Final Report - NFL Quarterbacks

Introduction

What aspects should quarterbacks focus on to ensure more touchdowns for their football team? As a recent fan of football, I've been invested in the statistics regarding quarterbacks. Although football is a team sport, the quarterback is debatably the most important player on the team. Due to this insight, I was curious towards which quarterback related factors affect the number of touchdowns a team receives per game. In this analysis, we will look at variables such as number of yards thrown, number of sacks, pass completion percentage, etc. It's important to note that the data used includes all in-season football games in 2014 and 2015. Multiple methods were used in this analysis because different models provide different insights. The two models I chose were used for two different purposes: clustering and feature importance. Interpretability for different models can also help the analyzer conclude multiple things from the data.

Exploratory Data Analysis

From the EDA, the first important thing I noticed was that the average number of touchdowns per game per team was 1 touchdown. This seems extremely different from football nowadays where teams are easily scoring 2 or 3 touchdowns per game. Football in 2023 could be much more fast paced than football in 2014 and 2015, however, I believe the factors pertaining to a quarterback helping his team score a touchdown are the same. I created a heat map to look at the correlation between variables and saw that yards was highly correlated with our response variable, number of touchdowns. Other variables that were also highly correlated with number of touchdowns were attempts, completions, and the longest throw. After seeing this correlation, I was intrigued to build a random forest model and see if this correlation lines up with the feature importance.

Overview of Models Tried

Model	Description	Hyperparameters	Results
K-means	An unsupervised machine learning algorithm that clusters similar data points into groups.	• Number of clusters (n_clusters)	• Silhouette score (-1 to 1) • Davies-Bouldin score
Random Forest	An ensemble learning method used for regression and classification tasks.	• Number of trees (n_estimators) • Random number generator seed (random_state)	• Accuracy (0 to 1) • R-squared (0 to 1)

Discussion on Model Selection

The tree-based, ensemble random forest model performed better than the K-means model when comparing results statistics. The K-means model had a silhouette score of 0.51 and the random forest model had an R-squared value of 0.91. One of the reasons why I chose a random forest model is because it's known to combat overfitting even with a high number of trees. However, for K-means one difficulty is choosing the number of clusters. This can be tricky because the user is telling the algorithm how many clusters to find instead of the algorithm finding how many clusters it thinks best fits the data. The only thought I had towards clustering these quarterbacks was grouping them by skill level, but it seemed the higher the cluster number was, the lower the silhouette score was, which is why I chose 3 clusters. It had a better silhouette score with clusters that made sense. The three groups it classified seemed to be related to the number of touchdowns per game. The first group consisted of quarterbacks who played in games where no touchdowns were scored. The second group had quarterbacks who played in games where 1-2 touchdowns were scored. The third group had quarterbacks who played in games where at least 2 touchdowns were scored.

Other models such as a Naive Bayes classifier or Support Vector Machines (SVM) weren't suitable for this analysis because certain assumptions weren't met and there could've been limitations in capturing relationships within the data. With Naive Bayes, independence between features is assumed, which is not the case with most if not all of the features in the quarterback dataset. For SVMs, the algorithm won't perform well if the data isn't linearly separable in higher dimensions. The relationships between the quarterback features might not be linear, causing problems with an SVM model.

Detailed Discussion on Best Model

The best model in this analysis was the random forest model with an R-squared value of 0.91. I altered the number of trees as well as the random number generator values, however it didn't seem to make that much of a difference for this dataset. One of the main reasons why I chose a random forest model is to analyze feature importance. From this analysis, it looks like yards is the most important feature with a feature importance value of 0.44. Yards per attempt was the next most important feature with a feature importance value of 0.12. This value is much lower than the value for yards, indicating that number of yards really is the most important variable for quarterbacks to ensure their team scores touchdowns.

Conclusion

In conclusion, the random forest model performed the best. From this model, I was able to determine feature importance, which lined up with the heat map correlations in the exploratory data analysis indicating that yards is the most important variable. This model also did well at predicting, which we know from its R-squared value of 0.91. However, we shouldn't forget about the K-means model. From this model, we can see which quarterbacks were part of high-scoring, medium-scoring, and low-scoring teams in relation to the number of touchdowns.

One future recommendation is to do a deeper analysis of each of the 3 clusters from the K-means model. Analyzing the feature ranges of the 3 groups could tell us more about how the algorithm determined which quarterback belonged in which group. Another recommendation is to use the random forest model to predict how many touchdowns a team will have nowadays in 2023. Doing this could assess the accuracy of the model overtime.