Loan Approval Prediction

Om G. Patel
Arizona State University
Tempe, Arizona
opatel14@asu.edu
ASU Id : 1233379110

Suraj Rao
Arizona State University
Tempe, Arizona
surajrav@asu.edu
ASU Id : 1233990435

Darren A. Ferrao
Arizona State University
Tempe, Arizona
dferrao@asu.edu
ASU Id : 1237189124

Mohammad A.K.Dalwai
Arizona State University
Tempe, Arizona
mdalwai@asu.edu
ASU Id : 1237671684

Abstract—This project aims to develop a robust machine learning pipeline for loan approval prediction using a dataset containing more than 58,000 samples and 13 original features. The primary goal is to classify applicants as Approved or Default, thereby reducing credit risk for financial institutions while also supporting automated decision-making and improving fairness in lending practices. There was a significant class imbalance between approved and default target classes. To address this problem, multiple resampling techniques such as SMOTE and SMOTETomek were applied to improve minority class representation. A wide range of machine learning models were evaluated, including Logistic Regression, Decision Trees, Random Forests, SVM, KNN, Naive Bayes, and advanced ensemble methods such as XGBoost, LightGBM, Gradient Boosting, and AdaBoost. Models were assessed using accuracy, precision, recall, F1-score, and ROC-AUC. Ensemble models consistently outperformed the others, with a tuned XGBoost classifier achieving 95.17% accuracy, a 0.812 F1-score, and a ROC-AUC of 0.953, demonstrating strong predictive capability and stability. Overall, the developed pipeline represents a reliable and scalable approach to loan risk assessment, effectively combining balanced data, engineered features, and optimized ensemble models to make precise predictions.

I. Introduction

Identifying the risks involved in granting loans to applicants is a critical task for financial institutions, as manual assessments are time-consuming and often introduce some level of human bias. Introducing AI and ML systems to evaluate applicants based on their data can accelerate the process and provide more efficient evaluations by identifying common applicant patterns. The specific problem addressed in this project is a binary classification task that distinguishes between approved and default applicants. The main challenges in this problem include complex feature relationships and significant class imbalance. Accurate loan approval prediction models help financial institutions optimize resource usage, speed up the application process, and maintain stable lending portfolios.

Previous work has used machine learning models for credit scoring and loan default prediction using behavioral and transactional financial data [1], and XGBoost models have also been applied for default prediction in conventional banking credit scoring [2]. Ensemble methods combining multiple algorithms have shown superior performance in handling imbalanced datasets [3], and recent studies have demonstrated that SMOTE-based approaches significantly improve minority-class prediction [4].

The system proposed in this project is a data mining pipeline that incorporates data preprocessing, feature engineering, class balancing, model training, and evaluation. In this architecture, raw data is transformed into meaningful variables, with resampling techniques applied to address imbalance. The dataset used in this project comes from a Kaggle

Playground series and contains approximately 58,000 samples with demographic attributes, employment information, credit history, and more. The data mining pipeline consists of data cleaning, encoding, standardization, feature engineering, class balancing, modeling, and evaluation. Incorrect decisions can lead to substantial financial loss. Traditional Model evaluation results show that conventional models struggled with low recall for default cases, whereas tree-based methods performed better. Ensemble models produced the significant results overall, with the tuned XGBoost model achieving 95.17 percent accuracy, a 0.812 F1-score, and an ROC-AUC of 0.953.

II. Problem Formulation

The aim of this project is to determine whether a loan applicant is likely to repay a loan or default, using patterns learned from past applicant information. Since there are only two possible outcomes, the task naturally becomes a binary classification problem. The model is trained on historical lending decisions and then used to score new applicants based on patterns it has learned.

A. Key Definitions

Input Features: The dataset includes a mix of demographic information, such as age and years of employment, along with financial variables (including income, loan amount, and interest rate), and indicators of previous credit behavior. To better represent a borrower's financial stress, several additional features were engineered, including ratios such as income-to-loan and debt-to-income, as well as metrics like loan burden.

- Target Variable: The loan outcome is represented by the field loan_status, where:

  0 = Approved

  1 = Default

- Dataset Source: The data comes from the Kaggle Playground Series competition on loan approval. It contains 58,645 samples with 13 original features and 12 additional engineered features.

B. Problem Description

The goal is to use historic loan outcomes to estimate the probability that a new applicant may default. Incorrectly approving risky applicants can result in financial losses for lenders, while rejecting applicants who are actually safe may lead to lost business opportunities. Because of this, the model must strike a careful balance between minimizing false approvals and not being overly conservative.

C. Challenges

Several challenges arise when modeling this dataset:

- Severe class imbalance: About 86% of records are Approved and only 14% are Default. This can lead most models to simply predict every sample as Approved and still achieve high accuracy, despite being useless in practice.
- Complex relationships between financial variables, where simple linear models struggle to capture risk patterns.
- Fairness and explainability considerations, since real world lending must justify predictions.

D. Project Objective

The main goal is to create a fair model that improves recall for default cases. This means reducing the number of risky applicants who are wrongly labeled as safe. We also want to keep high overall accuracy and stability across evaluation metrics, such as precision, F1-score, and ROC-AUC.

E. Assumptions

- The records provided in the dataset are representative of real loan decision patterns.
- Applicants' financial and credit details are factual and reflect real world lending conditions.

- The historical patterns present in the dataset generalize to future applicants.

III. Overview of the Proposed System

The proposed system is a full data mining and machine learning pipeline. It is designed to process raw loan application data and make predictions about whether a loan should be approved or is likely to default. The pipeline includes data cleaning, preprocessing, feature engineering, class balancing, model training, evaluation, and optimization.

The process begins with collecting and cleaning raw data. Then, we transform the dataset into a format that works for machine learning. Financial datasets often have missing information, inconsistent formats, uneven numerical ranges, preprocessing was necessary to ensure quality inputs for modeling. Numerical features were standardized to reduce scaling differences, and categorical variables were label encoded to allow compatibility with machine learning algorithms.

To capture financial risk more realistically, additional engineered features were introduced. Examples include income-to-loan ratio, loan burden, debt-to-income ratio, and high interest loan indicator. These features capture risk dynamics that influence repayment behavior and significantly improved model performance.

A major system component addresses the challenge of class imbalance since only about 14% of records represent default cases. To prevent models from becoming biased toward the majority class, two resampling techniques (SMOTE oversampling and SMOTETomek hybrid sampling) were applied. These methods balanced the dataset and led to substantial improvements in recall for default cases.

The processed and balanced data was then used to train and evaluate a wide range of predictive models, including traditional machine learning algorithms (Logistic Regression, K-Nearest Neighbors, Naive Bayes, Support Vector Machine), tree-based models (Decision Tree, Random Forest), and advanced ensemble methods (Gradient Boosting, LightGBM, XGBoost, AdaBoost). Model performance was assessed using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrices.

Finally, the system includes hyperparameter tuning to optimize the best performing model (XGBoost). It also follows with feature importance and interpretability analyses. These steps make sure that the chosen model performs reliably, generalizes well, and remains defendable for real-world lending decisions.

Overall, the proposed system offers a scalable and data-driven method for automating loan approval decisions. It combines preprocessing, feature engineering, class balancing, and ensemble learning to achieve strong predictive performance that is suitable for use in financial settings.

IV. Technical Details

A. Feature Engineering

To prepare the dataset for machine learning, we applied several preprocessing steps and transformations to improve data quality and model performance. The original dataset had 13 main features, including demographic attributes (person_age, person_home_ownership), financial indicators (person_income, loan_amnt, loan_percent_income, loan_int_rate), and credit-related information (cb_person_cred_hist_length, cb_person_default_on_file).

The first step in preprocessing was to remove the unused id field. We converted categorical variables into numeric form using Label Encoding to ensure compatibility with the algorithms. We standardized the numerical features using StandardScaler to lessen the impact of scale differences between income level features and smaller numeric values like credit history length.

Beyond basic preprocessing, extensive feature engineering was performed to capture complex financial conditions and improve prediction capability. Two ratio based features were introduced in Preliminary Analysis and Baseline Models income_to_loan_ratio and age_employment_ratio which provided more meaningful insight into an applicant's affordability and career stability. In Enhanced Modeling and Final Model Selection, 12 additional engineered features were created to represent high risk profile characteristics more explicitly. Examples include debt_to_income_ratio, loan_burden (interaction between loan cost and income pressure), high_interest_loan (binary indicator for interest rates above median), and high_risk_combination (flagging applicants with both low loan grade and history of default). Log transformations were applied to highly skewed variables (person_income, loan_amnt), and polynomial terms were added to capture non linear credit behavior patterns.[5]

After final processing, the dataset grew from 13 original features to 25 total features. This change contributed significantly to the performance improvements seen in later experiments. The feature importance analysis in Enhanced Modeling and Final Model Selection showed that many engineered features ranked as the top contributors, confirming their inclusion.

B. Predictive Modeling

Model development followed an incremental approach starting with baseline classifiers and progressing to advanced ensemble methods. The dataset was split using an 80/20 stratified train test strategy to preserve the class distribution. Initial experiments in Preliminary Analysis and Baseline Models evaluated three base models, Logistic Regression, Decision Tree, and Random Forest. Random Forest achieved the strongest baseline performance with 95% accuracy and F1-score of 0.78, but the recall for the Default class remained comparatively low, demonstrating limited sensitivity to minority class samples due to severe class imbalance (≈ 86% Approved, 14% Default).

To tackle this problem, we used SMOTE oversampling and SMOTETomek hybrid resampling techniques in Enhanced Modeling and Final Model Selection. These methods improved the balance of the training distribution, reduced bias toward majority class predictions, and significantly boosted recall for default cases. With balanced datasets, we trained and evaluated additional models, including SVM, K-Nearest Neighbors, Naive Bayes, Gradient Boosting, AdaBoost, XGBoost [7], and LightGBM [6]. Ensemble models consistently produced the best results because they can learn complex feature interactions and handle noisy financial data well.

After comparing different datasets and metrics, XGBoost appeared to be the best model. We used GridSearchCV to fine-tune the hyperparameters by adjusting settings such as tree depth, learning rate, and the number of estimators. The optimized XGBoost model reached an accuracy of 95.17%, an F1-score of 0.812, precision of 0.913, recall of 0.731, and a ROC-AUC of 0.953. It outperformed all baseline models and achieved the project goal of improving recall while maintaining accuracy.[8]

Finally, we included interpretability using SHAP (SHapley Additive Explanations) to look at how features affect model decisions. This followed the explainability framework introduced by Lundberg and Lee [9]. The SHAP results showed that loan grade, loan percent income, interest rate level, and features based on engineered ratios had the highest predictive importance. The learning curve analysis showed stable convergence with no signs of overfitting, confirming that the model performs well in general.

V. Experimental Evaluation

A. Dataset Description

The dataset used in this project contains applicant demographic information, financial background, and loan characteristics. The target variable is loan_status, where 0 = Approved and 1 = Default. As shown in Figure 1, the dataset is highly imbalanced (majority Approved), which encourages the use of imbalance handling strategies during training and evaluation.

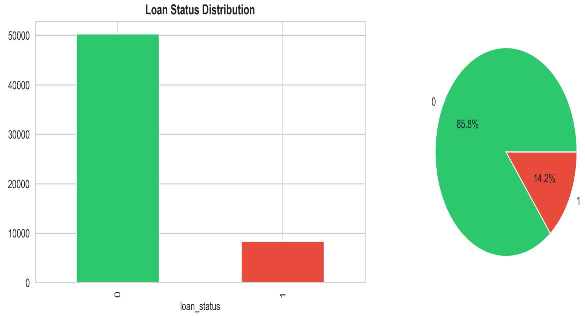

Fig.1 Loan status Distribution

The numerical features such as person_age, person_income, loan_amnt, loan_int_rate, and loan_percent_income display significant skewness and outliers. For example, both income and loan amount reach high levels, which is typical in financial data. This requires transformations and robust modeling choices.

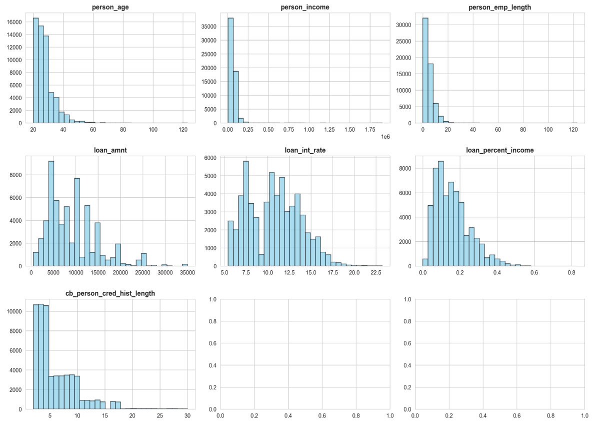

Fig.2 Numerical Distribution

The categorical variables (person_home_ownership, loan_intent, loan_grade, cb_person_default_on_file) show uneven category distributions, where some categories dominate. This impacts model learning and motivates careful encoding and feature engineering.

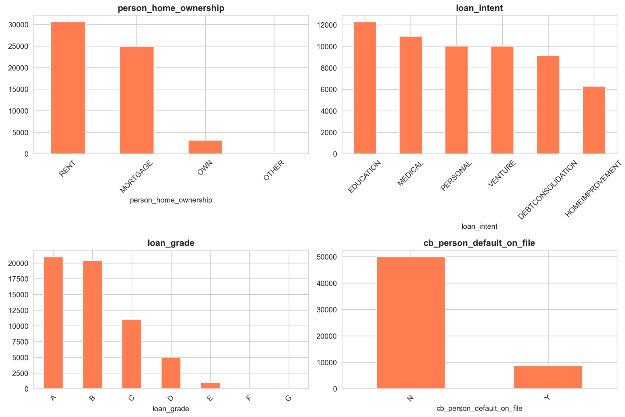

Fig.3 Categorical Distributions

We also calculated the correlation matrix to examine linear relationships. Although no single feature displays a very strong correlation with the target by itself, several variables, such as loan_int_rate, loan_percent_income, and loan_grade, show a moderate correlation. These variables serve as useful predictors when used together in non-linear models.

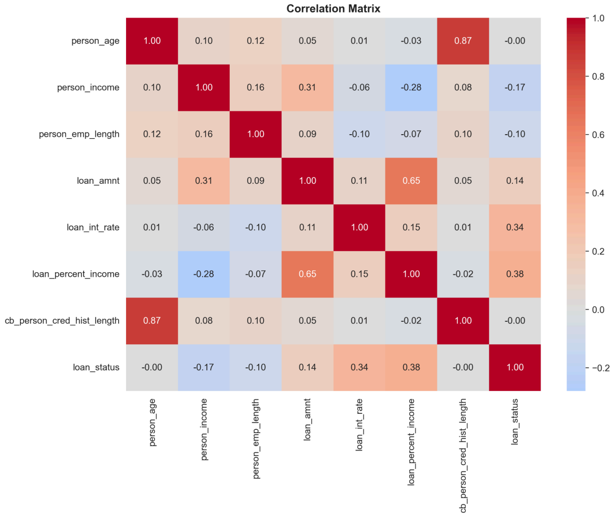

Fig. 4 Correlation Matrix

To better understand class differences, Figures 5 and 6 compare results across categories and display numerical distributions based on loan status. Defaulting borrowers generally exhibit higher loan_int_rate, higher loan_percent_income, and more risky loan_grade.

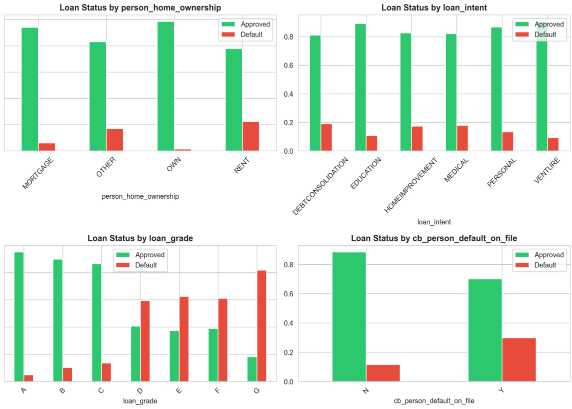

Fig. 5 Status by categories

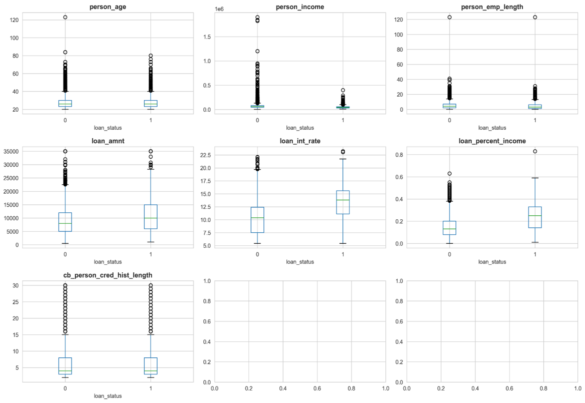

Fig. 6 Boxplot by status

Since imbalance is a major issue, Enhanced Modeling and Final Model Selection additionally evaluates SMOTE and SMOTETomek. Figure 7 shows the differences in class distribution between the original training split and the balanced versions used for modeling.

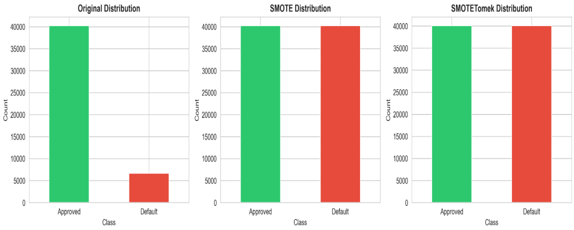

Fig. 7 Class balance comparison

B. Evaluation Metrics

Because this is a binary classification task with strong class imbalance, we evaluate models using multiple metrics:

- Accuracy: overall correctness
- Precision (Default class): among predicted defaults, how many are truly defaults
- Recall (Default class): among true defaults, how many were identified
- F1-score: balance between precision and recall
- ROC-AUC: ability to separate Approved vs Default across thresholds

Due to imbalance, recall, F1-score, and ROC-AUC are emphasized more than accuracy alone.

C . Baseline Methods

To ensure fair comparison, we define baseline models and a proposed final model.

Preliminary Analysis and Baseline Models :

- Logistic Regression is pretty simple to understand and use, which makes it a great starting point for many problems. But here's the catch—it struggles when your data doesn't follow a straight line. If the relationships in your data are more complex or curvy, Logistic Regression just won't cut it.
- Decision Tree (DT): captures interactions, but may overfit.
- Random Forest (RF): stronger non linear baseline with improved stability

Extended Comparisons (Enhanced Modeling):

In Enhanced Modeling, we expanded evaluation to include SVM, KNN, Naive Bayes, AdaBoost, Gradient Boosting, LightGBM, and XGBoost, trained on Original, SMOTE, and SMOTETomek variants.

Proposed Method (Final):

The final selected model is tuned XGBoost trained with imbalance handling and feature engineering, achieving the strongest overall tradeoff between identifying defaulters and maintaining strong overall performance.

D. Results and Analysis

1. Preliminary Analysis and Baseline Models results compare LR, DT, and RF using the same train/test split and standardized features. As expected, Random Forest provides the strongest baseline performance overall, while Logistic Regression tends to under detect default cases due to imbalance. ThePreliminary Analysis and Baseline Models comparison is summarized in Figure 8.

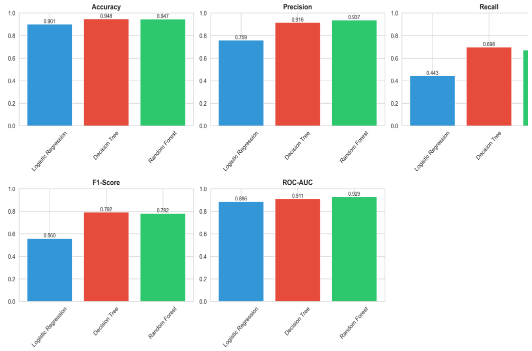

Fig. 8 Model Comparison

Confusion matrices (Figure 9) highlight error patterns. In particular, LR typically produces many false negatives for defaults (risky for lenders), DT improves default capture but can introduce more false positives, and RF provides a better balance.

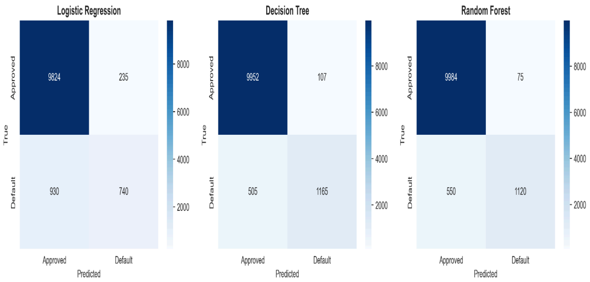

Fig. 9 Confusion matrix

Feature importance for the Preliminary Analysis and Baseline Models Random Forest (Figure 10) indicates that loan_percent_income, loan_grade, loan_int_rate, and income based ratios are among the most informative predictors.

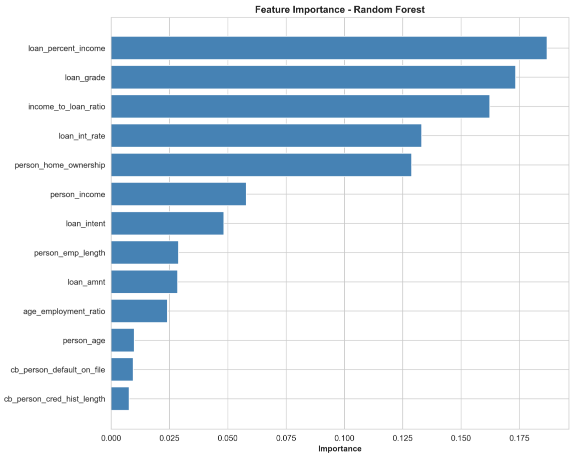

Fig. 10 Feature importance

2. Enhanced Modeling and Final Model Selection Results (Imbalance Handling + More Models + Tuning) expands the experimental evaluation in three key ways:

A. Impact of Class Balancing:

Training on SMOTE and SMOTETomek improves the model’s ability to identify defaults (higher recall), compared to training on the imbalanced original dataset. Figure 7 shows the balancing effect, and Figure 11 compares performance differences across datasets.

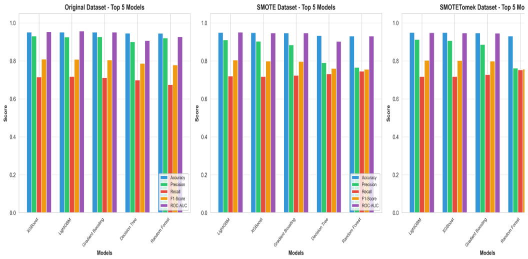

Fig. 11 Dataset comparison

B. Broader Model Comparison:

We evaluated 10 models across multiple training variants. The top performing models are consistently tree based ensembles, especially XGBoost and LightGBM, due to their ability to learn complex interactions in tabular financial data. The top results across all trained configurations are summarized in Figure 12.

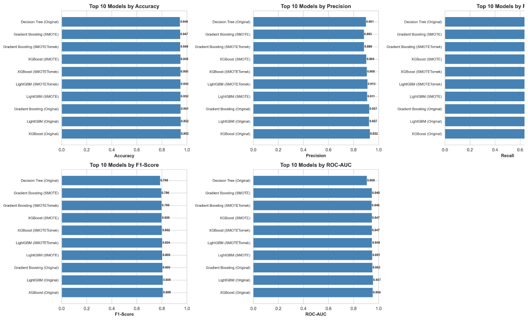

Fig. 12 Top model comparison

C. Best Model Selection + Hyperparameter Tuning:

After selecting the best candidate based on F1-score and ROC-AUC, we performed hyperparameter tuning (GridSearchCV). The final tuned XGBoost achieved is:

Accuracy: 95.17%
F1-score: 0.812
Precision: 0.913
Recall: 0.731
ROC-AUC: 0.953

3. Error Analysis: Confusion Matrices and ROC Curves (Enhanced Modeling and Final Model Selection):

To interpret the final model's classification behavior, we present confusion matrices for the top models (Figure 13) and ROC curves (Figure 14). These plots show that the selected ensemble approach provides strong separation between classes while maintaining an improved default detection rate compared to Preliminary Analysis and Baseline Models baselines.

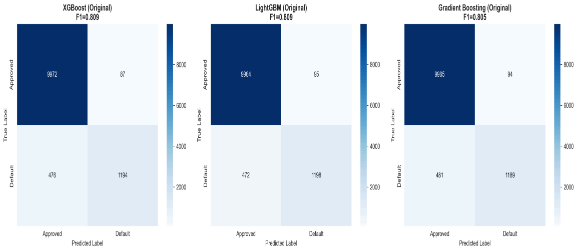

Fig. 13 Confusion matrix

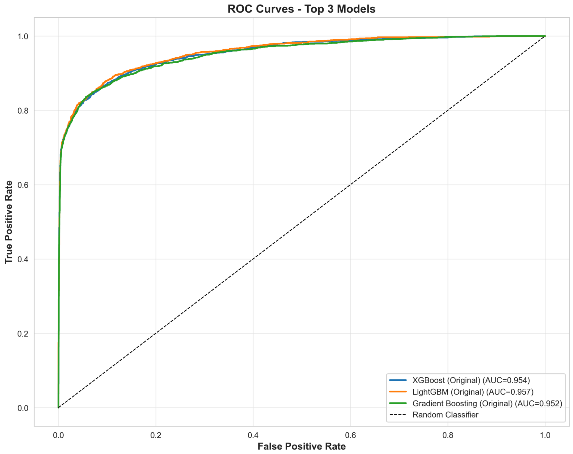

Fig. 14 ROC curves

4. Component wise Evaluation (Feature Engineering + Pipeline Gains):

Feature engineering contributed significant predictive improvements by capturing affordability and risk stress more directly (e.g., income_to_loan_ratio, debt_to_income_ratio, loan_burden, high_interest_loan). Figure 15 give a brief summary of feature importance on the best model, and Figure 16 provides interpretability using SHAP.

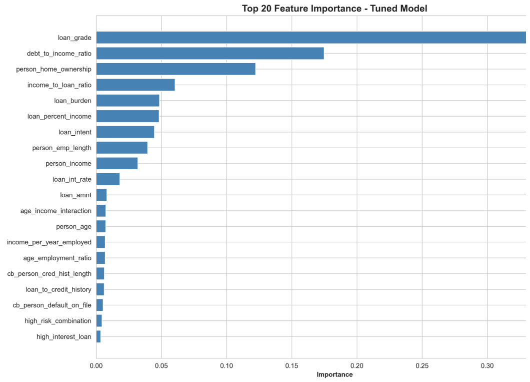

Fig. 15 Feature Importance

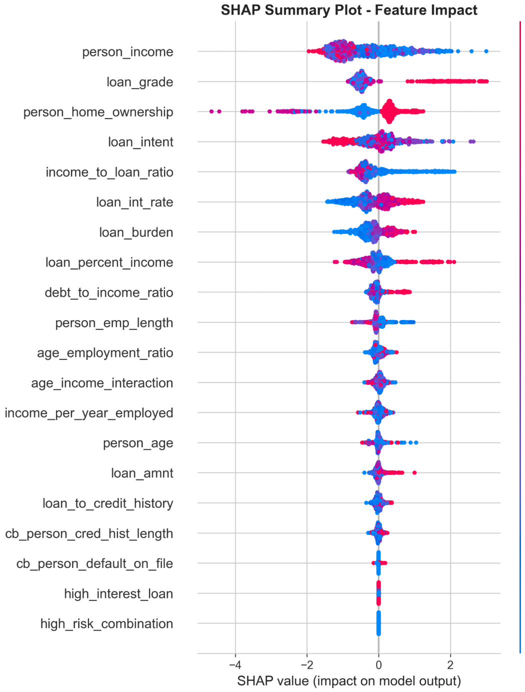

Fig. 16 SHAP summary plot

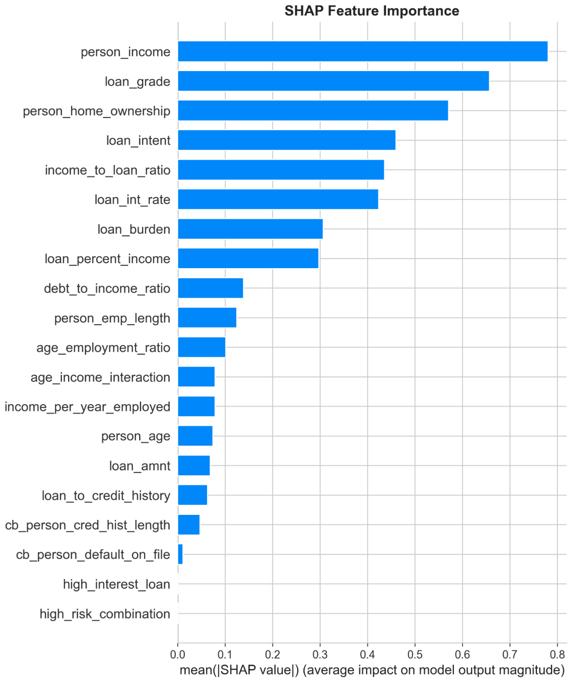

Fig. 17 SHAP bar plot

We compare performance improvement from Preliminary Analysis and the Baseline Models to Enhanced Modeling and Final Model Selection in Figure 18. This highlights overall gains from balancing classes, adding features, and tuning.

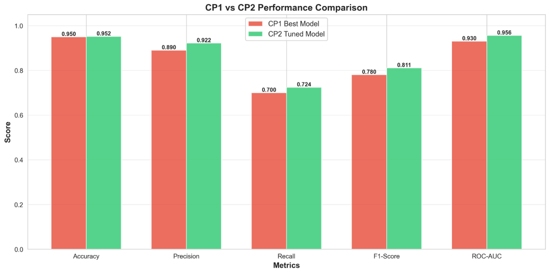

Fig. 18 Preliminary Analysis and Baseline Models vs Enhanced Modeling and Final Model Selection

Finally, learning curves (Figure 19) show stable convergence and no strong signs of overfitting, supporting generalization.

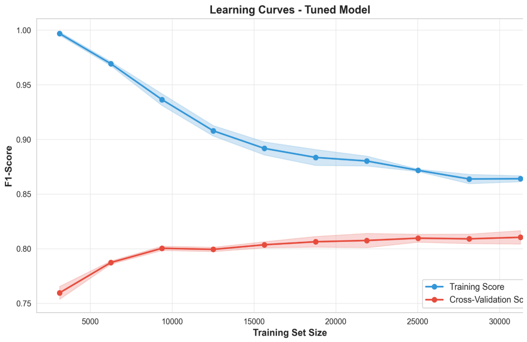

Fig. 19 Learning curves

Overall, the experimental results show clear improvements as we progressed through different stages of the modeling pipeline. Starting with baseline classifiers applied to the original imbalanced dataset, we moved toward more advanced ensemble models trained on balanced data with enhanced features.

The baseline models, particularly Logistic Regression and Decision Tree, had difficulty detecting default cases effectively. However, when we implemented resampling strategies and shifted to ensemble methods—specifically XGBoost, LightGBM, and Gradient Boosting—we observed significantly stronger performance across key metrics including recall, F1-scores, and ROC-AUC values.

After fine-tuning the hyperparameters, XGBoost proved to be the strongest model. These results highlight three key factors that were essential to achieving the best performance: creating meaningful features, addressing the imbalance in our dataset, and carefully optimizing the model's parameters.

VI. Conclusion

In this project, we built a complete machine learning pipeline to predict loan approval outcomes using a real-world dataset with over 58,000 applicant records. Through detailed exploratory analysis, we developed a strong understanding of the key factors that affect loan default behavior. These factors include income level, loan-to-income ratios, interest rates, loan grade, and the length of credit history. One major challenge we faced was the significant class imbalance, as most applicants fell into the approved class. This imbalance showed us that traditional accuracy metrics were not enough. We needed to focus more on recall, precision, F1-score, and ROC-AUC to accurately evaluate model performance.

Our experiments evaluated a range of machine learning models, from simpler baseline approaches like Logistic Regression and Naive Bayes to more sophisticated ensemble methods including Random Forest, Gradient Boosting, LightGBM, and XGBoost. The results clearly showed that tree-based ensemble models outperformed their simpler counterparts, especially when it came to capturing the complex, nonlinear patterns typical of financial data.

When we applied resampling techniques such as SMOTE and SMOTETomek, we saw notable improvements in the models' ability to identify minority class defaults. This reinforced how crucial it is to properly address class imbalance when working with this type of data.

The final tuned XGBoost model achieved strong overall performance, including 95.17% accuracy, a 0.812 F1-score, and a ROC-AUC of 0.953. These results demonstrate that with proper preprocessing, feature engineering, and model optimization, machine learning can be a powerful tool for supporting loan approval decisions. Feature importance and SHAP based explanations further validated that the model’s predictions align with real world lending logic, improving trust and interpretability.

Overall, this project illustrates how data-driven modeling can support safer and more efficient lending practices by reducing the risk of approving high risk applicants while still ensuring fairness and reliability in automated decision making.

A. Future Work

While the final model performs well, several opportunities remain to improve and extend this work:

- Advanced Imbalance Handling: Techniques such as ADASYN and cost-sensitive boosting could further enhance our ability to detect minority default cases. Previous research has shown that cost-sensitive learning approaches and different variants of focal loss can significantly reduce the costs associated with misclassifying cases in imbalanced credit scoring tasks.[8].

- **Richer Feature Set:** Incorporating additional financial and behavioral features could further enhance model performance. Features such as credit utilization ratios or broader economic indicators would provide deeper insights into applicant risk profiles. Research in interpretable machine learning has shown that thoughtfully crafted features are especially important when working with imbalanced datasets, as they help models better distinguish between classes.[5].

- **Model Optimization:** Advanced optimization frameworks such as Bayesian Optimization or Optuna can deliver better results than manual tuning or GridSearchCV. Research on ensemble-based credit scoring has demonstrated that combining these methods with systematic hyperparameter search leads to improved overall performance.[3].

- Deep Learning Approaches: Models like TabNet, TabTransformer, or NODE can improve results by capturing complex nonlinear interactions. Ensemble models currently lead in credit scoring research [4], but deep neural architectures for tabular data are getting better quickly.

- Deployment: Future systems should incorporate fairness metrics such as disparate impact or equalized odds to ensure ethical lending practices. SHAP based explainability methods already provide a strong foundation for transparent decision making [9].

- Fairness and Bias Analysis: Deploying the final XGBoost model as an API or dashboard would enable real time scoring. Prior studies highlight the value of user facing loan decision tools to improve accessibility and reduce human bias in lending [3].

These steps would help transform the model into a production ready system suitable for real world loan risk assessment.

Overall, this project showed that machine learning methods, especially ensemble models like Random Forest can offer valuable support for assessing loan risk and making more informed lending decisions.

VII. Team member responsibilities

Om G. Patel	CP1: Cleaned and prepared the dataset. Handled missing values, encoded categorical variables, and performed standardization. Created new features such as income_to_loan_ratio and age_employment_ratio. Implemented baseline models including Logistic Regression and Decision Tree. CP2: Implemented class imbalance handling techniques including SMOTE and SMOTETomek. I compared dataset distributions and created visualizations CP2_01 for class balance comparison and CP2_03 for dataset comparison.

	Final Report: Wrote Abstract, Introduction, and Problem Formulation sections. Defined the classification problem, described challenges including class imbalance, and outlined project objectives.
Suraj Rao	CP1: Developed and implemented the Random Forest model with hyperparameter tuning. Performed model comparison across baseline models and generated confusion matrices (08) and feature importance plots (09). CP2: Implemented and trained 10 machine learning models (Logistic Regression, Decision Tree, Random Forest, SVM, KNN, Naive Bayes, Gradient Boosting, AdaBoost, XGBoost, LightGBM) across three sampling strategies. Performed GridSearchCV hyperparameter tuning for XGBoost and created top 10 model comparison visualization (CP2_02). Final Report: Wrote Overview of Proposed System and Technical Details sections. Documented the complete ML pipeline, feature engineering steps, model architectures, and hyperparameter tuning methodology.
Darren A. Ferrao	CP1: Created all EDA visualizations including target distribution (01), numerical distributions (02), categorical distributions (03), correlation matrix (04), status by categories (05), boxplots by status (06), and model comparison (07). CP2: Performed model performance analysis including confusion matrices (CP2_04), ROC curves (CP2_05), feature importance (CP2_06), SHAP summary plots (CP2_07), and SHAP bar plots (CP2_08) for interpretability. Final Report: Wrote Experimental Evaluation section including Dataset Description, Evaluation Metrics, Baseline Methods, and Results and Analysis. Interpreted all visualizations and documented performance improvements.
Mohammad A.K.Dalwai	CP1: Analyzed feature importance from Random Forest, identified top predictive features, and wrote CP1 conclusions summarizing key insights and areas for improvement. CP2: Conducted CP1 vs CP2 comparison (CP2_09), created learning curves (CP2_10), synthesized results across all 30 models, and documented final XGBoost performance (95.17% accuracy, 0.812 F1 score, 0.953 ROC-AUC).

VIII. References

[1] R. Abi, “Machine learning for credit scoring and loan default prediction using behavioral and transactional financial data,” *World Journal of Advanced Research and Reviews*, vol. 26, pp. 884–904, 2025, doi: 10.30574/wjarr.2025.26.3.2266.

[2] H. Suftandar, M. Wasesa, and U. Putro, “XGBOOST model for default prediction in credit scoring of conventional bank,” *Jurnal Ilmiah Manajemen, Ekonomi, & Akuntansi (MEA)*, vol. 9, pp. 2164–2172, 2025, doi: 10.31955/mea.v9i2.5920.

[3] N. Uddin, M.K.U. Ahamed, and M.A. Uddin, "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] D. Guo, X. Zhou, H. Li, Q. Zhou, and J. Gong, "Ensemble-based machine learning algorithm for loan default risk prediction," *Mathematics*, vol. 12, no. 21, p. 3423, October 2024.

[5] Y. Chen, R. Calabrese, and B. Martin-Barragan, "Interpretable machine learning for imbalanced credit scoring datasets," *European Journal of Operational Research*, vol. 312, no. 1, pp. 357–372, January 2024.

[6] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[8] X. Zhang, Y. Han, W. Xu, and Q. Wang, "A focal-aware cost-sensitive boosted tree for imbalanced credit scoring," *Expert Systems with Applications*, vol. 208, p. 118147, December 2022.

[9] S. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” in *Proc. NeurIPS*, 2017, pp. 4765–4774.

Dataset Link:
https://www.kaggle.com/competitions/playground-series-s4e10

Code Link:
https://github.com/SurajRao21/Data-Mining-Project