Loan Default Prediction

Executive Summary

Name: Hewei Shen
Email: shenh23@wfu.edu

Business Question

The goal of this project is to develop and compare machine-learning models that can accurately predict default loan transactions in peer-to-peer lending on the Lending Club platform.

Model Comparison

We developed 5 classification models using Logistic Regression, Random Forest, Gradient Boost, Neural Network, and Stacking Classifier. Based on the AUC score on the test set, we chose the Gradient Boost model as our best model, with an AUC of 0.894 and an F1 score of 0.563.

Feature Evaluation

The Top 5 most important features of the Gradient Boost model are the last credit pull date, the self-reported annual income of the borrowers, the interest rate of the loan, the issue date of the loan, and the loan purpose. Among them, the last credit pull date has the most significant importance.

The FICO score does not have a significant predicting power in all 5 models. In fact, the FICO score in the model may even hurt the predicting accuracy in some cases.

Understand ROC AUC and Precision-Recall Curve

In the model report, we provided ROC AUC and Precision-Recall curve for each model. The ROC curve illustrates the trade-off between the True Positive Rate (TPR) and the False Positive

Rate (FPR) across various threshold values for classification. We can use the ROC curve to select an optimal threshold based on their specific business objectives and risk tolerance.

We can use the PR curve to evaluate the model's performance, especially in scenarios where class imbalance exists. PR curve helps in selecting an appropriate threshold that maximizes both precision and recall, balancing the identification of true positives while minimizing false positives.

Operational Strategy at 5% FPR

To operate and maintain a 5% false positive rate, the company needs to set a threshold of 0.7132 in our best model - the Gradient Boost model. This means if the predicted probability is larger than or equal to 0.7132, then the transaction would be classified as default. Additionally, the recall at the 5% FPR is 0.4234.

Understand 5% False Positive Rate

Operating at a 5% FPR means that out of all the default transactions detected by the model, only 5% of them are actually current.

A high FPR will negatively affect the customer experience by investigating the default behaviors of a current customer. For example, a current customer may be asked to provide backup documents or a higher interest rate. This might result in customer churn. The investigation is also time-consuming and costly.

Business Recommendations

To prevent default, set the tracking mechanism on the Top 5 important features instead of the FICO scores. The operation team should proactively reach out to customers who have a high default probability or have their default probability changed significantly in a short period.