Predicting Quarterback Success in the NFL - Data Storytelling

Exploratory 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 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.

A rule of thumb to consider for this analysis is having at least 10-20 observations per exploratory variable used in the model. Since there are 63 observations and 15 features/variables, several needed to be dropped. In order to determine which features to drop, multicollinearity of the features was investigated by making a Seaborn pairplot. From the pairplot, it was quite apparent that multicollinearity existed when looking at the scatterplots between various features. To go even further, the VIF(variance inflation factor) of the variables was calculated. The initial VIF can be seen in the table below.

Constant	3435.4
Y/A_NFL	2.9
AY/A_College	377.6
Att_College	484.6
Cmp_College	427.4
G_College	48.6
Int_College	23.7
Pct	44.6
Rate_College	431.2
TD_College	73.8
Y/A_College	164.1
YPC	4.6
Yds_College	262.3
rush_Att	10.9
rush_TD	10.2
rush_Yds	15.6
Y/G_College	17.4
above_nfl_ypa	3.1

These high VIFs make sense as many of these metrics are calculated with other ones. The need to drop features is obvious and this will be determined in the next phase of the project.