Predicting Quarterback Success in the NFL

Jacob Ieyoub
Springboard
Houston, TX
jieyoub@gmail.com

***Abstract* - The success of an NFL quarterback is hard to predict and it is evident when looking at the poor historical performance of teams selecting quarterbacks in the NFL Draft . There are many factors that play into the outcome of the career of an individual quarterback. If it was as easy as going after the biggest player with the strongest arm, then Jamarcus Russell would be in his thirteenth NFL season rather than fizzling out of the league after thirty-one games in three years. It is one of the most important positions in all of pro sports so to be able to predict if a college quarterback will be successful in the NFL would lead to better decision making when building rosters, provide struggling franchises with hope and ultimately lead to a better product on the field across the league. The goal of this project is to use college quarterback statistics to create a model for better predicting quarterback success by defining a threshold for what we will consider success.**

1. Introduction - The inspiration for this project is this video from The Ringer's Ryen Russillo. A similar study from FiveThirtyEight's Josh Hermsmeyer is referenced for some direction, but since this project is using open source data the features are different. NFL is king and to be able to find a franchise quarterback is the first step in the path to winning. Many owners and general managers in the league would find a model that has success in discerning which quarterback prospects will translate to the NFL quite useful on draft day. Not only would it give confidence in who they are selecting but it could even lead to situations where they are able to fill other needs earlier in the draft knowing an undervalued quarterback that the model predicts will be good will be available later. The data used for this project was collected from pro-football-reference.com and sports-reference.com/cfb. It consists of passing NFL and College passing data from 2009 to 2019. The approach was to determine our criterion of success in the NFL and then build a model that uses that specified criterion as our target variable and college passing data as the features.

2. Data Wrangling - To begin, the NFL data was concatenated into a single DataFrame. Once concatenated, the DataFrame was subsetted to only contain players with more than 100 passing attempts.. The data was formatted by season therefore there were duplicate names for players that played in multiple seasons. The "Player" columns of these datasets contained different entries that included the following: "Name*\PlayerCode", "Name/PlayerCode" and "Name*+\PlayerCode". Since these player codes did not correspond to the player codes used in the college data, the Player column was split to only contain the player names.

In this DataFrame there is a "QBrec" column that has the starting record by season for each player. The entries in this column contained record data that was incorrectly put into a date format. This included "mm/dd/yyyy" and "mm-dd-yyyy." The wins and losses were determined from these values and then a "win_pct" column was created. The next step for this data was to use .groupby() aggregated by average to determine the career averages for each player. For columns such as 'Int', 'TD', 'Yds', 'G' and 'Sk', the career totals were used to replace the career seasonal average that was output by .groupby(). Once this was done, several columns that are not needed for the scope of this project were dropped. The resulting dataset was 132 rows x 20 columns.

For the college passing data, the data was concatenated into a single DataFrame and as with the NFL players, the college players had to be split from their player code. Like the NFL data, the necessary career totals were determined and this replaced the career averages calculated from using .groupby() aggregated by average. The resulting DataFrame is 635 rows by 14 columns.

After creating these two DataFrames, they were exported into a .csv file. In a separate Jupyter Notebook, an inner merge was performed on the "Player" column to obtain a DataFrame that consists of all the quarterbacks that were drafted between 2009 and 2019. This DataFrame will be used for our model and is 63 rows x 35 columns. It consists of NFL career and college career statistics.

3. Exploratory Data Analysis - The data was first explored and analyzed to determine the target variable from the NFL passing dataset. This was done by using bivariate analysis to evaluate the correlations of several NFL passing metrics vs NFL win percentage. A correlation matrix was made and then eight 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.

Y/A_NFL Correlations	r-value
AY/A_College	0.11
Att_College	0.14
Cmp_College	0.18
G_College	0.17
Int_College	0.14
Pct	0.2
Rate_College	0.17
TD_College	0.26
Y/A_College	0.11
Yds_College	0.17
Y/G_College	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. This will not be the endgame of predicting success, but will serve as the threshold for this project. 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, forty had below a 7.1 average YPA and twenty-three 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 subgroup. 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.

Diff of Subgroup Mean	p-values
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

4. Conclusion - 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 different machine learning models to see which provides the best prediction for quarterback success in the NFL. Additional features may be explored if more open source data is found. One the model is built, it will also be intriguing to look at quarterbacks that are predicted to meet the threshold but do not to see the external factors that may have affected their potential (i.e. historically bad franchises, poor coaching, coaching instability, etc.).