Loan Approval Prediction using Machine Learning

Team Members:
Yael Sisniega & Yoselin Reyes

This project focuses on reproducing and extending a research paper on loan approval prediction using machine learning classification techniques. The main objective was to evaluate whether the original methodology generalizes well to similar datasets and whether performance can be improved through systematic model tuning and the inclusion of additional classifiers.

Motivation

Loan approval is a critical decision-making process in the financial sector, where incorrect decisions can lead to financial loss or unfair denial of credit. We chose this research paper due to its practical relevance and its use of classical machine learning models, which allowed us to study interpretability, performance, and robustness across datasets.

Research Paper Details

The selected research paper proposes a supervised learning framework for predicting loan approval using structured applicant data. The paper evaluates multiple classification models and reports performance using accuracy-based metrics. Our work aims to reproduce the key steps of this methodology and assess its effectiveness on related datasets.

Dataset Details

The datasets used consist of applicant-level information such as demographic attributes, financial indicators, and credit-related variables. The target variable represents whether a loan application was approved or rejected. The datasets contain a mix of numerical and categorical features and vary in size and distribution, enabling robust cross-dataset evaluation.

Data Preprocessing and Feature Engineering

Data preprocessing involved handling missing values using appropriate imputation strategies, encoding categorical variables through one-hot encoding, and scaling numerical features where required. Redundant identifiers were removed to avoid data leakage. These steps were implemented using scikit-learn pipelines to ensure consistency across models.

Steps Reproduced from the Paper

We reproduced the core steps described in the paper, including dataset preparation, baseline model training, and evaluation using standard performance metrics. Decision Trees and Logistic Regression were implemented following the original methodology to establish comparable baseline results.

Contributions

Our primary contributions include testing the methodology across multiple related datasets, introducing repeated train-test evaluation to assess robustness, tuning model hyperparameters using GridSearchCV, and incorporating additional classification models such as Support Vector Machines and Random Forests.

Significant Improvements

The inclusion of hyperparameter tuning and ensemble learning through Random Forests resulted in significant performance gains. The tuned Random Forest model achieved the highest mean accuracy, precision, and ROC-AUC across repeated experiments, demonstrating improved generalization and ranking performance compared to baseline models.

Challenges

One of the main challenges encountered during this project was the lack of full consistency across datasets. While all datasets addressed loan-related decisions, their target variables and feature definitions were not always directly comparable. For example, some datasets explicitly labeled loan approval, while others inferred rejection through zero loan amounts. Additionally, some datasets contained incomplete or missing information for key variables.

These inconsistencies made direct model comparison more difficult and required careful preprocessing and interpretation to avoid misleading conclusions. Despite these challenges, the methodology was adapted to maintain fairness and analytical rigor.

Conclusion and Future Scope

In conclusion, our results show that extending the original methodology with ensemble learning and systematic tuning can significantly improve model performance. Future work could explore fairness analysis, model interpretability techniques, and validation on real-world loan datasets to further strengthen the applicability of the approach.