NFL Capstone - Machine Learning

This section of my capstone project consisted of trying out a few different machine learning models in order to identify which one would be best suited for my problem - Whether or not we can identify great NFL quarterbacks based solely on their college statistics.

To begin, I scaled all of the numeric data from 0.0 to 1.0. By scaling the data, I hoped that it would make the multiple linear regression model that I used in the data storytelling/statistics notebook more robust. After the scaling process, I ran the model just as I had done last time, only to find that the results were nearly identical. The R-squared and adjusted R-squared values were the same. I tried to take away variables whose p-values were noticeably high, and continued to take away and rerun the model a few more times until I was left with a few variables that made the model as optimal as possible from an explained variance standpoint. However, the adjusted R-squared at the end of this process was still only 0.051, which points to a fairly bad model. Basically, what it is telling me is that the model explains only 5% of the variability of the response data around its mean. So 95% of that variance is explained by data that I have not use and/or do not have access to.

I then tried to perform a logistic regression model. Before I began, I established a response column to help the model divide the players into two categories: If a player had a career Approximate Value of greater than or equal to 0.15, he was considered a 'good' quarterback. If his career AV was less than 0.15, he was considered 'bad'. I ran the model and visualized the confusion matrix in the form of a heatmap. It told us that the diagonal values (101, 0) were actual predictions and the non-diagonal values (24, 0) were incorrect predictions. Our model's accuracy score was around 0.80, which is pretty good, but the precision and recall scores were at 0, which is not very good at all. What this says is that our model, while fairly accurate, is not precise at all. This equates to a very weak model, and AUC score (Which was at 0.6) helped us confirm that.

By far, the best model that I used was the random forest model. After running the model, I received an accuracy score of 0.744. I was even able to visualize the feature importance score for all of the individual features and take out some of the less important ones. This helped me adjust my model and end up with an accuracy score at 0.776!

So why were the models overall fairly weak? The most likely explanation is that our data only explains a small percentage of the variance of the response data around its mean because there is more valuable data out there than can help explain more of that variance. The NFL combine, for example, is an annual scouting event where NFL scouts measure physical attributes of players (Speed, vertical, strength, height, weight, etc.). Perhaps this information is more important in helping to predict which quarterbacks will become NFL greats.

If I had more time with this project, I would explore combine data along with other data sets and use them in conjunction with college statistics to see if that could help improve the accuracy and precision of the models I used and to see if that could help to better predict which quarterbacks will become great ones in the NFL.