Credit Risk Prediction Using Classification Models

Table of Content

Table of Figure

Introduction

In this task, the objective is to develop predictive models using three different classification methods to identify the key features influencing whether a client is likely to experience serious delinquency within the next two years. The dataset has undergone thorough pre-processing, including handling of missing values and treatment of outliers, to ensure high data quality for accurate model development.

To support robust performance evaluation, two versions of the dataset were prepared: one with the original feature values and another transformed using Weight of Evidence (WOE), a technique particularly well-suited for enhancing the interpretability and effectiveness of logistic regression models.

Model Selection

Given the binary target variable (Delinquent: 1 indicates delinquency, 0 indicates no delinquency), three classification models were chosen for the analysis:

1. **Logistic Regression**: A well-known, interpretable model suited for binary classification tasks. It performs best with WOE-transformed inputs, as this transformation encodes variables to reflect their relationship with the target, enhancing predictive accuracy.
2. **Random Forest**: An ensemble learning method that builds multiple decision trees and aggregates their predictions. It is robust to overfitting, capable of handling nonlinear relationships and variable interactions, and does not require transformations like WOE.
3. **XGBoost (Extreme Gradient Boosting)**: A powerful gradient boosting method known for high predictive performance. It includes regularization techniques to combat overfitting and is highly efficient for managing large datasets

Model Evaluation

Metrics	Logistic Regression	Random Forest	XGBoost
ROC AUC Score	0.8284	0.8569	0.7860
Accuracy	0.78	0.81	0.92
Precision (0)	0.97	0.98	0.94
Precision (1)	0.19	0.22	0.32
Recall (0)	0.79	0.82	0.98
Recall (1)	0.70	0.72	0.16
F1-score (0)	0.87	0.89	0.96
F1-score (1)	0.30	0.34	0.21
Macro Avg (F1)	0.59	0.62	0.59
Weighted Avg (F1)	0.83	0.85	0.91

Table 1 Error metrics

Among the three models evaluated—Logistic Regression, Random Forest, and XGBoost—Random Forest demonstrated the most balanced and effective performance for predicting serious delinquency within two years (Table 1). While Logistic Regression offered strong recall (0.70) for identifying delinquent clients and allowed the generation of a scorecard for interpretability, it suffered from low precision (0.19), indicating a high number of false positives. XGBoost achieved high accuracy and precision for non-delinquent clients but failed to detect most delinquents, with a recall of just 0.16. In contrast, Random Forest achieved the highest ROC AUC score (0.8569), strong recall (0.72), and better precision (0.22) for delinquent cases, making it the most reliable model for distinguishing risky clients from safe ones. Given the business need to prioritize minimizing false negatives in financial risk prediction, Random Forest is the most suitable model for deployment.

Conclusion

In conclusion, the Random Forest model outperformed others in predicting serious delinquency, with the highest ROC AUC and recall. Logistic Regression followed closely and enabled the creation of an interpretable scorecard. Key features like Revolving Utilization, PastDue variables, and Age were consistently important across both models, reinforcing their relevance in credit risk assessment. Combining Random Forest's accuracy with Logistic Regression's interpretability offers a strong, practical approach to credit scoring.