Predicting Quarterback Success in the NFL - Statistical Data Analysis

Exploratory and Statistical Data Analysis was performed to determine the target variable from the NFL passing dataset. This was done by using bivariate analysis to evaluate the correlation of several NFL passing metrics vs NFL win percentage. A Correlation Matrix was made and then 8 NFL passing metrics that had a Pearson coefficient > 0.5 were then related to win percentage by making a scatter plot for each. It was clear that they all had a positive correlation. These 8 metrics were Passer Rating, AY/A (Adjusted Yards per Attempt), ANY/A (Adjusted Net Yards per Attempt), TD% (Touchdowns / Attempts), QBR (Total Quarterback Rating), NY/A (Net Yards per Attempt), Y/A (Yards per Attempt) and Completion Percentage.

The next step in this analysis was to determine what the target variable for determining NFL success is. Each of the 8 metrics were put into a correlation matrix with the available college quarterback data between 2009-2019. After doing this, YPA (Yards per Attempt) was chosen as the target as it had the most consistent Pearson coefficients across all the college passing metrics as seen in the table below.

r-value	AY/A	Att	Cmp	G	Int	Pct	Rate	TD	Y/A	YPC	pass Yds	rush Att	rush TD	rush Yds	Y/G
Y/A	0.11	0.14	0.18	0.17	0.14	0.2	0.17	0.26	0.11	0.27	0.17	0.21	0.2	0.23	0.036

After choosing YPA, the NFL Average YPA between 2009-2019 was calculated to be 7.1. This will be the threshold of success for a quarterback in the NFL. To simplify the analysis, a categorical column was made in the dataset that had a value of either 1 for above the NFL YPA average or 0 for below that average. Of the 63 quarterbacks that had more than 100 passing attempts in the NFL and drafted between 2009-2019, 40 had below a 7.1 average YPA and 23 had at least a 7.1 average YPA.

Next, the distributions were examined by looking at distribution plots and box plots. They were made by grouping the quarterbacks that were above and below the average and seeing the distributions of each metric with each group. This provided a great visualization that showed most passing metrics having a higher mean value for the quarterbacks with a greater than NFL Average YPA of 7.1. This difference shows that these metrics translate to NFL YPA.

One outlier that was quite apparent was that of Kyler Murray's college Passer Rating and AY/A. These were the highest overall in the whole dataset of quarterbacks used. His NFL YPA is 6.9, which is quite close to 7.1. This outlier will be kept for now as his NFL YPA is close to the average and therefore his college statistics are relevant.

Bootstrap Inference was then used to examine the features further by looking at the subgroups in each. The p-values of the difference of the means for the subgroups was calculated and the following was found.

AY/A_College	0.0799
Att_College	0.0679
Cmp_College	0.0415
G_College	0.0055
Int_College	0.0578
Pct	0.0377
Rate_College	0.0328
TD_College	0.0049
Y/A_College	0.0555
Yds_College	0.0196
Y/G_College	0.6254

When looking at the p-values, it can be seen that most of our features are right around the conventional threshold of 0.05. We will keep all of these except for Y/G_College, which is not very close to the threshold. The rushing statistics will also be dropped as it does make much sense that these would influence the YPA of an NFL Quarterback.

Next, a Seaborn pairplot was made to examine the relationships amongst our features. It was evident that several had strong correlations with each other. This will be explored further when selecting the features of our machine learning model. It should also be noted that both the continuous target variable (Y/A_NFL) and categorical target variable (above_nfl_ypa) were retained for later to see which approach works best.

This leaves us with 10 features moving forward that include:

1. AYA_College
2. Att_College
3. Cmp_College
4. G_College
5. Int_College
6. Pct
7. Rate_College
8. TD_College
9. Y/A_College
10. Yds_College

It is expected that several of these features will be dropped for our model. The next step will be to explore difference machine learning models to see which provides the best prediction for quarterback success in the NFL.