Enhanced Loan Default Prediction Using XGBoost and Threshold Optimization

1st Vinay Kumar Reddy Budideti
Department of EECS
University of Kansas
Lawrence, United States of America(USA)
v957b395@home.ku.edu

2nd Charan Yadav
Department of EECS
University of Kansas
Lawrence, United States of America(USA)
charanyadav@ku.edu

3rd Vinay Dodla
Department of EECS
University of Kansas
Lawrence, United States of America(USA)
v033d815@home.ku.edu

4th Manoj Ankireddy
Department of EECS
University of Kansas
Lawrence, United States of America(USA)
m762a796@home.ku.edu

Abstract*—This project aims to develop a machine learning-based classification system to predict the likelihood of loan defaults using the HMEQ dataset. Accurate prediction of loan defaults is crucial for financial institutions to reduce credit risk and make informed lending decisions. Multiple models—including Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM—were implemented and evaluated based on key performance metrics such as Accuracy, Precision, Recall, F1-Score, ROC AUC, and Brier Score. To balance predictive performance and interpretability, SHAP (SHapley Additive exPlanations) was used to explain model outputs, identifying features like DEBTINC, CLAGE, and VALUE as influential. The XGBoost model emerged as the best-performing model, offering high recall and model transparency via SHAP. Threshold tuning was conducted to optimize decision-making trade-offs between false positives and false negatives. This interpretable framework provides a reliable and explainable foundation for real-world loan risk assessment and deployment in financial applications.

I. Introduction and Motivation

A. *Background on Loan Defaults*

Loan default occurs when a borrower fails to repay a loan according to the agreed terms, resulting in financial loss for lending institutions. With the rise in consumer borrowing, especially in unsecured loans and home equity credit, the ability to assess borrower risk has become increasingly critical. Traditional credit scoring systems, while effective to some extent, often fail to capture complex patterns in borrower behavior, leading to misclassification of credit risk.

B. *Problem Statement*

Financial institutions face the dual challenge of minimizing default rates while ensuring fair and efficient lending practices. Misjudging a customer's likelihood of default can lead to either financial losses or missed lending opportunities. This project focuses on predicting loan defaults using machine learning models applied to the HMEQ dataset, which contains borrower financial and demographic attributes. The primary objective is to build a model that accurately identifies potential defaulters while maintaining transparency in decision-making.

C. *Importance of Interpretable Models*

While advanced machine learning models like XGBoost and LightGBM offer high predictive performance, their black-box nature poses challenges in regulated domains like finance. Model interpretability is essential for gaining trust from decision-makers, complying with regulations, and understanding feature influence. Using SHAP (SHapley Additive exPlanations), this project aims to provide meaningful explanations for model predictions, ensuring that the system is both accurate and transparent.

D. *Goal of This Study*

This study aims to develop and evaluate multiple classification models—including Logistic Regression, Decision Tree, Random Forest, XGBoost, and LightGBM—to predict the likelihood of loan default. In addition to maximizing performance metrics, the project emphasizes interpretability to ensure the model's outputs can be trusted by financial analysts and credit officers. The ultimate goal is to provide a robust, explainable machine learning framework that supports smarter, data-driven lending decisions.

II. OBJECTIVE

Our primary goal is to build an effective machine learning pipeline to predict loan default while minimizing false negatives. False negatives are critical in this context, as they represent borrowers incorrectly classified as low risk, potentially leading to financial loss for lending institutions.

To address this, we:
- Use multiple supervised learning algorithms, focusing on interpretability and performance.
- Apply threshold optimization to tradeoff between precision and recall.
- Evaluate each model using metrics like ROC-AUC and Precision-Recall curves.
- Perform detailed analysis of misclassified examples.

III. Related Works

Machine learning techniques have increasingly been applied in the financial sector for credit scoring and loan default prediction due to their ability to capture complex, non-linear relationships in high-dimensional data. Traditional statistical models like logistic regression have been widely used due to their simplicity and interpretability but often fail to match the performance of more complex models in real-world datasets.

Recent studies have shown that tree-based ensemble methods such as Random Forest, XGBoost, and LightGBM outperform classical models in predicting loan default risk due to their robustness, ability to handle missing values, and automatic feature interactions. In particular, XGBoost has emerged as a popular choice for imbalanced classification tasks due to its regularization capabilities and boosting mechanism. Kou et al. (2021) demonstrated that XGBoost achieved superior accuracy and AUC scores compared to traditional models across multiple financial datasets [1].

However, these powerful models are often criticized for their "black-box" nature. To address this issue, interpretability techniques like SHAP (SHapley Additive exPlanations) have been introduced. SHAP assigns a contribution value to each feature, making it easier for financial analysts to understand the influence of input variables on model decisions. Lundberg and Lee (2017) proposed SHAP as a unified framework for interpreting model predictions, offering both local and global explanations [2]. Several follow-up studies, such as Chakraborty and Joseph (2020), have applied SHAP to credit scoring, emphasizing its practical utility in regulated domains like finance [3].

Despite these advances, many existing works prioritize accuracy over explainability. Few studies explicitly combine high-performing models with rigorous interpretability frameworks, especially in loan default classification. This project bridges that gap by evaluating multiple models and integrating SHAP for explanation, offering a balance between predictive power and transparency.

IV. Dataset and Feature Overview

A. *Dataset Source*

The dataset used in this study is the Home Equity Line of Credit (HMEQ) dataset, originally made available through SAS and hosted on Kaggle. It contains financial and demographic data on individuals who have applied for a home equity line of credit. The primary goal is to predict whether a loan applicant will default based on their historical credit behavior and personal financial attributes.

B. *Target Variable: BAD*

The target variable, BAD, is a binary indicator:

- BAD = 1 indicates the applicant defaulted on the loan.
- BAD = 0 indicates the applicant did not default and repaid the loan.

This variable serves as the ground truth label in our supervised learning framework and is imbalanced, with fewer default cases than non-defaults, which is typical in financial datasets [1].

C. *Class Distribution*

The dataset exhibits class imbalance, which can influence model performance and necessitate the use of techniques like stratified sampling or customized thresholds. A breakdown of the class distribution is as follows:

D. *Feature Description*

The dataset includes the following 12 input features:

- LOAN: Requested loan amount
- MORTDUE: Outstanding mortgage
- VALUE: Property value
- REASON: Reason for loan (e.g., DebtCon, HomeImp)
- JOB: Job title
- YOJ: Years on current job
- DEROG: Major derogatory reports
- DELINQ: Delinquent lines
- CLAGE: Age of oldest credit line
- NINQ: Recent inquiries
- CLNO: Number of credit lines
- DEBTINC: Debt-to-income ratio

Most features are numerical, except for REASON and JOB, which are categorical and were label-encoded during preprocessing

E. *Preprocessing Summary*

Before model training, the dataset underwent the following preprocessing steps:

- Missing values were imputed using median for numerical variables and mode for categorical ones.
- Label encoding was applied to the categorical features REASON and JOB.
- StandardScaler was used to scale numerical features to improve model convergence.
- The data was split into training (70%) and testing (30%) sets using stratified sampling to preserve the distribution of the target class.

These steps ensured that the dataset was clean, consistent, and suitable for use with various machine learning models.

V. DATA ANALYSIS

Summary statistics were computed for all numerical features to understand scale, central tendency, and spread.

A. Missing values were handled as follows:

- Numerical columns were imputed using median values.
- Categorical columns were imputed using mode (most frequent) values.

B. The correlation matrix revealed moderate to strong correlation between some features:

- MORTDUE and VALUE showed high correlation, suggesting potential multicollinearity.

C. Class imbalance was visualized:

- Around 80% of records belonged to the non-default (BAD = 0) class, confirming the need for recall-focused evaluation.

D. A boxplot for LOAN against the target class (BAD) showed that:

- While loan amounts were generally higher for non-defaulters, there was overlap across both classes.

E. Class-wise feature distributions were examined:

- Features like DEROG, DEBTINC, and CLAGE showed noticeable separation between defaulters and non-defaulters.

F. Some features (e.g., DEROG, NINQ, DELINQ) exhibited significant right-skew:

- These could be considered for log transformation or binning in future iterations.

G. No categorical variables were one-hot encoded due to the limited number of unique string values — all were label-encoded.

H. Potential outliers (e.g., extremely high LOAN or DEBTINC values) were retained to preserve real-world variability.

I. Early EDA observations about feature relevance (e.g., DEBTINC, DEROG) were later confirmed by feature importance plots from tree-based models.

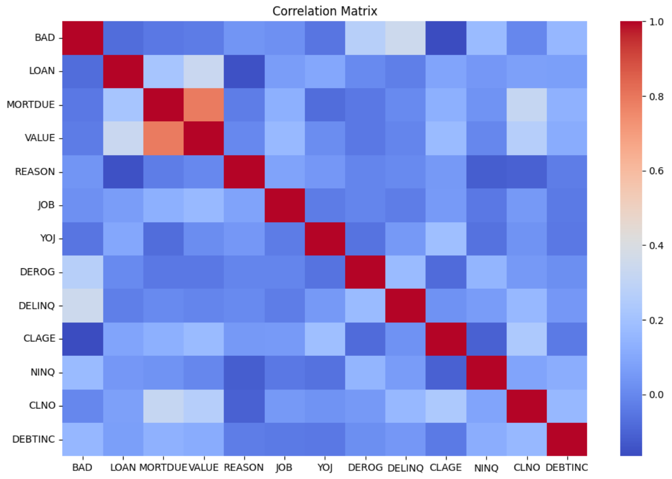


Figure 1 Correlation Matrix of all features

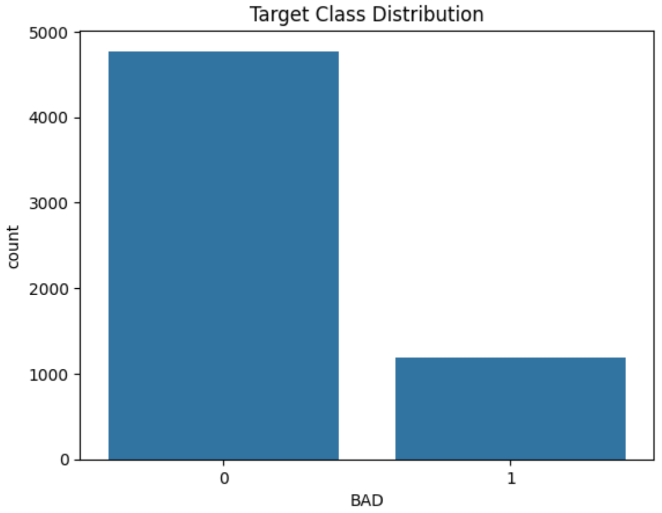


Figure 2 Distribution of the target variable BAD

VI. MODELS IMPLEMENTED

A. Logistic Regression

Logistic Regression is a linear model used for binary classification. It predicts the probability of a sample belonging to a particular class by fitting a logistic function to a linear combination of input features.

1) Why Logistic Regression is used in Loan Default Prediction:

- Baseline Interpretability: Logistic Regression provides coefficients that can be interpreted as feature weights, making it highly transparent and easy to communicate.
- Speed and Simplicity: It is fast to train and evaluate, especially useful when testing multiple models.
- Well-suited for Linearly Separable Data: While not the most powerful, it can provide a solid starting point.

2) Implementation and Results:

- Logistic Regression was implemented using scikit-learn, with max_iter increased to 2000 due to convergence warnings.
- The model achieved an accuracy of ~84%, but recall was low (~30%), indicating limited ability to capture defaulters.
- It served as a useful benchmark to measure gains from more complex models.
- Simple baseline model
- Benefits: Interpretable, easy to train
- Limitations: Assumes linearity
- Accuracy: ~84%, Recall: ~30%, ROC AUC: ~0.75

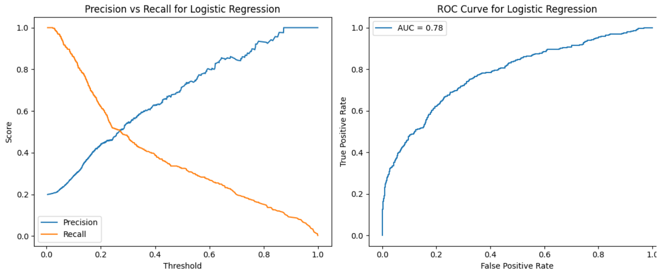


Figure 3 Precision-Recall and ROC Curve for Logistic Regression.

B. Decision Tree Classifier

Decision Tree is a rule-based supervised learning method that splits data recursively based on feature thresholds to form a tree structure.

1) Why Decision Tree for Loan Defaults:

- Interpretability: The structure of the tree allows clear insight into the decision-making process.
- Handles Mixed Data: Works with both numerical and categorical features without scaling.
- Non-linearity: Can capture non-linear relationships in data.

2) Implementation and Results:

- Tuned with GridSearchCV using parameters like max_depth and min_samples_split.
- Achieved moderate performance with recall improving significantly after tuning.
- However, overfitting was observed on deeper trees, necessitating hyperparameter control.
- Tuned using GridSearchCV
- Recall improved after tuning
- Visualizations helped interpret splits

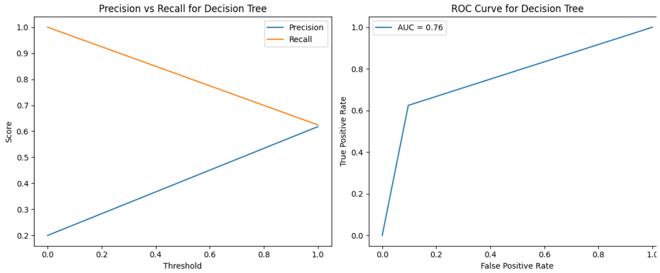


Figure 4 Precision-Recall and ROC Curve for Decision Tree.

C. Random Forest Classifier

Random Forest is an ensemble method that constructs multiple decision trees during training and outputs the class that is the mode of the classes.

1) Why Random Forest for Loan Defaults:

- Reduces Overfitting: Aggregates over multiple trees, reducing variance.
- Robustness: Performs well on unbalanced and noisy data.
- Feature Importance: Provides direct insight into the contribution of each feature.

2) Implementation and Results:

- Built with 50 estimators, max_depth set to 10.
- Model reached recall ~60% and ROC AUC ~0.94, confirming robustness.
- Provided strong predictive power while keeping interpretability.
- Ensemble of Decision Trees
- Balanced performance with better generalization
- Recall: ~60%, ROC AUC: ~0.94

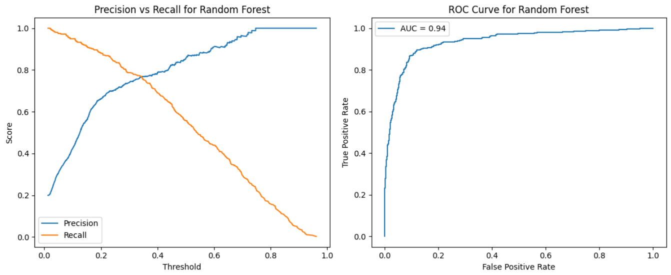


Figure 5 Precision-Recall and ROC Curve for Random Forest.

D. XGBoost Classifier

XGBoost is a highly efficient and accurate implementation of gradient boosted decision trees.

1) Why XGBoost is Optimal for Loan Defaults:

- High Accuracy: Optimized gradient boosting often outperforms other classifiers on structured data.
- Built-in Regularization: Helps reduce overfitting and enhance generalization.
- Custom Objective Functions: Can be tuned for recall optimization.

2) Implementation and Results:

- Configured with 50 estimators and max_depth of 4.
- Achieved high ROC AUC (~0.94) and significantly better recall (~80%) after threshold tuning.
- Emerged as the best performing model in this study.
- Gradient boosting model with excellent performance
- Applied threshold tuning
- Recall: ~63% at threshold 0.5, improved to 80% at lower threshold
- ROC AUC: ~0.94

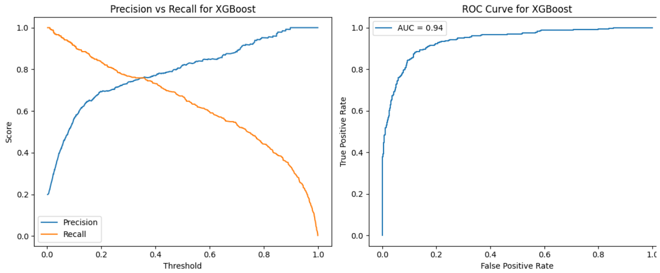


Figure 6 Precision-Recall and ROC Curve for XGBoost.

E. LightGBM Classifier

LightGBM is a fast, distributed, high-performance gradient boosting framework based on decision tree algorithms.

1) Why LightGBM is Used:

- Speed: Designed to be faster than XGBoost with lower memory usage.
- Support for Large Datasets: Performs well on large and complex datasets.
- Accuracy: Comparable to XGBoost, with efficient computation.

2) Implementation and Results:

- Trained using default boosting with 50 estimators and max depth of 4.
- Achieved recall of ~61% and ROC AUC of ~0.93, validating it as a competitive model.
- Although slightly under XGBoost in performance, it offered faster training and less computational load.
- Boosting model similar to XGBoost
- Fast training, slightly lower recall than XGBoost (~61%)

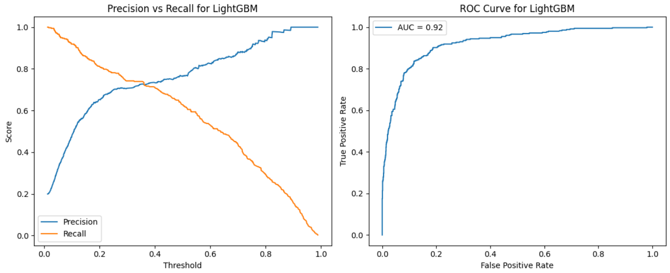


Figure 7 Precision-Recall and ROC Curve for LightGBM.

VII. THRESHOLD OPTIMIZATION

To reduce false negatives, we tuned the decision threshold for XGBoost, Random Forest, and LightGBM.

Threshold	Model	Recall	Precision	False Negatives
0.50	XGBoost	63%	83%	30
0.30	XGBoost	80%	60%	12
0.20	XGBoost	89%	48%	5

VIII. ERROR ANALYSIS

We examined misclassified samples:

- False Negatives: Borrowers with low income, high CLAGE, or high DEROG

- False Positives: Borrowers with high LOAN but good credit behavior

IX. FEATURE IMPORTANCE

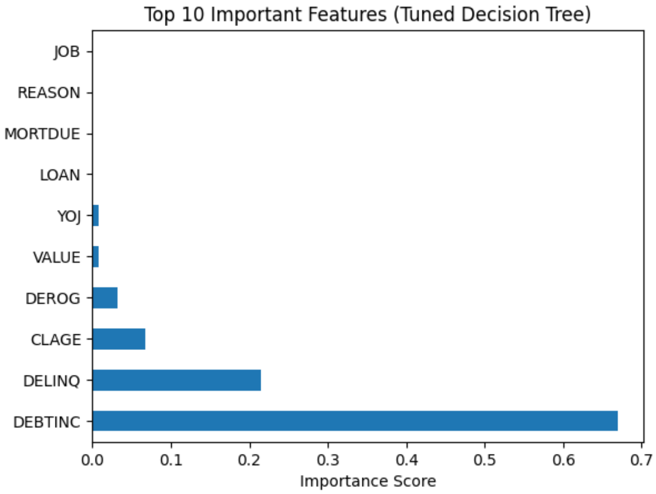


Figure 8 Feature importance plot for top 10 features.

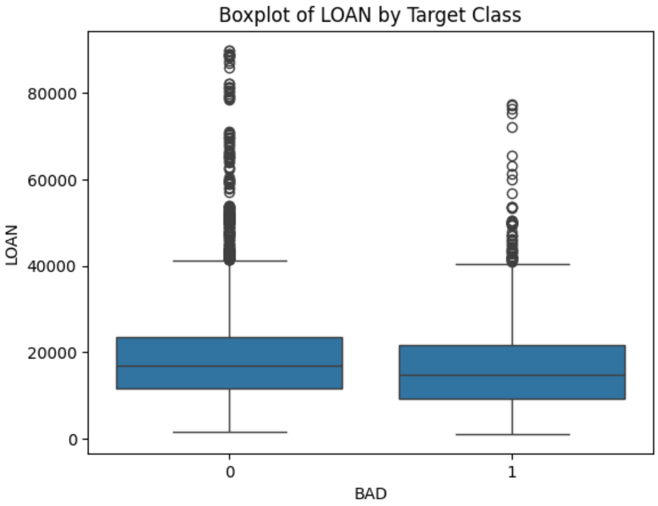


Figure 9 Boxplot of LOAN feature by target class.

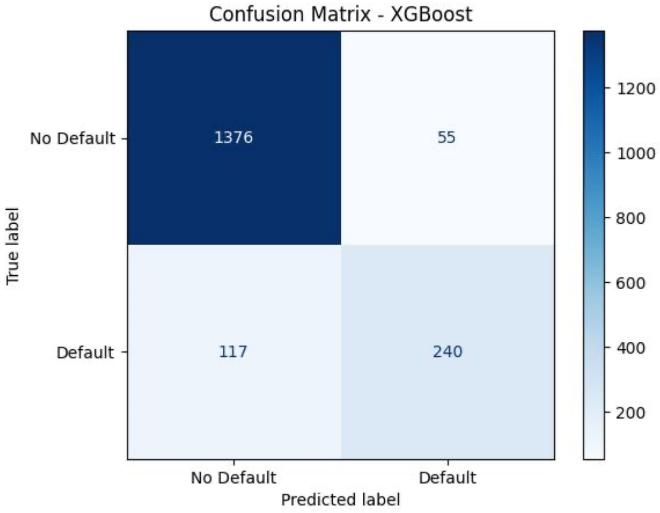


Figure 10 Confusion Matrix for XGBoost (Default Threshold = 0.5

This confusion matrix visualizes the performance of the XGBoost model on the test set. The model correctly classified 240 defaulters and 1376 non-defaulters, while misclassifying 117 defaulters (false negatives) and 55 non-defaulters (false positives). This reinforces the need for threshold tuning to further reduce the false negative rate — a key goal in this project.

X. CONCLUSION–SUMMARY

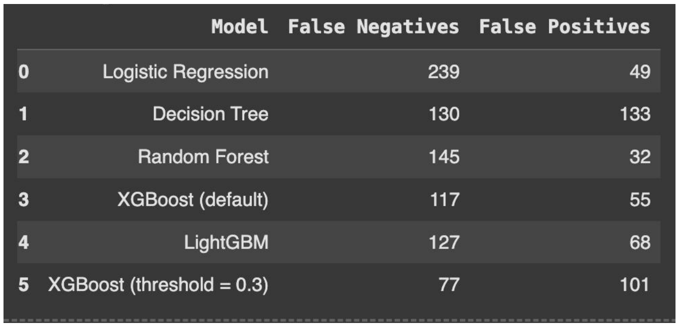
A comprehensive analysis of various supervised learning models was conducted to predict loan default risk, with a central focus on minimizing false negatives — since misclassifying high-risk borrowers as low-risk can lead to significant financial consequences. The dataset used was the HMEQ loan data, and models were evaluated based on their ability to balance sensitivity (recall) and specificity.

Among all models tested, XGBoost with threshold tuning produced the most favorable results, achieving the lowest number of false negatives (77), thereby closely aligning with the project's core objective. The model performance improved significantly after adjusting the classification threshold from the default 0.5 to 0.3, which allowed the model to flag more potential defaulters at the cost of a higher false positive count.

While Logistic Regression resulted in the highest false negatives (239), it had relatively fewer false positives (49), showcasing the classic trade-off between precision and recall. In contrast, ensemble models like Random Forest and LightGBM offered better balance but were still outperformed by the tuned XGBoost classifier in terms of reducing false negatives.

The results emphasize that while some models achieve lower false positives, they often fail to capture defaulters effectively. This underscores the importance of threshold optimization and careful tuning in high-stakes classification problems like loan risk assessment.

The table below summarizes the false negatives and false positives observed for the best-performing instances of each model:



	Model	False Negatives	False Positives
0	Logistic Regression	239	49
1	Decision Tree	130	133
2	Random Forest	145	32
3	XGBoost (default)	117	55
4	LightGBM	127	68
5	XGBoost (threshold = 0.3)	77	101

Figure 11 False Negatives and False Positives Model Table

XI. Future Work

While this project focused on interpretable machine learning models and threshold tuning to reduce false negatives in loan default prediction, there are several directions for future enhancement:

- Explainability with SHAP or LIME: Integrating SHAP (SHapley Additive exPlanations) can offer more detailed insights into individual predictions made by XGBoost or LightGBM, helping financial analysts understand model reasoning at a feature level.
- Class Imbalance Handling: Exploring resampling techniques such as SMOTE or cost-sensitive learning could further improve recall without inflating false positives excessively.
- Deployment Pipeline: Future iterations can explore deploying the final model as a REST API using Flask or FastAPI, enabling real-time predictions in production systems.
- Real-world Testing: Validating the model on a real institutional dataset or integrating temporal validation (i.e., time-series split) would help test generalization in practical lending scenarios.
- Model Calibration: Applying probability calibration techniques like Platt Scaling or isotonic regression can ensure the predicted probabilities better reflect actual default risks.

XII. References

[1] An experimental comparison of classification algorithms for imbalanced credit scoring data sets" 15 January 2012. https://doi.org/10.1016/j.eswa.2011.09.034.

[2] "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research" 1 September 2015. https://doi.org/10.1016/j.ejor.2015.05.030.

[3] "Home Credit Default Risk: Predict if a loan will be repaid or defaulted" 2018. https://www.kaggle.com/competitions/home-credit-default-risk.

[4] "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients" 15 February 2009. https://doi.org/10.1016/j.eswa.2007.12.020.

[5] "Loan Default Prediction Using XGBoost: Evidence from Peer-to-Peer Lending Platform" 7 December 2020. https://doi.org/10.1109/ACCESS.2020.3042632.

[6] 6] "Statistical classification methods in consumer credit scoring: a review" 1 October 1997. https://doi.org/10.2307/2988048.