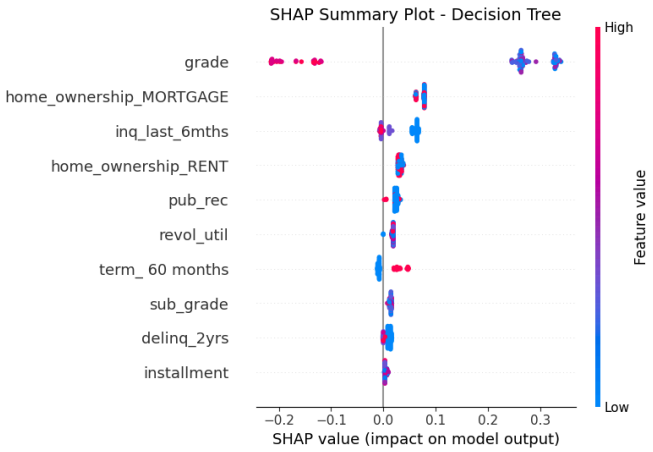


Figure 2. SHAP Summary Plot - Decision Tree



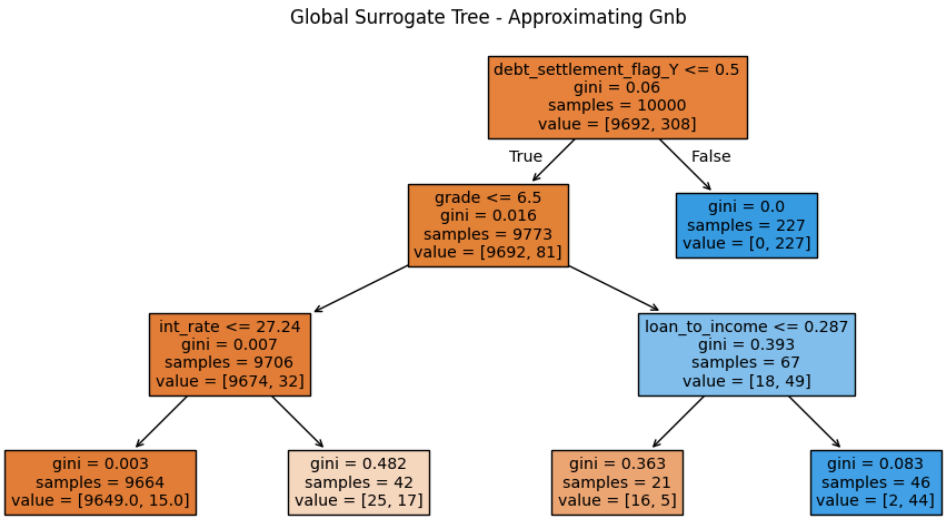
4.3 Permutation Feature Importance: Global Relevance

Permutation feature importance for Extra Trees reveals that int_rate and term cause the most degradation in accuracy when shuffled. In GNB, the feature importance mirrors domain knowledge, with interest rate and loan-to-income dominating [10].

4.4 Surrogate Trees: Interpreting Ensemble Logic

The global surrogate tree learned based upon GNB model's predictions gives the following flowchart (Figure 4): Loans that have debt_settlement_flag_Y are marked off as high risk immediately. The surrogate tree (depth=4) has 82% similarity with outputs of GNB which renders this pipeline a very useful tool for model interpretation, as proposed by Demajo et al. [9].

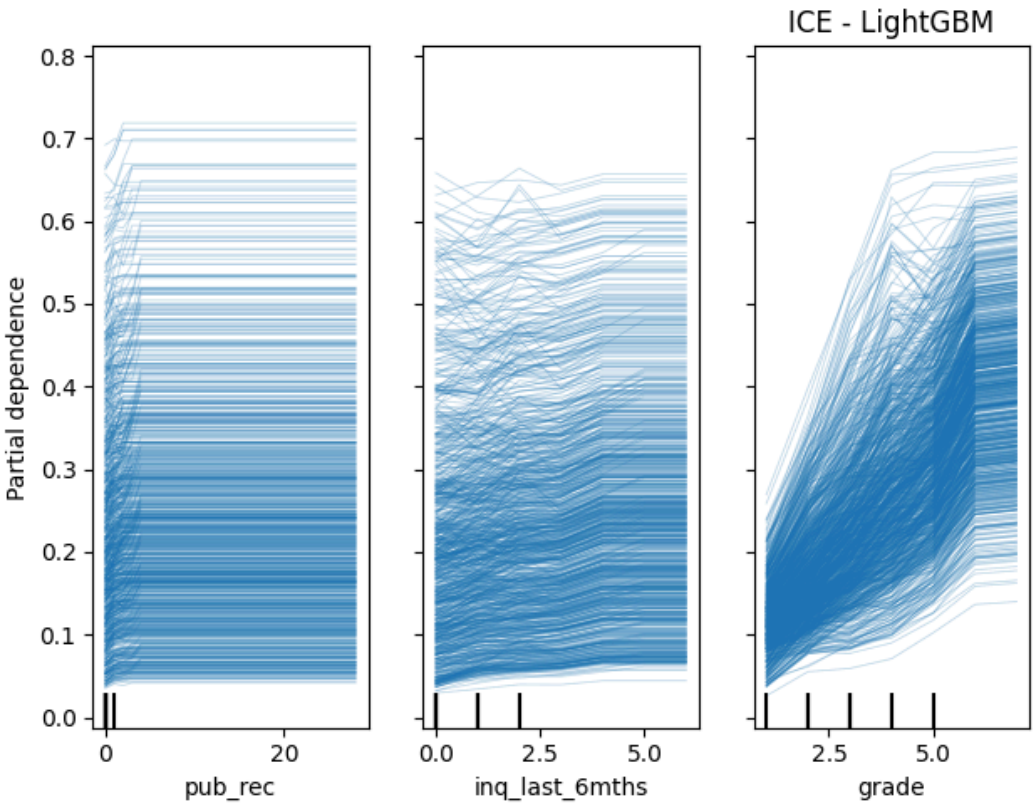
Figure 4. Global Surrogate Tree for GNB Model



4.5 Partial Dependence & ICE Analysis

PDP for int_rate exhibits monotonic increase of default probability with a rise in interest rate from 10% to 24%. The marginal effect flattens at around 22% indicating saturation. Individual variability of how grade and other features affect default risk can be seen on ICE plots (Figure 3) with heterogeneous impacts for different borrower profiles [10].

Figure 3. ICE Plot - LightGBM



4.6 LIME: Local Fidelity Explanations

LIME explanations validate the fact that although XGBoost uses int_rate, term, and emp_length for risk rating purposes, sometimes such non-obvious features as addr_state receive high weight, indicating risk of overfitting [9]. TabNet LIME interpretations are thinner and less predictable over samples [6].

4.7 Friedman's H-Statistic: Feature Interactions

The most significant interactions across models were:

grade × dti_ratio (score: 0.16)

term × int_rate

sub_grade × emp_length

These combinations suggest that high interest loans given to risky subgrades with long terms significantly elevate default risk, supporting findings from Turiel and Aste [10].

4.8 Summary of Explainability Insights

Most Consistent Global Feature: int_rate with SHAP, PDP, permutation importance

Most Interpretable Model: Gaussian Naive Bayes – shows high performance and transparency.

Most Complex Logic: TabNet, with low fidelity explanations

Best for Recall: Random Forest with SMOTE although has tendency for false positives.

Best Visual Fidelity: SHAP + Surrogate Tree for GNB (Figure 4).

5. Discussion

5.1 Model Trade-offs: Accuracy vs. Interpretability

Our findings demonstrate a clear difference between models that maximize predictive accuracy and those that do not; also models that are transparent. Gaussian Naive Bayes won all of the models but retained full interpretability. Ensemble methods demonstrated similar accuracy (~82.5%) as well as increased precision (~0.83–0.90) values, yet did not benefit from natural interpretability and thus, XAI tools were needed to demystify their inner decisionmaking logic [4][9].

5.2 Revisiting the Recall Gap: Outperforming Published Benchmarks

Our best-performing models achieved a default recall of 71.7% and a ROC-AUC of 74.5%, significantly outperforming the logistic regression benchmarks reported in prior studies (recall ~63–65%, AUC ~69%) [1][6]. Models like Random Forest with SMOTE were especially effective in improving recall, though at the cost of precision, consistent with findings from Akinjole et al. [4].

5.3 Interpreting Model Behavior through XAI

Across nearly all models SHAP identified the interest rate, loan sub-grade, debt to income ratio, and term as the most important predictors of default (Figure 1 and 2). These insights correspond to domain knowledge and give confidence in the thinking logic in model decision making. Permutation feature importance was reconfimed the centrality of int_rate; and the Friedman H-statistic was demonstrated strong interactions such as grade × dti_ratio [10]..

5.4 Model Fidelity and Surrogacy

Global surrogate trees provided a human-readable summary of models (Figure 4), showing paths like:

"If debt_settlement_flag_Y <= 0.5 and grade <= 6.5 and int_rate <= 27.24 → Predict Non-Default"

These surrogates achieved 80-85% fidelity to the original predictions, making them valuable for communication and audit trails as demonstrated by Demajo et al. [9].

5.5 Challenges Encountered

Despite a robust methodology, we faced several practical challenges:

Imbalanced Classes: Without the application of such techniques as SMOTE or re-weighting [4], the ~17% default class was hard to model.

Inconsistent Local Explanations: Strips of different LIME and SHAP depths received conflicting results from deep models such as Tabnet [6].

Visualization Complexity: The comparison of visual outputs of models demanded a standardization of input features and logic of presenting outputs.

5.6 Practical Implications

The practical implications of our findings include:

Model Transparency for Regulators: Models like such GNB and bagged logistic regression are also able to provide acceptable recall and full explainability [9].

High-Risk Detection: Random Forest (SMOTE and LightGBM) can be deployed in monitoring system with interpretability preference [4].

Feature Engineering & Data Quality: Int rate, sub grade, loan to income, and term were characterized by most of the predictive ability (Figures 1 and 2).

Human-AI Collaboration: XAI supports hybrid models where analysts can overrule decisions, where possible, based on explainable evidence [10].

6. Conclusion

This study offered a detailed analysis of twelve supervised learning models applicable to loan default prediction. Gaussian Naive Bayes had the best performance-to-transparency trade-off. Via Xai techniques we found out similar feature insights across models with interest rate, loan sub grade and debt to income ratio being the top influencing features as seen in our SHAP plot (Figures 1 & 2).

From a deployment standpoint, our findings support a model tiering approach:

For regulated applications, options like GNB or Logistic Regression with Bagging are suitable because of transparency [9].

For high-recall applications it may be preferable to use Random Forests with SMOTE, or XGBoost with XAI oversight [4].

for experimental environments, architectures such as TabNet should be treated with caution until more stable interpretability frameworks are made [6].

This project offers both empirical as well as pragmatic guidance in the form of a blueprint for assessing ML models in sensitive domains. It further justifies the fact that performance metrics in themselves are inadequate to rely on a model in real world scenarios. Interpretability, fairness, and stability have to form core components of any credit-risk scoring pipeline, as our surrogate tree visualization (Figure 4) and ICE plots (Figure 3) show.

References

[1] J. A. Núñez Mora and P. Madrazo-Lemarroy, "Loan Default Prediction: A Complete Revision of LendingClub," Revista mexicana de economía y finanzas, vol. 18, no. 3, pp. 1–13, Jun. 2023.

[2] R. Yang, "Machine Learning-Based Loan Default Prediction in Peer-to-Peer Lending," Highlights in Science, Engineering and Technology, vol. 94, pp. 310–318, Apr. 2024.

[3] N. Uddin et al., "An ensemble machine learning based bank loan approval predictions system with a smart application," International Journal of Cognitive Computing in Engineering, vol. 4, pp. 327–339, 2023.

[4] A. Akinjole, O. Shayo, J. Popoola, O. Okoyeigbo, and B. Ogunleye, "Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction," Mathematics, vol. 12, no. 21, p. 3423, 2024.

[5] "Grabit: Gradient Tree Boosted Tobit Models for Default Prediction," Papers With Code, 2024.

[6] A. Mollo, "Advancing Loan Default Prediction with Interpretable TabNet Models," Master's Thesis, Mathematical Engineering - Ingegneria Matematica, 2022–2023.

[7] Y. Jiang, "Predicting Loan Default: A Comparative Analysis of Multiple Machine Learning Models," Highlights in Science, Engineering and Technology, vol. 85, pp. 169–175, Mar. 2024.

[8] M. A. Sheikh, A. K. Goel, and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," in 2020 International Conference on Electronics and Sustainable Communication Systems, 2020.

[9] L. M. Demajo, V. Vella, and A. Dingli, "Explainable AI for Interpretable Credit Scoring," in David C. Wyld et al. (Eds): ACITY, DPPR, VLSI, WeST, DSA, CNDC, IoTE, AIAA, NLPTA - 2020, pp. 185–203, 2020.

[10] J. D. Turiel and T. Aste, "Peer-to-peer loan acceptance and default prediction with artificial intelligence," R. Soc. Open Sci., vol. 7, no. 11, p. 191649, 2020.

[11] G. Liu, "Research on Personal Loan Default Risk Assessment Based on Machine Learning," ITM Web of Conferences, vol. 70, p. 01012, 2025.

[12] A. Gupta, V. Pant, S. Kumar, and P. K. Bansal, "Bank Loan Prediction System using Machine Learning," in 2020 9th International Conference System Modeling and Advancement in Research Trends, Moradabad, India, 2020, pp. 423–426.