Presented by: Eshaan Vora

05/13/22

Stout: Full-Stack Data Analyst

Executive Summary of Case Study:

Case Study 1:

This dataset includes various information about loans that have already been made including the characteristics of the loan receiver.

Specifically, there is significant information about individual’s delinquencies, debt profile, and credit-worthiness.

The most predictive variable in modeling the interest\_rate was the “sub-grade” feature and this makes sense because the “sub-grade” is a measure of investment quality and therefore, the lower the investment grade, the more risk must be compensated for by a higher interest rate.

My first step in cleaning the data was to understand the scope of the missing data and which features were most affected and so I included a function to print number of “N/A” values per feature. I found that a majority of missing values were due to the fact that most individuals were individual filers and so I separated the dataset into single filers and joint filers. This allowed me to address the bulk of the missing values without compromising predictive quality. I then checked each string variable’s factor level and dropped variables that only had 1 unique value as that would have no effect on model performance.

I built OLS and Lasso Regression models to understand which variables were most important and found out that the model error started plateauing at 36 variables, meaning there were some variables, out of the 43 variables in the model, which had little to no affect on model performance. I isolated and removed some of these variables such as “loan\_status”, “term”, or “total\_debt\_limit” I also built a Random Forest Model to provide an ensemble algorithm to help prevent overfitting. If I had more time, I would focus on validating and tuning the model more, as there are useful metrics that could be applied such as “root mean squared error (RMSE)”

Case Study 2:

There seemed to be significant growth in new customers every year marred by the large amount of customers that did not purchase again. It was an interesting and rewarding experience trying to figure out and breakdown the data transformation steps that is required to convert the information given into the desired output.