Applying Statistics to the Capstone Project 1

In my first Capstone Project, I am trying to build a model that would predict “Payment in Full” on leases provided through the Lending Club platform.

Previously, I had looked at the variables in the Lending Club data set to try to get an idea as to potential correlations between the independent and dependent variables.

In my “Applying Statistics to Capstone Project 1.ipynb” Jupyter notebook, I generated a subset of the entire data set to be able to build a quick model to see which variables in the data set would show up as being important/significant to the model. I did this because I have over 150 variables in this data set and needed a simple way to focus my attention on variables that would eventually be relevant to the model I would later build.

* To eliminate some of the variables up front, I looked at a Pandas Profiling file I had previously generated to find variables with categorical columns that are unique, have entirely missing values, are rejected, are unsupported, have high cardinality, or are date fields.
* I then had to isolate a list of categorical variables to be able to generate dummy variables for modeling, which consists of changing categorical variables into binary flag variables.
* With this list I generated the dummy variables
* I then found all the variables that had NaN values and imputed the NaN values

With the data in place to start building a model, I decided to build a RandomForestClassifier model. At this point I used the feature\_importances\_ variable and found that most of the top 45 variables in the model were Performance variables that I would have not had access to at time of lease application, so I had to remove those.

After removing them and rerunning a RandomForestClassifier model, I used the feature\_importances\_ variable again to see that the top 45 features were all ok to leave in the model.

I then built a barplot in Seaborn to look at the features in order of importance to focus my research on feature correlations. I looked at the top 3 variables and tried to see if they were correlated with the Perf response variable.

I used the Pearson correlation coefficient to see how correlated the variables were to the Perf response variable. The coefficient takes on values between -1 and 1, with 0 meaning no correlation while 1 is positively correlated and -1 is negatively correlated.

* The top variable last\_fico\_range\_low had a Pearson correlation coefficient of 0.5791423235578518, which means it is somewhat highly correlated.
* The second variable last\_fico\_range\_high had a Pearson correlation coefficient of 0.7102556649422719, which means it is even more highly correlated
* The third variable dti (Debt to Income ratio) had a Pearson correlation coefficient of -0.04553847373302067, which means it is not correlated as the coefficient is extremely close to 0.

From this analysis, I would have thought that the most relevant feature for the model would have also been the most correlated to the response variable. As we can see from the results on the comparison of the three most relevant variables in the model, the second most relevant is the most correlated to the response variable, while the first is still strongly correlated, and the third is not correlated.