A LOAN DEFAULT PREDICTION MODEL USING MACHINE LEARNING

Shubham Aher[1] , Gaurav Jadhav[2] , Dr. Priya Shelke[3] , Prof. Amol Dhumane[4]

[1-4]Assistant Professor , Department of CSE(IoTCSBT), Vishwakarma Institute of Information Technology, Pune

Email: shubham.22310391@viit.ac.in , gaurav.22310069@viit.ac.in

Keywords: loan default, Ml, Random forest, XGBoost.

Abstract

Loan defaults pose significant financial risks for lending institutions. To mitigate such risks, accurate borrower credit assessments are crucial before loan approval. This study explores how machine learning can enhance loan credit evaluation using a Kaggle dataset. We evaluate four classification models — Logistic Regression, Random Forest, XGBoost, and Decision Tree — based on their predictive accuracy. Among them, XGBoost achieved the highest performance, while Logistic Regression lagged behind. Our results suggest that incorporating advanced ensemble techniques like XGBoost into credit risk assessment can substantially improve decision-making for financial organizations.

1. Introduction

In the financial domain, credit risk is a persistent issue associated with loans, credit cards, and mortgages. Borrowers may fail to fully repay their loans, posing significant challenges for lenders. Accurately evaluating the creditworthiness of applicants is vital for the sustainability and profitability of financial institutions, especially in a competitive environment. Sophisticated assessment techniques are required to identify reliable borrowers. Despite these efforts, defaults remain common, leading to substantial financial losses for lending organizations [1]. The integration of artificial intelligence (AI) and big data analytics offers powerful tools to mitigate these risks. Unlike traditional credit scoring methods, which often rely on expert judgment and are limited in scalability, machine learning (ML) techniques provide more accurate and data-driven alternatives for credit evaluation [2]. By analyzing personal, financial, and historical data of borrowers, AI enables institutions to make better-informed lending decisions. The use of ML in credit risk assessment enhances prediction accuracy and operational efficiency, thereby reducing default rates and improving cash flow stability.

Historically, credit analysis involved manual evaluation based on limited datasets, which proved time consuming and less accurate. In contrast, ML models can process large-scale datasets rapidly, uncovering complex patterns that traditional methods may overlook [3]. As a result, ML is increasingly used to automate credit scoring and risk prediction, reducing operational costs and improving consistency in decisions.For instance, Kumar et al. evaluated models such as Decision Tree, Random Forest, SVM, K-Nearest Neighbors, and AdaBoost, showing promising results [4]. Similarly, Singh et al. applied XGBoost, Random Forest, and Decision Tree classifiers to public-sector loan data, with Random Forest achieving an accuracy of 91.7%, along with high sensitivity and specificity [5]. Although progress has been made, there remains significant potential for further enhancement in predictive performance. This research employs Logistic Regression, Random Forest, XGBoost, and Decision Tree models to predict loan defaults using a Kaggle dataset. Special attention is given to feature selection, as the choice of input variables significantly influences model accuracy. A poorly selected feature set may obscure valuable insights and reduce prediction quality [6]. Moreover, model performance can vary with different feature combinations, highlighting the need for careful engineering. This study aims to contribute to the ongoing development of ML-driven credit assessment systems, supporting financial stability and reducing institutional losses.

2. Literature Survey

The integration of machine learning into the financial sector has transformed how credit risk is assessed. A large body of literature has investigated various models and frameworks for predicting loan defaults, focusing on improving accuracy, interpretability, and scalability of predictions. This section reviews key contributions categorized into five thematic areas. Early explorations emphasized understanding the patterns within loan approval datasets and laying the groundwork for predictive analytics. Pandit conducted one of the foundational studies by applying data mining techniques to identify defaulters, highlighting the importance of pre-modelling analysis and data quality. Zhang and Yang proposed a machine learning model tailored for personal loan default prediction and demonstrated its effectiveness using a real-world financial supervision dataset.

As machine learning evolved, researchers focused on deploying these models in operational banking environments. introduced a data-driven credit risk management framework that integrated ML throughout the credit lifecycle, from application to repayment. Kumar et al. implemented multiple algorithms such as SVM and KNN to predict loan eligibility in real-time banking workflows. Singh et al. developed a modernized loan approval system that blends traditional rule-based checks with predictive machine learning models to improve loan disbursement timelines.

Feature selection and ensemble models have received attention for their performance on structured financial data. Hegde et al. [7] demonstrated that engineering variables like income-to-loan ratios and applying correlation analysis significantly improved classifier performance.Zhu et al. [8] used a Random Forest algorithm and validated its effectiveness in default prediction tasks on large datasets. Gupta et al. [14] showed that combining decision trees with robust preprocessing pipelines enhances prediction stability. Boosting methods such as AdaBoost and XGBoost have emerged as leaders in classification accuracy for imbalanced financial data.Chen and Guestrin [9] introduced XGBoost, which improved execution speed, memory optimization, and regularization for loan default tasks.Cao et al. [10] conducted a theoretical and empirical analysis of AdaBoost, concluding its strong potential in binary classification problems with skewed data.Shaheen and ElFakharany [11] tested various algorithms and confirmed that decision-tree-based ensembles outperform naive classifiers in predictive accuracy and precision. Recent research emphasizes financial impact and interpretability alongside predictive performance. Zhang et al. [12] proposed a profit-driven model that incorporates business objectives directly into the loss function, enhancing the relevance of predictions for lenders. Ahmed and Rajaleximi [17] focused on model transparency and the importance of regulatory compliant scorecards in ML implementations.Kohv and Lukason [15] categorized predictor variables into behavioral, macroeconomic, and demographic domains to study their differential impacts on loan default risks.Fati [13] and Aphale & Shinde [16] emphasized the growing relevance of ML in cooperative banks and underbanked sectors, especially for automating loan approvals.

3. Proposed system

To develop a machine learning-based predictive system capable of classifying whether a loan applicant is likely to default, using financial and demographic data. The system is designed to support financial institutions in making informed lending decisions by minimizing the risk of bad loans. This project adopts a data-driven approach to predict loan default outcomes using a labeled dataset of loan applications. The objective is to train models capable of classifying each application as either a default (1) or a non-default (0). Given the nature of financial data, the dataset exhibits class imbalance, with significantly fewer default cases. To address this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to enhance model performance, particularly on the minority class. A comparative analysis of multiple machine learning algorithms was conducted to determine the most effective model. The models evaluated include Logistic Regression (used as a baseline), Random Forest Classifier, Decision Tree Classifier, and XGBoost Classifier. These models were trained and assessed using standard performance metrics such as Accuracy, Precision, Recall, F1 Score, ROC-AUC, and the Confusion Matrix, providing a well-rounded view of each model's capabilities. To enhance predictive power, feature engineering was applied, including the derivation of new variables such as the Loan_to_Income_Ratio, which captures the financial burden of the loan relative to the borrower's income. The overall modeling pipeline was designed with scalability and generalizability in mind, ensuring that the system can be extended or integrated into future automated decision making platforms in financial services.

4. System Design and Architecture Diagram

The proposed system follows a modular and sequential pipeline that begins with raw data ingestion and ends with performance evaluation of multiple predictive models. The design ensures that every stage — from preprocessing to model training — is independently manageable, reproducible, and upgradable. The system architecture comprises the following components: Data Acquisition - The dataset (Loan_default.csv) is sourced from Kaggle. It contains borrower-level data including demographics, income, loan amount, employment status, and the target variable: Default. Data Preprocessing - During preprocessing, the non informative identifier "LoanID" was dropped to clean the dataset. Categorical variables were transformed using One-Hot Encoding to make them suitable for machine learning algorithms. Additionally, a new feature called "Loan_to_Income_Ratio" was engineered to capture the relative burden of the loan on the borrower's income. Handling Class Imbalance - To address class imbalance, SMOTE (Synthetic Minority Over-sampling Technique) was applied. This technique helps in balancing the distribution between default and non default classes, which is especially important since financial datasets typically contain fewer default cases. Feature Scaling - Numerical features were scaled using StandardScaler. This ensures that all features contribute equally in distance-based models and gradient-based calculations, thereby improving model performance and convergence. Model Training- After preprocessing and balancing, the data was split into training and testing sets using a 70:30 ratio. Four classifiers were then trained and tested: Logistic Regression, Random Forest, Decision Tree, and XGBoost. Model Evaluation - The trained models were evaluated on the test set using several metrics, including Accuracy, Precision, Recall, F1 Score, ROC-AUC Score, and the Confusion Matrix. These metrics provided a comprehensive understanding of each model's performance.

5. Experimental Details Implementation of Algorithms Used

This section outlines the implementation of the machine learning models used for predicting loan default. Each algorithm was chosen based on its suitability for classification problems and its proven performance in financial risk modeling.

5.1 Dataset Overview

The dataset used in this study is sourced from Kaggle and is titled the "Loan Default" dataset. It contains approximately 255,346 rows, each representing an individual borrower's financial and demographic details. The target feature is Default, a binary variable where 0 indicates no default and 1 indicates default.

The dataset includes several attributes such as age, income, employment type, loan amount, loan tenure, and past repayment history. After loading the data into a pandas DataFrame, the column LoanID was removed, as it served only as an identifier and held no predictive value for modeling purposes.

5.2 Preprocessing Pipeline

To ensure data quality and compatibility with machine learning algorithms, a structured preprocessing pipeline was implemented.

Null Value Treatment: Missing values were handled appropriately—either by dropping rows with excessive nulls or by applying suitable imputation techniques to preserve data integrity.

Categorical Feature Encoding: Categorical variables were encoded to convert them into a machine-readable format. Binary categorical features (such as gender) were transformed using Label Encoding, while features with multiple categories (such as employment type) were processed using One-Hot Encoding.

Feature Scaling: Numerical features like income, age, and loan amount were scaled using StandardScaler. This normalization step ensures that all features contribute proportionately during training and enhances model convergence, especially for gradient-based algorithms.

5.3 Machine Learning Algorithms

Logistic Regression
A statistical linear model used as a baseline. It models the probability of default using the logistic function. Its coefficients provide interpretability into how each feature affects the likelihood of default.

Decision Tree Classifier
A rule-based classifier that splits data based on entropy or Gini impurity. Although prone to overfitting, it's valuable for understanding decision boundaries. At each node, it selects a feature xix_ixi and threshold ttt that best splits the data using a criterion like: Gini Impurity , Entropy.

Random Forest Classifier
An ensemble of decision trees trained on bootstrap samples with random feature selection. This reduces overfitting and improves generalization.

XGBoost Classifier
An advanced gradient boosting algorithm known for handling imbalanced data and non-linear interactions. Regularization terms help avoid overfitting.

6. Data Splitting and Validation Strategy

The dataset was split into training and testing sets using a 70:30 ratio, where 70% of the data was allocated for training the models and 30% for evaluating their performance. To maintain the original distribution of the target classes (default and non-default) in both subsets, stratified sampling was employed during the split. This is crucial in imbalanced classification problems to prevent biased model learning.

For hyperparameter tuning and performance validation, cross-validation techniques were applied. Specifically, Grid Search combined with K-Fold Cross-Validation (with k set to 5) was used to identify the best parameter configurations for the models. This step, while optional in the notebook, contributes significantly to optimizing model accuracy and generalizability.

All models were trained on the same dataset and evaluated using consistent metrics to ensure fairness in comparison.

7. Performance Metrics Summary

Below is a tabular comparison based on the results obtained

Model	**Acc**	**Precision**	**Recall**	**F1-Score**	**ROCAUC**
Logistic Regression	**0.89**	0.87	0.91	0.89	0.89
Random Forest	0.91	**0.90**	0.91	0.91	0.92
Decision Tree	0.88	0.86	0.89	0.87	0.89
XGBoost	**0.92**	**0.91**	**0.93**	**0.92**	**0.94**

1. True Positives (TP): 58134 (correctly predicted defaults)

2. True Negatives (TN): 66658 (correctly predicted non defaults)
3. False Positives (FP): 85 (incorrectly predicted defaults)
4. False Negatives (FN): 1015 (missed defaults)

8. Discussion

Among the models evaluated, XGBoost emerged as the best-performing classifier. It achieved the highest scores across all evaluation metrics, demonstrating its robustness and effectiveness for imbalanced binary classification problems. Its ability to handle complex feature interactions and focus on difficult-to-classify examples contributed to its superior performance.

The Random Forest model followed closely behind, delivering strong precision and generalizability. Its ensemble nature helped reduce variance and capture a wide range of patterns in the data, making it a reliable choice for practical applications.

Despite its simplicity, Logistic Regression proved to be a competitive model, especially in terms of the F1 score. This performance can be attributed to effective class balancing and feature scaling, which helped the linear model better separate the two classes. Lastly, the Decision Tree model offered the highest level of interpretability. While its overall performance was slightly lower, it achieved a decent recall, making it a valuable option in scenarios where transparency and explainability of decisions are critical—such as in financial risk assessment or regulatory environments.

These results confirm the suitability of ensemble models — particularly boosting techniques — for financial risk modeling in structured datasets.

9. Conclusion

The objective of this research was to build a predictive system to identify loan default risks using a set of supervised machine learning models. Starting from a real-world dataset obtained from Kaggle, we constructed a full ML pipeline comprising data cleaning, feature engineering, class imbalance correction via SMOTE, model training, and detailed evaluation.

The dataset exhibited significant class imbalance, which was effectively resolved using SMOTE, resulting in a balanced training ground for all models. Feature engineering — particularly the creation of a Loan-to Income Ratio — contributed to better model learning by providing a more interpretable relationship between a borrower's financial obligation and income.

Among all models tested, XGBoost outperformed others in accuracy (92%), precision (91%), recall (93%), F1-score (92%), and ROC-AUC (94%). This is attributed to XGBoost's ability to capture complex non-linear relationships and its robust handling of overfitting via regularization.

Random Forest also demonstrated strong performance and could serve as an effective alternative when computational resources are constrained.

Practical Implications

Deploying such a system in real-world financial institutions can significantly reduce loan default risks, streamline loan approval processes, and enhance overall decision-making by providing real-time predictive insights.

Future Work

Explainability: Integrate SHAP or LIME to explain model decisions to non-technical stakeholders. Real-Time Scoring API: Deploy the model in production using tools like Flask or FastAPI for live loan applications. Extended Fe atures: Include behavioral and transactional data for improved prediction (e.g., credit card usage, bank balances). Model Monitoring: Establish continuous monitoring to detect model drift and update models as patterns change over time.

10. References

[1] Ashish Pandit, Data Mining on Loan Approved Dataset for Predicting Defaulters, M.Sc. Thesis, 2016.

[2] Zhang Liying, Yang Ruoji, "Application of ML in Loan Default Prediction", Financial Supervision Research, 2022.

[3] Mingrui Chen et al., "Data-Driven Credit Risk Management Process", ACM Proc. Int. Conf. Software and System Process, 2017.

[4] C. Naveen Kumar et al., "Customer Loan Eligibility Prediction", ICCES, 2022.

[5] Vishal Singh et al., "Modernized Loan Approval System", CONIT, 2021.

[6] Pratheeksha Hegde N et al., "Predictive Analysis of Loan Data", AIDE, 2022.

[7]Zhu, L., et al. (2019). *Loan default prediction using random forest*, Procedia CS.

[8] Lin Zhu et al., "Predicting Loan Default using Random Forest", Procedia Computer Science, 2019.

[9] Chen T, Guestrin C., "XGBoost: A Scalable Tree Boosting System", ACM SIGKDD, 2018.

[10] Ying Cao et al., "Research on AdaBoost Algorithm", Acta Automatica Sinica, 2013.

[11] S.K. Shaheen, E. ElFakharany, "Predictive Analytics for Loan Default", ICCTA, 2018.

[12] Lifang Zhang et al., "Profit-Driven Loan Default Prediction", Expert Systems with Applications, 2023.

[13] S. M. Fati, "ML-Based Loan Status Prediction", Journal of Hunan University Natural Sciences, 2021.

[14] A. Gupta et al., "Bank Loan Prediction System", SMART Conf., 2020.

[15] K. Kohv, O. Lukason, "Best Predictors of Loan Defaults", Risks Journal, 2021.

[16] A. S. Aphale, S. R. Shinde, "Loan Approval in Cooperative Banks", IJERT, 2020.

[17] M. I. Ahmed, P. R. Rajaleximi, "Credit Scoring in Financial Institutions", IJARCET, 2019.