Loan Approval Prediction Using Machine Learning

Hasti Aksoy

September 29, 2025

1 Introduction

Access to credit is a fundamental driver of economic growth, enabling individuals and businesses to invest, consume, and manage financial risks. However, the decision to approve a loan application is a critical task for financial institutions, as it directly affects profitability and risk management. Traditionally, these decisions have relied on manual assessments and rule-based credit scoring systems. While effective to some extent, such approaches are often limited in their ability to capture complex, non-linear relationships between borrower attributes and loan repayment behavior.

With the increasing availability of large-scale financial data, machine learning has emerged as a powerful tool for improving loan approval prediction. By leveraging demographic, financial, and behavioral features, machine learning models can provide more accurate and automated assessments of credit risk. This not only reduces default rates but also promotes financial inclusion by ensuring that worthy applicants are not unfairly denied credit.

In this project, we build a predictive system for loan approval using a dataset containing over 58,000 loan applications with diverse features, including applicant demographics, employment history, credit background, and loan details. The dataset is first preprocessed to ensure data quality, followed by feature engineering to enrich the information available to the model. Subsequently, several machine learning algorithms are applied to predict whether a loan should be approved or rejected. The ultimate goal is to design a robust and interpretable predictive model that can support financial institutions in making data-driven lending decisions.

2 Data Preprocessing and Feature Engineering

The original dataset contained 58,645 records and 13 features, including demographic attributes (e.g., `person_age`, `person_income`), credit history (`cb_person_cred_hist_length`, `cb_person_default_on_file`), and loan characteristics (`loan_amnt`, `loan_int_rate`, `loan_percent_inco` `loan_grade`, and `loan_status`). A thorough inspection confirmed that the dataset was free from missing values and duplicate entries.

To prepare the data for modeling, several preprocessing steps were performed. First, the unique identifier column `id` was removed as it carried no predictive value. Categorical features such as `person_home_ownership` and `loan_intent` were one-hot encoded, while the ordinal feature `loan_grade` (ranging from A to G) was mapped to integer values. The binary variable `cb_person_default_on_file` was transformed into a numeric indicator (0 for No and 1 for Yes).

In addition, multiple domain-inspired features were engineered to capture important financial relationships. These included the **debt-to-income ratio (DTI)**, the **credit history length normalized by age**, and interaction features such as **age** $\times$ **employment length** and **loan amount** $\times$ **interest rate** (denoted as risk factor). Binary flags were also created to highlight borrowers with relatively high income, short employment history, or young age.

After preprocessing and feature engineering, the dataset expanded to 26 features, providing a richer representation of borrower risk profiles for predictive modeling. The same transformation pipeline was applied to the test dataset to ensure consistency.

3 Exploratory Data Analysis (EDA)

A comprehensive exploratory data analysis was conducted to better understand the structure of the dataset and the relationships between features and the loan approval outcome.

The target variable, `loan_status`, was highly imbalanced: approximately 85.8% of applicants did not default, while only 14.2% defaulted on their loans (Figure 1). This imbalance indicates that predictive models need to account for skewed class distributions to avoid biased performance.

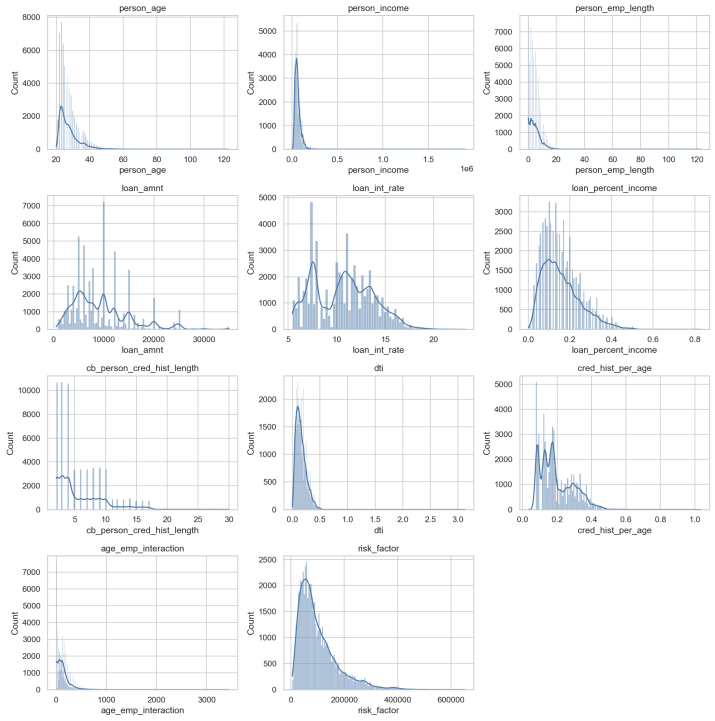

Figure 1: Example of univariate distributions for numerical features.

Univariate analysis of numerical features revealed skewed distributions across most variables. For example, applicant age, income, employment length, and credit history length exhibited right-skewed patterns, indicating a concentration of borrowers with relatively younger age, lower income, shorter job tenures, and shorter credit histories. Loan-specific attributes such as loan amount and interest rate showed multimodal distributions, reflecting different lending categories.

Further analysis conditioned on the target variable highlighted important differences between default and non-default groups. Defaulters tended to have higher loan interest rates, larger debt-to-income ratios, and higher loan percent income ratios (Figures 2, 3, 4).

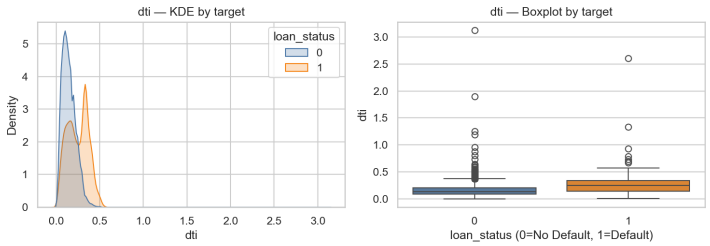

Figure 2: Distribution of Debt-to-Income ratio by loan status.

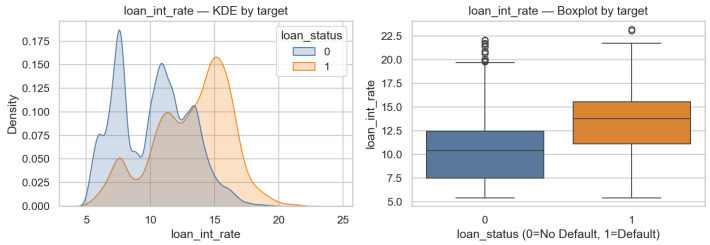

Figure 3: Distribution of loan interest rate by loan status.

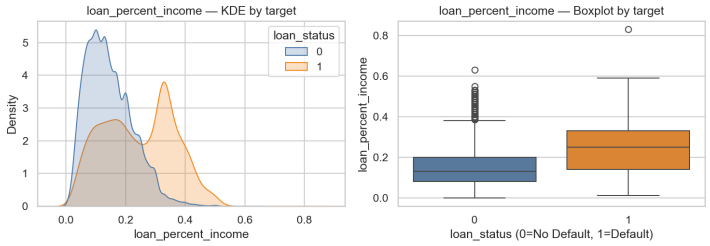

Figure 4: Distribution of loan percent income by loan status.

Categorical features also revealed meaningful patterns: applicants who rent or fall into the "other" home ownership category had higher default rates compared to those with mortgages or owned homes. Similarly, individuals with a record of prior defaults showed a nearly 30% default rate, more than double that of applicants without such history. Loan purpose also influenced default behavior, with medical and debt consolidation loans exhibiting higher default proportions relative to education or personal loans.

Correlation analysis further quantified these relationships. Features such as loan grade, loan percent income, debt-to-income ratio, interest rate, and risk factor were among the strongest predictors of default status, while variables like age and credit history length demonstrated weaker correlations. These findings not only validate the engineered features but also provide valuable insights for model development.

4 Modeling

We developed three supervised learning baselines for binary loan default prediction—**Logistic Regression (LR)**, **Random Forest (RF)**, and **XGBoost (XGB)**—implemented as scikit-learn pipelines to ensure reproducibility and leakage-free preprocessing.

- **Logistic Regression (LR):** Implemented with `lbfgs`, `max_iter=2000`, and `class_weight="bala` Standardization was applied to all numeric features.
- **Random Forest (RF):** Configured with `n_estimators=500`, `class_weight="balanced"`, and `n_jobs=-1`. Median imputation handled missing values, and scaling was unnecessary.
- **XGBoost (XGB):** Implemented with `n_estimators=800`, `learning_rate=0.05`, `max_depth=6`, and subsampling/column sampling for regularization. Early stopping was employed using the validation set. Where XGBoost was unavailable, a fallback to `HistGradientBoosting` was used.

For all models, the decision rule was based on the predicted probability for the positive class (default). A default threshold of 0.5 was applied for validation, while the evaluation utility also supported F1-optimized thresholds for potential business tuning.

5 Results

Table 1: Validation performance of baseline models.

Model	ROC-AUC	PR-AUC	Accuracy	F1	Precision	Recall
Logistic Regression	0.9000	0.6816	0.8183	0.5647	0.4285	**0.8275**
Random Forest	0.9353	0.8457	0.9498	0.7974	0.9369	0.6940
XGBoost	**0.9534**	**0.8711**	**0.9513**	**0.8074**	0.9243	0.7168

Logistic Regression achieved the highest recall (82.7%), capturing the majority of defaulters but at the cost of lower precision (42.8%). Ensemble models reversed this trade-off: both RF and XGB achieved very high precision (>92%), correctly labeling most predicted defaults, but recalled fewer true defaults (69–72%).

Among the ensemble methods, XGBoost outperformed Random Forest across nearly all metrics, with the highest ROC-AUC (0.953), PR-AUC (0.871), and F1 (0.807). The PR-AUC gains are especially important under class imbalance, showing that XGBoost better distinguishes rare defaulters from the majority.

6 Conclusion and Future Work

In this project, we developed a machine learning framework for predicting loan default using a dataset of over 58,000 applicants. Through systematic preprocessing, feature engineering, and exploratory analysis, we identified several critical predictors of default risk, including loan grade, debt-to-income ratio, loan percent income, and interest rate.

We benchmarked three baseline models—Logistic Regression, Random Forest, and XGBoost—and observed distinct trade-offs. Logistic Regression achieved the highest recall, capturing most defaulters but at the expense of precision. Ensemble models, particularly XGBoost, provided superior overall performance, achieving the highest ROC-AUC (0.953), PR-AUC (0.871), and F1 score (0.807). These results highlight XGBoost's robustness in handling complex, non-linear relationships and imbalanced data distributions.

The study demonstrates that predictive modeling can significantly enhance the loan approval process, helping financial institutions manage risk more effectively while promoting fairer access to credit. However, model selection and threshold tuning should ultimately align with organizational priorities—whether the goal is to minimize default risk, maximize loan approvals, or strike a balance between the two.

Future work could extend this study by:

1. Incorporating categorical embeddings or more sophisticated encoding strategies for non-numeric features.

2. Exploring ensemble stacking or hybrid models that combine the recall strength of logistic regression with the precision of boosting methods.

3. Performing cost-sensitive learning to explicitly model the asymmetric costs of false positives versus false negatives.

4. Applying explainability techniques (e.g., SHAP values, feature importance) to provide transparency for regulatory compliance and business stakeholders.

5. Validating the model on out-of-time datasets to assess temporal robustness and generalization in real-world loan portfolios.

In conclusion, the XGBoost model presents a strong baseline for loan approval prediction, offering a balance of accuracy, interpretability, and adaptability for future deployment in financial decision-making systems.