Sahil Rangwala
MSDS 422

1. Discuss what your models tell you in layman's terms.

The dataset came out with a lot of interesting insights prior to modeling. There was not much sparsely empty data, only not defined within Cabin and Age primarily. This could be due to the tragicness of the event, this data was lost. In order to accurately build our models, I needed to clean the data to remove these null values as well as reshape it for the models. I found that there existed a lot of Null values within the Cabin feature as well as a few empty values in certain categories for individuals. Each feature was handled case by case. The data seemed like many different features could possibly predict the survival rate of the individuals. Due to this, my gut reaction is that KNN would fare the best.

Once model building began, from the precision and recall metrics, it is seen that logistic regression was an okay predictor of survival rate. With a precision score of 0.768 and a recall score of 0.698, the logistic regression indicates some sort of prediction that is accurate but definitely, there are models that could offer more of a confident answer. Quadratic has worse of a score which is surprising as I would assume this model fares better than logistic regression. QDA is used to help group data into known groups, thus I would assume it does better in a situation where multiple factors influence the grouping. Surprisingly, through the hidden test cases, the QDA approach fared better than logistic regression.

Nonetheless, KNN does the best as it accurately classifies the data even with the multi classes that exist. KNN is known to perform well where it is used for multiclass classification. The KNN algorithm assumes that similar things exist in close proximity thus in a scenario where many factors could influence the outcome I would assume this performs the best. Additionally, 5 neighbors were chosen as different numbers between 5 - 11 neighbors were tested but 5 offered the best performance. Possibly, further outlier analysis needs to be done in order to clean the data further. Outliers can greatly influence KNN performance as they can misclassify points.

The scores of the three model techniques are below.