How NFL Quarterback Performance Leads to Points Scored

**Motivation**

Since the introduction and success of the Moneyball strategy implemented by Billy Beane and the Oakland Athletics, data analysis has been a hallmark of the sports industry, and its influence has only increased. Twenty years after the initial implementation of Moneyball, we can see sports organizations go well beyond the surface level box-score statistics, and now utilize teams of data analysts and scientists to ensure maximum performance on the field. Due to the money and influence it possesses on American culture and interest, hardly any other sport utilizes data to the extent of American football and the National Football League (NFL), and no other position in that sport receives as much attention and analysis as the quarterback position.

The quarterback of a football team is by far the most important player. Numerous studies have been done to try and evaluate quarterback performance so that franchises can use that to draft and develop the best quarterback possible, which in turn would lead to success for potentially decades. Due to the importance of getting a franchise quarterback, there is no shortage of research done regarding this topic. In a 2014 analysis by Steven Hoerning, Bobak Moallemi, and Matthew Wilson, they conducted a predictive analysis to predict the number of touchdown passes a quarterback would throw in an upcoming game; while higher than ESPN’s forecasts accuracy of 33%, it still operated at only a 38.6% accuracy rate (Hoerning, 1). People have also built models to try and predict the success of whether a quarterback will be a draft bust or not, with a 2017 report by Amit Patankar and Aseem J Monga using data from college to predict draft busts with a 73.7% accuracy (Patankar 5). The model then predicts the results of the 2018 quarterback draft class, albeit to mixed results. It successfully predicted Lamar Jackson to be an NFL-ready quarterback, which several draft experts and NFL franchises did not agree with; however, it did also predict Sam Darnold and Josh Rosen to be NFL-ready quarterbacks, which ultimately was not correct for either (Patankar 5). Other projects have been created to develop alternative methods to evaluating quarterbacks, such as Matthew Reyers’s inclusion of parameters to evaluate the impact of mobile quarterbacks, and him choosing to later focus on all the possible decisions a quarterback can make to evaluate. It leads to interesting conclusions, as it shows how Alex Smith, despite having a good quarterback rating (QBR), may not be playing up to that rating in terms of value (Reyers 12-16). There are scores of more reports and analyses attempting to evaluate and predict excellent quarterback play, however, I was not able to find research addressing the exact topic I will embark on. The problem I am attempting to address is how the in-game performance of quarterbacks can predict the number of points a team will score in a game, and I will attempt to find out the most significant and least significant variables that can predict this.

**Problem Framing**

As stated previously, the problem I will address is the prediction of points scored using valuable quarterback statistics, and the most impactful variables that contribute to points scored. I found this to be a regression problem, one where finding out the most impactful variables to create the best model is paramount to the outcome of this project. My primary data source is a spreadsheet illustrating all single-game quarterback performances from 1996-2016, with some of the most important statistics being passing yards, passing touchdowns, and interceptions, and it possesses a response variable in team points scored. To design the data for usage, cleaning was necessary; the specifics will be addressed in a later section. Some easily determined inputs are that passing touchdowns contribute quite a bit towards how many points a team will score; this makes sense as that variable directly contributes to greater amounts of points scored. To contrast, it is easy to see that aside from a few gunslinger quarterbacks like Brett Favre, the number of interceptions thrown influences the number of points scored, albeit in a negative way. This makes sense, as interceptions are literally impossible to score from, and it ends the offensive possession. There is a big ability to learn from this data set, as other stats present themselves with much less certainty when it comes to predicting points scored. Using tools like XGBoost and Random Foresting, we will find out how much impact other variables have in predicting points scored. There is potential for bias to be shown in the data. A primary source can come from the people who created the data, which is often done by home-team box scorers. This can lead to certain stats like passing yards and longest pass to be a little more inflated than they actually are so that the QB’s stats can be elevated. Another source of potential bias can come from the year of when the data was taken, as the NFL over the time range of the data set has become far more willing to design their offenses around the pass compared to the run. This may skew some of the results towards the later years of the survey, when quarterback play became a lot more prioritized.

**Data Overview**

The data set I will use comes from Kaggle, and it is a game-by-game list consisting of quarterback stats from 1996 to 2016, which makes the data set 10,756 points long. The data set includes fourteen different variables, with three being categorical and eleven being quantitative. The categorical variables give the quarterbacks’ names, whether they played in a home or away game, and the year that the game took place. The quantitative variables give data points for important quarterback counting stats, like completions, attempts, passing yards, passing touchdowns, interceptions, and additional points that are typically used when evaluating quarterback performance. One quantitative measure however addresses the team as opposed to the individual quarterback play, which is game points, which measures how many points a team scored in that game. This is different compared to the other quantitative variables, as team points scored also factors in the running game and non-offensive scores, like kickoff returns and interceptions returned for touchdowns.

Some notable summary statistics I decided to examine were passing yards and quarterback rating (QBR), as those variables are often thought of as being the most important. Based on the density plots of both graphs (in html file), we can see the median of passing yards in a game is a little over 200 yards and the typical QBR is a little under 100. I also decided to plot a scatterplot between passing touchdowns in a game and points scored to see if there was a relationship that can be examined more in depth. According to the scatterplot, we can see a clear upward trend between the two variables, meaning that the relationship between passing touchdowns and points scored is a positive relationship, and showing that more passing touchdowns typically lead to more points being scored by the team.

The data set needed to be cleaned prior to analysis. The first obstacle was ensuring that there were no listed games where a quarterback threw less than 14 pass attempts. I chose to do this because oftentimes, backup quarterbacks will enter games and throw a very small number of passes. Depending on the outcomes of those small number of attempts, it can lead to extreme results in the data. In addition, I thought 14 would be a good cutoff because according to Pro-Football Reference, the largest source of public data for the NFL, in years post-1978, a quarterback must average 14 pass attempts over a 16-game regular season to qualify for most season-long quarterback ratings. The data was initially over 13,000 points long; however, the 14-attempt cutoff lowered the count by over three thousand. There was also an initial issue with the “long” column, where there would be accurate measurements, followed immediately by the presence of a “t” aside the number. Once those figures were removed from the column, the data was ready for analysis.

**Methods**

Due to my problem being regression-based, I decided that using both Random Forest and XGBoost be the best methods to solve the presented problem. Both situations required splitting the data into training and test data; using the set.seed function and random sampling input, about 20% of the data was used for test data and 80% of the data was used for training data. The split and the samples that were sorted into each category would remain the same throughout the project. To see the effectiveness of the model both pre- and post-tuning, its Random Mean Square Error (RMSE) measure was examined, with smaller measurements of RMSE resulting in more effective models.

Before any complicated analysis, I tried a simple linear regression model with game points being the response variable and all other variables being used as predictors; once that was completed, I took the RMSE of the model to serve as a benchmark for the remainder of the analysis. The Random Forest Model started off with a simple bagging model with mtry (number of variables randomly sampled at each split) equaling 12 and the number of trees equaling 200. After the initial bagging model was created, a tuning process was conducted to get both the optimal number of trees and the optimal node size that results in the smallest RMSE measurement. The values of the number of trees that were evaluated were 100, 500, and 1000, with 200 already being measured in the initial bagging model, and the node sizes that were tested were 50, 100, 200, 500, and 1000. Once the best combination was found, I then extracted the corresponding RMSE value and ranked the importance of the variables. Likewise, similar methods were used during the XGBoost process. Prior to any deep analysis with the XGBoost, the test and training data had to be converted to matrices. Once that was completed, I could train the XGBoost model with the training data and get the RMSE for it. Afterwards, the XGBoost model underwent an extensive tuning process where the optimal value for max depth, minimum child size, eta, gamma, subsample, and column space were identified and applied to the final model, where the RMSE was found for that too and the variables were ranked in terms of importance.

**Results**

The RMSE for the simple linear regression model was 7.59, which was an adequate amount of error. Ultimately, this simple model could not be used for deeper analysis and finding the most important variables, so further work was done. The initial bagging model with mtry = 12 and ntree = 200 resulted in an RMSE of 6.98, which was a massive increase in model effectiveness; the initial bagging model also possessed a decent fit, as 50.7% of the variance could be explained by the model. Tuning was then used to figure out the model that would lead to the smallest RMSE measure, and the parameters that were tested were the number of trees and node size. Regardless of which value I used for the number of trees (100, 200, 500, 1000), the RMSE value either stayed at 6.98 or 6.99, with 200 and 500 equaling 6.98 and 100 and 1000 equaling 6.99. Because of the lower RMSE value, either 200 or 500 trees would have sufficed for usage in the final bagging model; I ultimately decided with 200 trees because I was aiming for as low of a number of trees as possible while keeping the error at its lowest value. After the number of trees was decided on, the tuning for node size was completed, with 100 trees producing the smallest measure of RMSE at 6.91. Once the correct combination was figured out, it was then analyzed to see that it possesses an RMSE equaling 6.91, and the importance of each variable was plotted. While there were possible options for the measure of importance, I decided to use %IncMSE because it is a more reliable and less biased measure than IncNodePurity. The rankings showcased passing touchdowns (td) to be the most important variable when it came to predicting game points, with QBR (rate) and yards per attempt (ypa) being second and third most important respectively. Smallest loss (loss) and passing yards (yds) closed out the top five, ranking as fourth and fifth respectively.

Once the training and test data were converted into matrices, the initial XGBoost model could be trained, resulting in a RMSE of about 7.34. As with the initial bagging model, the XGBoost model also possessed a lower measure of error than the simple linear regression model. Afterwards, significant tuning was done to the model to create the model with the least error. The first parameters tuned were maximum depth and minimum child weight. According to the data frame that shows each different combination of the two parameters and the resulting RMSE measures, the combination that resulted in the least amount of error was a max depth equaling 15 and a minimum child weight of 1, as it had an RMSE measure of 0.836. The gamma value was then tuned next, where the data frame shows that a gamma value of 0.15 results in the smallest RMSE measurement of 7.51. Column space and subsample were the next two parameters tuned, where the data frame showed the combination of subsample equaling 0.9 and column space equaling 0.6 resulting in an RMSE measure of about 7.365, which was the lowest in the data frame. Eta was the final parameter measured, where five models, each consisting of a different eta parameter (0.3, 0.1, 0.05, 0.01, 0.005) were created and then plotted to see which eta value consistently has the lowest RMSE value. According to the graph, the eta value of 0.1 consistently possessed the lowest RMSE value, meaning it would result in the lowest error for the model. With all the parameters properly tuned, the RMSE for the model was extracted and the variables were ranked in terms of importance. The RMSE of the model was about 7.28, making it a better model than the initial XGBoost model. The variable importance graph showed a similar output in terms of the three most important variables, with touchdown passes, QBR, and yards per attempt again ranking as first, second, and third respectively. However, the importance graph showed passing yards (yds) and passing attempts (att) as the fourth and fifth most important variables respectively, creating an interesting difference between the Random Forest and XGBoost processes.

**Discussion**

The most important results from both the Random Forest and XGBoost models were that touchdown passes, QBR, and yards per attempt are the most important quarterback statistics that lead to points scored (can be found on corresponding .html file). These variables make sense as to why they would be the most significant, as touchdown passes are a type of scoring play that football teams can do, QBR is an overall measure of how well the quarterback played, and a high yards per attempt value indicates that the quarterback is gaining a lot of yards whenever he throws the ball. Ultimately, it means that the better the quarterback plays and the better his statistics are for these three categories, the more points his team will likely score. Teams could use this to implement highly efficient passing attacks, as giving the quarterback easy throws that net several yards an attempt will allow the offense to score in productive amounts.

An interesting insight that could be interesting to analyze further is the discrepancy between the Random Forest and XGBoost model’s ranking of pass attempts, as while the XGBoost model ranked it in the top five in terms of importance, the Random Forest model ranked as the second-least important variable. This is interesting as the interpretations of each graph essentially lead to opposite conclusions, as the XGBoost likely will state that throwing the ball more times is more crucial to score points than the Random Forest analysis, which would lead to different outputs in a practical setting. For example, a coach who bases his passing offense off the Random Forest model will not prioritize the amount of times the quarterback will throw the ball and will likely run a more balanced offensive attack compared to a coach who bases his offense off the XGBoost analysis, who will likely use a very pass-heavy scheme to try and score points. Due to the Random Forest analysis having a smaller RMSE output and therefore a smaller amount of error, using the output from the Random Forest would likely lead to a more effective output.

Another interesting aspect of the analysis is that both the XGBoost and Random Forest models highlight the lack of importance of other stats like completions and sacks given up possess when it comes to team scoring. This can make sense, as an increase in sacks allowed can be the result of simply choosing to pass the ball more often, and a higher amount of completions does not necessarily mean that the offense is moving the ball in an efficient manner; oftentimes, a high amount of completions means that the offense is facing a large deficit, meaning that they need to throw more often in an attempt to try and get back into the game. These scenarios refer to the importance of being efficient with passing plays an offense, rather than choosing to throw the ball in an inefficient manner.

**Conclusion**

Based on the analysis displayed and discussed in this report, the most important variables of quarterback play that lead to team points scored are touchdown passes, quarterback rating, and yards per attempt, as these three variables highlight the importance of the efficiency of passing plays as opposed to increasing counting statistics. If given more time and resources, a possible next step could be to examine if this was the case for NFL games played prior to 1996, where offenses relied a lot less on the pass and far more on the run game. Having a wider range of seasons can also showcase the impact of certain rule changes that allowed the game to be more quarterback friendly, which would also be a highly interesting analysis to conduct. Doing a categorical analysis with figuring out which quarterback statistics are the most important for wins and losses would also be a logical next step, as it could lead to interesting analysis on how truly important quarterback play is for winning consistently in the NFL; combined with the expansion of the number of seasons, it would allow for general quarterback importance to be evaluated across the history of the NFL.

**Contribution**

I did this project by myself.

**Bibliography**

Hoerning, Steven, et al. 2014, pp. 1–5, *Forecasting NFL Quarterback Performance*.

Patankar, Amit, and Aseem J Monga. “Projecting NFL Quarterback Readiness Final - Cs229.Stanford.Edu.” *Stanford.Edu*, 2017, cs229.stanford.edu/proj2017/final-reports/5231213.pdf.

Reyers, Matthew. “Quarterback Evaluation in the National Football League.” *Summit Research Repository*, 20 Aug. 2020, summit.sfu.ca/item/20853.

**R-Code**

On a separate .Rmd file and .html file.