

Loan Approval Prediction - Project Report

1. Project Overview

The Loan Approval Prediction project involves predicting loan eligibility based on various factors such as the applicant's age, income, employment length, loan intent, and credit history. This project uses machine learning models to predict the likelihood of loan approval, employing feature engineering and ensemble methods to optimize model performance.

2. Dataset Information

The dataset contains various features including demographic information, financial details, and credit history of loan applicants. The key features include:

- person_age: Age of the applicant
- person_income: Annual income of the applicant
- loan_amnt: Amount of loan applied for
- loan_intent: Purpose of the loan
- loan_grade: Risk grade assigned to the loan
- cb_person_cred_hist_length: Credit history length
- loan_status: Target variable indicating loan approval (1) or rejection (0)

3. Feature Engineering

To enhance the performance of the models, new features were created based on existing ones.

Some of the engineered features include:

- income_to_loan_ratio: Ratio of the person's income to the loan amount
- cred_hist_loan_amnt_ratio: Ratio of credit history length to loan amount
- age_income_ratio: Ratio of the person's age to their income
- age_cred_hist_interaction: Interaction between age and credit history length

4. Model Building

Several machine learning models were used in this project, including:

- XGBoost: A gradient boosting algorithm known for high performance
- Random Forest: An ensemble of decision trees
- LightGBM: A fast, gradient boosting framework
- CatBoost: A boosting algorithm designed for categorical data

The models were stacked using Logistic Regression as the meta-model, allowing us to combine the strengths of multiple classifiers.

5. Model Evaluation

The models were evaluated using ROC-AUC score, precision, recall, and F1-score.

- The Stacked Model achieved a ROC-AUC score of 0.9584.

Detailed performance metrics such as precision and recall were also calculated to ensure balanced classification.

6. Hyperparameter Tuning

GridSearchCV was used to perform hyperparameter tuning for the Logistic Regression meta-model.

The key parameters tuned were:

- C (Regularization strength)
- Solver (Optimization algorithm)
- Penalty (Regularization type)

The best ROC-AUC score after tuning was achieved with the following parameters:

C = 0.1, solver = 'liblinear', penalty = 'l2'.

7. Conclusion and Results

The Loan Approval Prediction project successfully implemented several machine learning models to predict loan eligibility. The final stacked model with hyperparameter tuning achieved a high ROC-AUC score of 0.9584, indicating strong predictive performance.

Further improvements can be made by exploring additional feature engineering techniques, more advanced model tuning, or even incorporating more sophisticated ensemble methods.