In-Depth Analysis - Predicting Quarterback Success

For the modeling phase of the project, several machine learning techniques were applied to the project dataset. After analyzing the dataset, a hypothesis was developed that the success in the NFL of a college quarterback could be predicted accurately enough for use when drafting a quarterback by only using basic college passing statistics. Two different target variables were analyzed in the modeling to see if there was more success modeling to predict one vs. the other. The two target variables were YPA(Yards Per Attempt) and Passer Rating. The threshold of success for both of these target variables was a quarterback having a YPA or Passer Rating that was above the 20 year NFL average.

The dataset consists of quarterbacks drafted between 1998-2019. The features include Adjusted Yards Per Attempt, Attempts, Completions, Games, Interceptions, Completion Percentage, Passer Rating, Touchdowns, Yards Per Attempt and Passing Yards. The target variable was a binary classification where 1 denotes above the NFL average and 0 below the NFL average. Yard Per Attempt was the first target variable used.

Yards Per Attempt - As discussed in the Statistical Analysis, YPA was selected as the target variable because it showed the highest consistent correlation across the features along with a relatively significant correlation to NFL QB Win Percentage. The scikit-learn machine learning library was used to begin modeling the dataset. Starting off, a baseline classifier was created using the DummyClassifier class. This is a classifier that uses simple rules to gauge how accurately a model can be predicted by guessing. For YPA and all 10 features, the DummyClassifier yielded an accuracy of 60.4 % using the "most-frequent" strategy. Therefore, the goal was to have a model that would give a higher accuracy than 60.4%. A Logistic Regression model was the first model applied to the dataset. With all ten features, there was decent success with this model. Initial results using 5 fold cross-validation saw a mean score of 69.7%. This, however, needed to be further studied as from our statistical analysis we know there is multicollinearity between several of our features. By calculating the Variance Inflation Factor of our features, the dataset was reduced to two features, Adjusted Yards Per Attempt, Interceptions, Games and Touchdowns. This was with a threshold of 50.0 for the VIF. Logistic Regression was used again and the accuracy was 67.4%, which is a solid score. Next, the score of the test data was determined to be 50%. This is obviously very poor. Therefore, a DecisionTree classifier was created for our dataset to see if there is more success modeling the dataset. This provided a 5-fold cross-validation mean score of 53%. The model score on the test data was then determined to be 59%. This is much better than what Since a very good model with YPA as the target variable could not be built, Passer Rating was used as the target variable to see if there would be more success.

Passer Rating Passer Rating became the new target variable because it had the highest correlation amongst the basic NFL passing statistics to NFL QB Win Percentage. After seeing that the consistent correlation in the features to YPA did not lead to a good model, these correlations were disregarded. As with the YPA modeling, a DummyClassifier was created and this had an accuracy of 66.6%. Logistic Regression was then used with all ten features, ignoring multicollinearity at first, and the model predicted success with a 62% accuracy. The features were reduced like before and with the four features a new Logistic Regression model was created. The 5-fold cross-validation mean score of the Logistic Regression model with these four features was 68%. This is a decent result and since this was better than the results from the modeling with YPA as the target variable, it appeared that passer rating is a better target variable. This was explored further by building a DecisionTree model. With no parameter tuning, the DecisionTree model produced poor results with a 5-Fold cross-validation mean score of 48.8%. By setting the hyperparameters max_depth = 2 and criterion = ‘entropy’, the 5-fold score improved to 61%. The test data scored 77.3 %. This result was the best of any of the modeling so far. To verify this high accuracy, the model was run 100 times. The mean of the scores for each player was then calculated and if the mean was above 0.5 then the predicted result was rounded to 1, or above the NFL Average Passer Rating. If the mean was below 0.5 then the predicted results were rounded to 0, or below the NFL Average Passer Rating. The resulting DataFrame can be seen in the table below.

Player	predict	actual	NFL Pass Rating	True / False	Player	predict	actual	NFL Pass Rating	True / False
Patrick Mahomes	0	1	109.6	False	**Jacoby Brissett**	0	1	84.8	False
Deshaun Watson	1	1	101.4	True	**Cody Kessler**	1	1	84.8	True
Lamar Jackson	1	1	98.9	True	**Mason Rudolph**	1	1	82.0	True
Dak Prescott	1	1	97.0	True	**Sam Darnold**	0	0	80.9	True
Gardner Minshew	0	1	91.2	False	**Kyle Allen**	0	0	80.0	True
Nick Mullens	1	1	90.8	True	**Jeff Driskel**	0	0	78.8	True
Drew Lock	1	1	89.7	True	**Josh Allen**	0	0	76.6	True
Jared Goff	1	1	87.9	True	**Dwayne Haskins**	0	0	76.1	True
Daniel Jones	0	1	87.7	False	**C. J. Beathard**	0	0	75.5	True
Kyler Murray	1	1	87.4	True	**David Blough**	0	0	64.0	True
Baker Mayfield	1	1	86.2	True	**Josh Rosen**	0	0	59.4	True

The feature importances were also calculated and can be seen in the table below.

Completions	0.41
Completion Percentage	0.35
Yards Per Attempt	0.24

Conclusion The results provide some interesting takeaways. First of all, there were zero false positives. This means if it did predict a player to be above the NFL average passer rating, then that player was actually above every time. There were, however, four false negatives. This included Gardner Minshew, Daniel Jones, Jacoby Brissett and most notably Patrick Mahomes. Obviously, the model missing on Patrick Mahomes is unforgivable but this follows the old adage that all models are wrong and some are useful. It should also be reiterated that this was only using basic college passing statistics as the features. Many factors play into the success of a quarterback's success but the results using these basic features are promising. Now, being above the NFL Average Passer Rating does not necessarily mean a quarterback is successful. It does however mean that a quarterback has the ability to be productive. It is up to the front office and coaches to optimize that productivity with the right offensive system, productive offensive weapons, a good offensive line and a good defense. Finding a good quarterback is just the tip of the iceberg.

There are a few things to consider when looking at the feature importances. It can be safely said that the number of completions of a college quarterback does tell us whether or not they will be good in the NFL. To me it shows that experience is important as I am sure most quarterbacks that play more games will have more completions. The importance of completion percentage and yards per attempt is somewhat expected as an NFL quarterback needs to be accurate and efficient.

Further Considerations As mentioned above, several factors affect a quarterback's success. Some of the considerations that would be beneficial for improving the model would include metrics that are not freely available. There are two ways to look at this project. The perspective of a fan or from the perspective of a general manager. This could also be looked at as pre-draft or post-draft. Pre-draft considerations and potential features would include more advanced passing data such as Tight-Window Throws, Average Depth of Target, Completion Percentage on Deep Throws and Deep Throw Rate. Other factors could include the strength of the defenses these quarterbacks faced and also incorporating combine results to see if they are significant.