Capstone Statistical Analysis Report

The overall goal of the capstone project is to predict the interest rate of a single customer for a requested loan, given an amount of information about their financial, legal, and other statuses. Specifically, for this report I will be answering questions focused on the data analysis and manipulation done in order to draw some meaningful conclusions before applying the data to machine learning algorithms.

Firstly, the subject of cleaning needs to be addressed. In total, the dataset has 139 columns that act as categories and over 40,000 rows representing individuals who applied for a loan and received one. The simplest thing to address first is missing values. There were columns which only had about 10-40% filled out with potentially useful information. The argument can be made that there is still enough information overall to draw some meaningful conclusions, but we would have to create a different dataset for these specific columns and run completely different tests. Looking ahead, the time invested in these columns more than likely would not be well spent since the contribution to the accuracy is probably minimal. After removing these problem columns, in total, there were less than 100 rows that contained missing values. Instead of fabricating some zeros, which would not have had a large impact on a dataset this size, I found it appropriate just to remove those rows altogether. Next, was determining which categories could be considered useful in our analysis and predictions. Every column was analyzed and was either found significant or relevant enough to keep or discarded. By the end, I had 18 features. Lastly, some of the columns had percentage signs and numbers that were of type string, so fixing that was the final detail.

Next, I wanted to learn more about my dataset and see what information I could gather from it using graphs and functions. The first step I took was checking the distribution of our income to see if it was normal. Although it did have a slight skew to left, it was mostly normal so we can progress with that in mind. Then I wanted to test out a basic theory. Is income related to interest rate? That feels like a reasonable question to ask and intuitively the answer should be yes. But it was not. The correlation coefficient for the relationship between income and interest rate was 0.05. That number allows us to say nothing about their relationship except that they are not. The next logical step is to see if maybe employment length has something to do with it. Maybe if we separate people by their employment length and then try to relate income to interest rate, there will be something there. I continued to do so and created heatmaps, distribution plots, and calculated the correlation coefficient for each of the categories being 0-9 years of employment and then a 10+ years category. They all shared the same distributions with the same peak and with similar heat maps as well. The correlation coefficient never went over 0.11. Next, I wanted to see if maybe the mean of the interest rates of each of the employment length categories had any immediate trend. The values on that plot were so scrambled it could have been served at Denny’s with pancakes. I tried similar methods with home ownership too and nothing came from it.

To take things just a set further, I went ahead and attempted lasso regression to see if there was any dominating factor in determining interest rate. Although the raw regression did poorly, there were no huge contributing factors that stood out.