Executive Summary

Introduction

This project addresses the critical financial issue of credit card default by leveraging historical data to predict future delinquency. The data, sourced from the Kaggle competition "Give Me Some Credit," encompasses a broad spectrum of features. These include age, number of dependents, open credit lines, loans, monthly income, revolving credit utilization, debt ratio, and frequency of missed payments.

Methodology

One of the unique features of our project is the use of data engineering to enhance the dataset. The project began with a rigorous data cleaning process, which involved eliminating invalid values and outliers to ensure data integrity. A comprehensive correlation analysis between different features resulted in the creation of a new column through feature engineering. This column encapsulates highly correlated columns, thereby enhancing the dataset's analyzability and providing a more robust foundation for modeling.

Several predictive models were evaluated, starting with Linear Regression. Despite its simplistic interpretability, it did not yield satisfactory results. Logistic Regression was subsequently employed; however, even after hyperparameter tuning, it still fell short of the project's performance criteria.

The Random Forest model was then implemented, achieving a satisfactory AUC score of 0.81. However, due to its extensive runtime, the XGBoost model was explored. This model proved to be the most efficient and accurate, achieving an AUC score of 0.84 after hyperparameter tuning. As a result, we made the final test predictions using XGBoost. Furthermore, our model also returns the three factors that affected the credit score the most, allowing a consumer whose loan was rejected to understand the parameters that influenced their credit score.

Conclusion

The project successfully predicts the likelihood of credit card default and provides personalized improvement suggestions. These recommendations, based on the three most influential factors contributing to potential default, provide actionable insights for individuals. This dual approach of prediction and personalization could serve as a prototype for financial institutions worldwide, potentially transforming risk assessment and customer engagement strategies. Our three-feature identification adds a layer of transparency and can help consumers improve their creditworthiness for future applications.

Data Source: https://www.kaggle.com/competitions/GiveMeSomeCredit/data
Github Repository: https://github.com/tanujmath/CreditGuard