**Capstone Project – Statistical Analysis**

Feature Selection and Feature Extraction

There are many methods of feature selection – filter methods, wrapper methods and embedded methods. While the filter method may not be very accurate, the other two methods are computationally expensive and work when there are not more than 20 - 30 features.

Principal component analysis is a technique for feature extraction — so it combines our input variables in a specific way, then we can drop the “least important” variables while still retaining the most valuable parts of all of the variables. PCA also helps in visualizing multidimensional data. PCA reduces data by geometrically projecting them onto lower dimensions called principal components (PCs), with the goal of finding the best summary of the data using a limited number of PCs.

In this data set, the 2D features are numbers representing the difference between the surface map of 'normal' amino acid (wild-type) and the mutated one. The 3D features represent the difference in magnitude of the distance changes in the 3D structure. This kind of multidimensional data is very hard to visualize for any specific feature correlation. We can try feature selection and dimensionality reduction on the data set and then try to visualize the key components. These features are bound to have highly correlated variables and dropping the highly correlated variables is part of feature selection – filter method.

The correlation matrix of the numerical predictors was generated and the features showing high correlation are dropped - a correlation coefficient of 1 shows maximum correlation. Two different cut off scores were chosen – correlation coefficients of 0.7 and 0.9 - and the selected dataset from both cut offs were analyzed. When the cut off was 0.7, we dropped 3744 features from 5414 features whereas with a cut off of 0.9, we dropped 1625 features.

The numerical features of these 2 selected datasets and the entire data set was then standardized using the minmax\_scale. This transforms the features by scaling to a range between 0 and 1.

The scaled data were converted to data frames and then used for PCA. The PCA model was then analyzed for their explained variance ratio and a cumulative of the variation is plotted. This graph shows the number of components along the X axis and the percentage of variation explained along the Y axis. This graph was the same across all the 3 data sets – 2 datasets selected from correlation matrix cut offs and the entire dataset. Approx. 250 components were needed to explain 90% of the variation in data. Considering this, we can use just the PCA analysis for the feature selection/dimensionality reduction instead of using the correlation matrix and a cut off.

The PCA plots were plotted using the first 2 components. In the first plot, the ‘Type’ variable was also plotted as different colors. In the second PCA plot, the ‘count’ variable was the third dimension and the ‘Type’ was also included as different style for ‘active’ and ‘inactive’.

Since the data is skewed to begin with and only 30% of the variation is explained by the 2 principal components, we cannot see a good partitioning of the clusters but we can at least visualize the data.

The next step would be to run a baseline model using different classifying algorithms and use cross validation to get a good prediction score.