Matthew Sparacio

Final Project: Loan Default

12/9/22

Executive Summary

# Problem

The goal of this project is to predict the status of a loan, specifically which loans are likely to default. Being able to predict likely loans to default will allow us to address them and save money by preventing them. To accomplish this goal, we built a machine learning classification model that makes predictions about loan status.

# Key Findings

An important fact which is vital to the whole analysis is that loans default occurs only 15% of the time. This has a great influence on the model and the state of loans as a whole. The first key finding of the study was that as the grade of a loan decreases (from A to G), a loan becomes much more likely to default. A loan of grade A defaults only 10% of the time while a G grade loan defaults over 30% of the time. Credit score is another variable that often influences if a loan will default. The average current loan has a credit score almost 50 points higher than the average default loan. Lastly, I found it interesting that if the loan is being paid on a payment plan, they never default. If they are on a payment plan, the loan always gets paid.

# Model Summary

The best model that we created and used for predictions was a gradient boost model. This model was able to use almost every variable in the data set to make predictions with 97% accuracy on the test set. But it is important to understand that accuracy is a poor measure of performance in this case. A model using this data can predict that no loans will default and be 85% accurate because only 15% of loans default. Better metrics to use are precision and recall which are based on the number of true positive predictions vs false negative and false positive predictions. We are most interested in predicting if a loan will default and we can maximize the true positive rate while minimizing the false positive rate. The graph to the below shows the optimal operating range of the model for best predictive capabilities.

Chart

Description automatically generated

Because of industry regulations and the need for transparency from the model, we have insight into which factors influence the model’s prediction of if a loan will default and by how much and in which direction. The single factor that influences the model the most is the last payment amount. Having a last payment amount of $10,380 makes it very unlikely that the loan will default. Below is a graph of the most important predictors.

Chart

Description automatically generated

Although the model performs very well, it is not perfect. We compiled a list of the models 10 best predictions of default, the 10 worst estimates where we predicted default but it did not, and the 10 worst estimates where we predicted no default but it did default. The best predictions were mostly on 36-month term loans, had a middle of the road grade, and were early in their careers (low employment length). Their last payment amount was fairly low and their last credit pull was the shortest it could be. The worst false positives were often on higher grade loans with low interest rates. They tended to have higher fico ranges and some last payments were well above average. The worst false negatives was fairly diverse and middle of the road. Many of the people with these loans were renters which could be why we predicted they would default.

# Recommendations

Based on our analysis, we would like to offer some recommendations that our analysis brought to light. First, from the chart of variable importance, we can see that having a term of 60 months rather than 36 months makes loans much more likely to default, so we would recommend more shorter-term loans. Additionally, I would recommend periodically monitoring the dates related to a loan. These are almost all good predictors of default and are easy to set up flags for likely defaulting loans.