**Lab 2 – Feature Selection**

**Revisiting Data Preprocessing:**

After doing some analysis and brain storming I planned to revisit my preprocessing that I did previously. I realized that replacing the missing values in those columns will effect the results of my model in the future and plus after dropping the missing values and data points, the size of the dataset did not decrease drastically so I was actually not loosing enough information. Plus I can to realize after some analysis that the columns that had missing values were missing at the same sample majority of the time. So that sample did not contain much information so dropping that was not going to change anything. So in the end I decided to drop the missing values. There was one column named ‘’Loaned From’’ that had missing values in all of the data points so I dropped that column as well.

**Converting Data Types:**

After dealing with the missing values, there were some columns that should I have been float or in int but were actually an object. This was because the were units along with the numbers. Like ‘’$123K’’, so I replaced every “K” with a 1000 and “M” with 1000000 and removed the $ sign, then converted that into int or float Data type. There were about 7 columns that had data values like these so applied similar algorithms to deal with them and convert there Data type to numerical. There was one column named “Joined” that should be a datetime datatype so I changed that to datetime.

**Visualizing:**

After dealing with missing values and data types, I visualized every column in the dataset to see it spread, basic curve of the column any outliers, etc. The visualization can be seen in the Figure folder. Majority of the columns had a normal curve or a skewed normal curve, there were some outliers in the dataset. There were columns that had a lot of outliers like “Value” column, but I decided to keep them because in this column the curve was bimodual. With these visualizations I had a better understanding of the dataset.

**Label Encoding:**

Since there were few columns that had categorical data, so I label encoded them to convert them into numerical form. This was done because a machine learning model only accepts numerical values.

**Feature Selection:**

After completing the process of transforming the data into a form that can be used in a model, I started with feature selection. First step was to find correlation of columns with the target column which in this case is ”Release Clause”. The visualizing can be seen in the figures folder. The graph showed that there were 8 columns that showed strong correlation with the target column. Few of them are “Value”, “Wage”, “Potential”, “Overall”, etc.

After correlation I applied various algorithms of Sklearn library to find out the best features. The algorithms I used were “Variance Threshold”, “SelectKBest”, “Mutual Info”, “Percentile”, “RandomForest”, “Chi-Squared”, “PCA”, “Genetic Algorithm”. I visualized some of the results of these algorithms. After applying these algorithms I came to realize that “Overall”, “Potential”, “Value”, “Wage”, “Reactions”, “Composure” were the best features for this dataset.