Predicting Loan Repayment with Gradient-Boosted Decision Trees

Trisha Wu

1. Introduction

Predicting whether a borrower will repay a loan is a core problem in consumer finance. Modern machine learning methods can capture nonlinear interactions among income, debt burden, credit history, and demographic factors. Using the Kaggle Playground S5E11 dataset, supplemented with a historical loan dataset of 20,000 records, this project develops a model for the binary task of **loan_paid_back** (1 = repaid, 0 = default).

Our goal is to build a strong predictive model and examine which engineered features most strongly relate to repayment behavior. A LightGBM classifier with 5-fold cross-validation achieves an out-of-fold AUC of **0.9249**.

2. Data

The dataset contains financial attributes (annual income, loan amount, debt-to-income ratio, interest rate), credit variables (credit score, grade_subgrade), and categorical factors such as employment status and loan purpose.

Basic exploratory analysis shows expected structural patterns: borrowers who repaid loans generally have higher credit scores, lower debt burdens, and more favorable credit grades. Employment status and loan purpose also correlate with repayment rates.

3. Methods

3.1 Feature Engineering

Feature engineering is central to model performance. Key transformations include:

Target-encoding features.
Using the auxiliary dataset, each feature receives two statistics—mean repayment rate and count frequency—which provide historical context for both numeric and categorical values.

Financial ratio features.
These capture a borrower's financial capacity relative to loan obligations:

- loan_to_income, total_debt, available_income, affordability
- monthly_payment, payment_to_income

Risk score.
A composite feature combining debt-to-income ratio, credit score, and interest rate.

Credit grade decomposition.
grade_subgrade is split into grade (A–G), subgrade number, rank, and a combined index to reflect the hierarchical credit-rating system.

Additional transformations.
Interaction features (credit_interest, income_credit, debt_loan) and log-scaled income and loan amounts. Missing numeric values are filled with medians, and categorical columns are converted to pandas categorical types for efficient LightGBM handling.

3.2 Model Training

We use **LightGBM (LGBMClassifier)** with a binary objective and AUC metric.
Training setup:

- **5-fold KFold cross-validation** (shuffle=True, fixed seed)
- **Early stopping** (100 rounds)
- **learning_rate = 0.01, n_estimators = 10000**
- Best hyperparameters:
    - num_leaves = 16
    - colsample_bytree = 0.2
    - subsample = 0.8
    - reg_lambda = 0.0

Out-of-fold predictions from all folds are used to compute the final score.

4. Results

The final LightGBM model achieves an **AUC of 0.9249**, demonstrating strong predictive ability. Across multiple hyperparameter trials, AUC values remained in the 0.9243–0.9249 range, indicating that performance is robust and driven largely by feature engineering rather than fine-tuning.

Features related to credit score, income-to-loan ratios, debt burden, and credit-grade decomposition consistently contributed most to model performance, aligning with intuitive financial risk factors.

5. Conclusion

This project shows that gradient-boosted decision trees combined with structured feature engineering can effectively model loan repayment behavior. Financial ratios and credit-grade features were particularly influential, highlighting how risk emerges from interactions among income, debt, and institutional credit assessments. The results suggest that machine learning provides a strong framework for understanding and predicting loan outcomes.