Project 3

Loan Defaults

Niklas Baldis

# Executive Summary

## Analysis

Five total models were created to attempt to predict accounts that will default on their loan. Exploratory Analysis of the dataset found that $5000 is usually the most fraudulent loan amount for accounts, with a frequency of about 400 of the nearly 3000 total accounts given that loan amount. Approximately $5000 is also the most commonly invested amount for fraudulent accounts, with a frequency of about 350 of the nearly 1750 accounts with that investment amount. Many numerical features also exhibit a right skew, indicating that lower-volume and newer accounts are usually more fraudulent. An optimized Gradient Boost model produced a ROC of 0.8992, a precision of 0.7083, a recall of 0.3666, and an F1 score of 0.4832. ROC is one of the strongest metrics to evaluate models because it defines how well the model can distinguish between the two categories of default accounts and current accounts. This ability to accurately separate the groups leads to a stronger predictive model on future datasets. The precision-recall curve of the model was a 0.64, indicating that the model strongly finds the defaulted accounts among all accounts and among all defaulted accounts.

## Recommendations

I would not use credit score as a predictor of fraud because the Gradient Boost model identified last payment amount as the most important variable to predicting fraud at nearly 41% of the model. Other variables like interest rate and the number of payments on the loan were also found to be more important. Operating at a 5% false positive rate means that out of every 100 accounts flagged as legitimate, 5 defaulted accounts are falsely labeled as legitimate. The 5% threshold for logistic regression correlates with a 100% true positive rate. GBM calculates a TPR of about 50%. Although there is a risk in allowing defaulted accounts to pass through the models, the model correctly identifies both defaulted and legitimate accounts at a 95% accuracy rate. However, the system still makes occasional mistakes, highlighting the need for fine-tuning to improve efficiency and customer experience. In this scenario, using a threshold of 0.389132 for predicted probability would capture about 50% of all defaulted accounts with a precision of 95% while incorrectly flagging 5% of legitimate accounts as defaulted.