Predicting NFL Success among NCAA Quarterbacks from 1999-2019

Matthew Murray

1 Introduction

1.1 Background

In pro football, it is implicitly assumed that the quarterback position is the most important, as he is the focal point of any NFL team's offense. Sports writer and former NFL player Scott Fujita wrote (about quarterbacks): "No other football position can have such an impact on the win-lose outcome of professional matches"(Fujita (2022)). Quarterbacks are also among the highest paid players in the NFL, with an average salary (as of April 2022) of $6.48 million ("NFL Positional Payrolls" (n.d.)), second to only left tackles, the protectors of a quarterback's "blind spot."

Due to the salience of the position, it is not surprising that NFL scouts place a very high priority on scouting NCAA quarterbacks and determining which quarterbacks' college success will translate to success at a higher level. Nonetheless, this task is much more difficult than expected. There has been no shortage of All-American caliber, star college quarterbacks who were lauded immensely for their talent and potential, but failed to fulfill said potential once they reached the pros. One notable example of such a player is Ryan Leaf, whom many refer to as the biggest bust in NFL history (Add Source). Leaf played his collegiate years at Washington State where he finished as a first team All-American and Heisman trophy finalist in 1997. After being drafted second overall in the 1998 NFL Draft by the San Diego Chargers, Leaf was out of the NFL by 2001. He finished his career with 14 touchdowns, 36 interceptions, and an abysmal starting record of 4-17. On the other hand, there are diamonds in the rough like Tom Brady. Brady, who is widely recognized as the greatest quarterback of all time, greatly exceeded expectations when he was drafted 199th overall in the 6th round of the 2000 NFL Draft. Today, Brady's awards and accolades are too long to list. All in all, finding the college quarterbacks who will flourish in the NFL is quite the tall task,

1.2 Dataset Description and Variables

The dataset for this report contains statistics for NCAA Division 1 quarterbacks from 1999 to 2019. The data was scraped from Sports Reference, a US company that runs several sports statistics-related websites, one of which is dedicated to college football. Excel files with the quarterback data were downloaded, converted into comma separated files (CSV's), and read into my RStudio coding environment. Each of the Excel files contained the top 100 NCAA quarterbacks for a given year in terms of passing yards per attempt. Several steps were then taken to clean and tidy the data. The first row was set as each dataset's column names, the unnecessary Rank column was deleted, the columns were renamed to deal with columns that had the same name (more specifically, columns referring to either passing or rushing statistics were labeled as such), and a new column was added to indicate the year that a quarterback played. These steps were taken for all 21 data sets before they were all merged into one dataset. Another issue with the data set that was remedied is that there were multiple entries for a single player, as most (if not, all) college football players play multiple years in college. To solve this problem, I merged the rows referring to the same player and either summed or averaged the variables for said player accordingly. For example, I took the *sum* of the games played variable and the *average* or *mean* of the completion percentage variable for a given quarterback. After these steps were taken, the dataset contains 1,130 observations, or in other words, statistics for 1,130 college quarterbacks. Of these quarterbacks, only 18.9% were selected in the NFL draft.

Additionally, data related to a quarterback's NFL career was imputed into the dataset. These variables were collected in an attempt to find data that defines a player's pro "success," which is admittedly difficult to quantify.

The data was taken from various Wikipedia pages. For example, the variables for *years on an NFL payroll* and *games started* were collected because the most successful NFL quarterbacks are likely to stick around and have longer careers, simply because teams will want their skillsets. The average career length of an NFL player is approximately 3.3 years, while that of an NFL quarterback is 4.4 years ("Average Playing Career Length in the National Football League (in Years)" (n.d.)). Meanwhile, the average career length of an NFL player who plays in at least one Pro Bowl — the NFL equivalent of an All-Star game — is 11.7 years (Mojica (2020)). I also collected data regarding whether or not a quarterback has made at least one Pro Bowl during his career.

This data was originally collected by myself, along with two other Duke undergraduate students, Sofia Carrascosa and Rex Evans, for a project in the Duke Sports Analytics Club (DSAC). However, since the data scraping, cleaning, and imputation took longer than expected, our group did not pursue a project together.

1.3 Objectives

My paper's broader objective is to determine which variables related to a quarterback's college statistics are most strongly associated with NFL success. In other words, I hope to identify which characteristics and statistical trends to look for when predicting whether or not a college quarterback will succeed in the pros. In doing so, I also hope to aptly define and quantify "success," which is admittedly subjective.

1.4 Exploratory Data Analysis

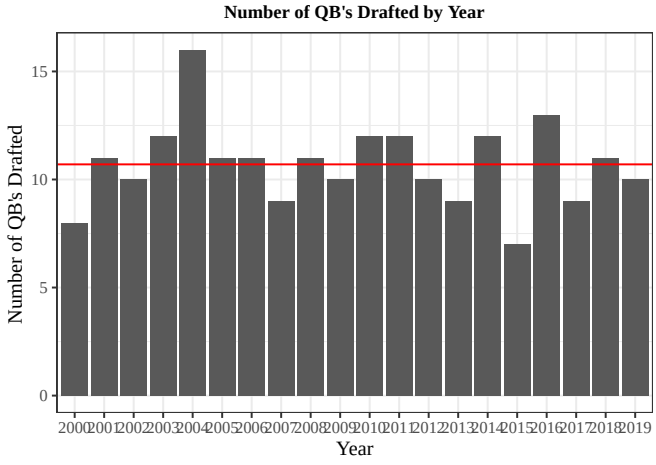


Figure 1

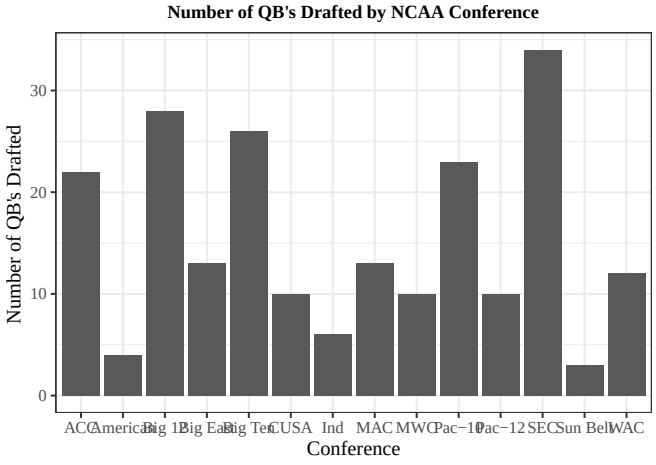


Figure 2

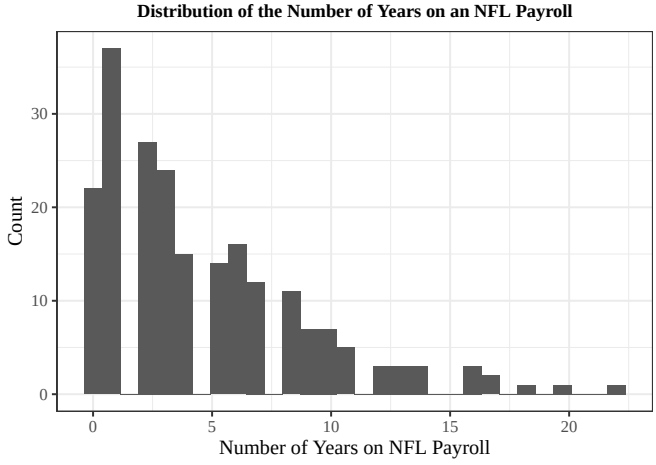


Figure 3

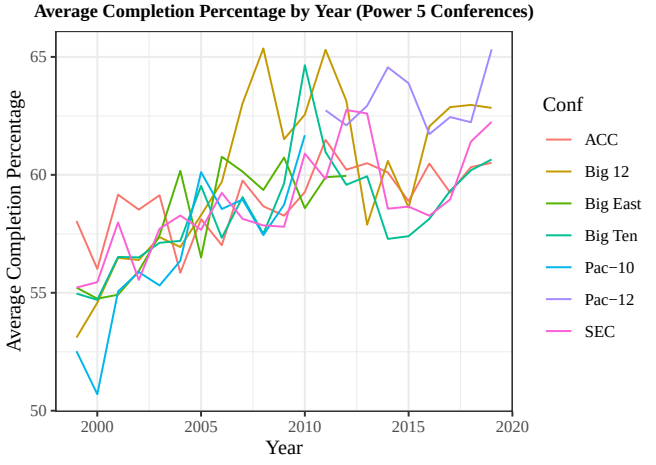


Figure 4

2 Methodology

2.1 Model Selection

The model of choice is a logistic regression model that seeks to predict whether or not a quarterback will make at least one Pro Bowl during his NFL career. The rationale behind this model is that the best quarterbacks in the NFL – which is what teams are looking for when scouting – will likely make *at least one* Pro Bowl during their respective careers in the pros. One can argue that there is subjectivity in the selection of players to the Pro Bowl since players need to be voted in by fans, players, and coaches, but by in large, the *best* quarterbacks seem to make the Pro Bowl at least once in their careers. Logistic regression was chosen due to its high interpretability and suitability for binary outcome variables. Our model was fit using the `glm()` function in R.

Alternative models that I considered involve different linear regression models. One option that I considered is a linear regression model that seeks to predict the number of years that a quarterback is on an NFL payroll. The reasoning behind this modeling choice is that the best quarterbacks will remain in the NFL for longer periods of times because teams will recognize their talent. In other words, this model is assuming that a "successful" NFL quarterback is one who remains in the NFL for a relatively long period of time. Nonetheless, the main limitation of this response variable is that quarterbacks can remain in the league for a long time as a backup, sitting second or third in a depth chart. While these players may appear to be successful because they remain in NFL for a while, they may not be seeing the field often. Furthermore, some quarterbacks, while extremely talented and successful, may not play in the NFL for very long due to concerns over the violent nature of the sport and various other personal considerations, a very notable example being Andrew Luck. Luck, a four-time Pro Bowler who was widely recognized as one of the best quarterbacks in the NFL, shocked the NFL world when he retired from football in 2019 at the young age of 29. He cited injury concerns as the main reason for his retirement.

Another linear regression model that I considered is one that predicts the number of games that a QB starts in the NFL. However, I decided to not use games started as my response variable due to the fact that this number can be influenced by many extraneous factors that my model cannot take into account such as injuries and overall team/organizational success (i.e. players who play for more successful organizations will start more games since they will be making deeper playoff runs).

2.2 Variable Selection

For variable selection, I decided against set model selection criterion such as *AIC* or *BIC*. The use of these criterion often involves greedy variable selection algorithms like backward elimintation and forward selection, and furthermore is a form of post-selection inference. Instead, I used literature and my knowledge about the game of football to conjecture a pool of variables that can potentially serve as useful predictors. Fron there, I used other statistical tools to determine which of these variables work best for the model. One thing to note is that when thinking about predictors, I wanted to choose variables that are stable in the sense that they translate well from the college game from the pro game. The predictors that I considered using in my model, as well as the reasons for their consideration, are discussed below:

- **Completion Percentage** - A quarterback's completion percentage – or accuracy for short – is widely regarded as one of the best barometer's of a quarterback's ability to exceed at the next level. *Bleacher Report* writer and NFL Scout Matt Miller writes: "Despite what some may say, accuracy is one of the few traits that I believe you cannot coach into a quarterback" (Miller (2013)). Serendipitously, accuracy is also one of the easiest traits to quantify compared to other important characteristics like vision, leadership, and pocket presence. However, one weakness of the completion percentage statistic is that quarterbacks can pad their completion percentage through frequent short, safe passes.
- **Average Touchdowns per Game** -
- **Average Interceptions per Game** -
- **Adjusted passing yards per Attempt** - To control for the shortcomings and limitations of completion percentage, I also decided to include adjusted yards per attempt in the model to help account for the difficulty of the passes that a quarterback attempts. *Adjusted* yards per attempt measures the average yardage of

a quarterback's passing attempts while accounting for touchdowns and interceptions. It is a more useful statistic than solely yards per attempt because it gives weight to more impactful passes, namely touchdowns and interceptions. Its formula is as follows: **Pass AY/An = Yds + 20*TD - 45*Int) / Att**. Some models choose to combine these two metrics (completion percentage and yards per attempt) to create a statistic for *expected* completion percentage, but for simplicity, I decided to include the two variables separately.

- **Passer Rating** - A quarterback's passer rating is an all-in-one type of metric that looks to take a hollistic approach to measuring a quarterback's passing performance. The NFL adopted it in 1973 so that there could be a statistic that can referred to when determining who was the best passer in the league in a given season (Kozlowski (2020)). Today, the NFL and NCAA have different formulas for competing this statistic, the principles and underlying reasoning behind each remain the same. The formula for NCAA passer rating is:

$$\frac{(8.4 * YDS) + (330 * TDP) + (100 * CMP) - (200 * INT)}{ATT}$$

- **Average Rushing Yards per Attempt** - A recent trend in the NFL has been teams' increasing affinity towards quarterbacks who are athletic and can run. Today, some of the most elite quarterbacks – Lamar Jackson, Russell Wilson, Josh Allen, and Patrick Mahomes, to name a few – are labeled "dual threat" quarterbacks, meaning that they are effective with *both* their arms and their legs. In the 2020 NFL season, quarterbacks rushed for 8,697 yards, the most in NFL history and 14.3% of all rushing yards (Mahoney (2021)).

- **Power 5 Conference or Not** - Additionally, I chose to create an indicator variable for whether or not a quarterback played in a Power 5 conference. The rationale for creating this variable is that competition is likely more difficult in Power 5 conferences, so including this variable is a way of accounting for the increased level of competition. Today, the Big Ten is the Power 5 conference that is often associated with creating the best quarterbacks, The creation of this variable was inspired by Josh Hermsmeyer; in his *FiveThirtyEight* article, he explains how he adjusts a quarterback's completion percentage for the conference that he plays in to account for the level of competition that he faces (Hermsmeyer (2019)). He uses the example of Russell Wilson, who in 2011 had a raw completion percentage of 73%. However, in the same year, the expected completion percentage for a quarterback in the Big Ten with the same number of passes and the same target depth is 57%, meaning that Wilson's accuracy is 16 percentage points above expected. Today, the Power 5 conferences consist of the following conferences:

1. Atlantic Coastal Conference (ACC)
2. Big Ten
3. Big 12
4. Pac-12
5. Southeastern Conference (SEC)

Nonetheless, in the past, the other conferences in the Power 5 include the Big East and the Pac-10 (which is now more appropriately named the Pac-12). Additionally, I included Notre Dame, while independent and not officially apart of a conference, as among the Power 5 schools. The rationale behind this grouping is that Notre Dame is currently a full voting member of the ACC (as they play 5 ACC schools each year), and for years – before its agreement with the ACC – Notre Dame has played a schedule with a plethora of Power 5 schools due to its annual rivalries with schools like USC, Stanford, and in past years, Michigan.

- **Whether or not the quarterback was under contract (at time of data collection)** - Lastly, I chose to create an indicator variable for whether or not the player was under contract at the time of data collection (Fall 2021). This variable was created to account for the fact that players who are still in the NFL, while exceptional and NFL-worthy talent, may not have as many years on payroll and games started as players who have already finished their careers.

2.2.1 Variable Correlation Matrix

Below is a correlation matrix displaying the correlations between all the variables discussed above:

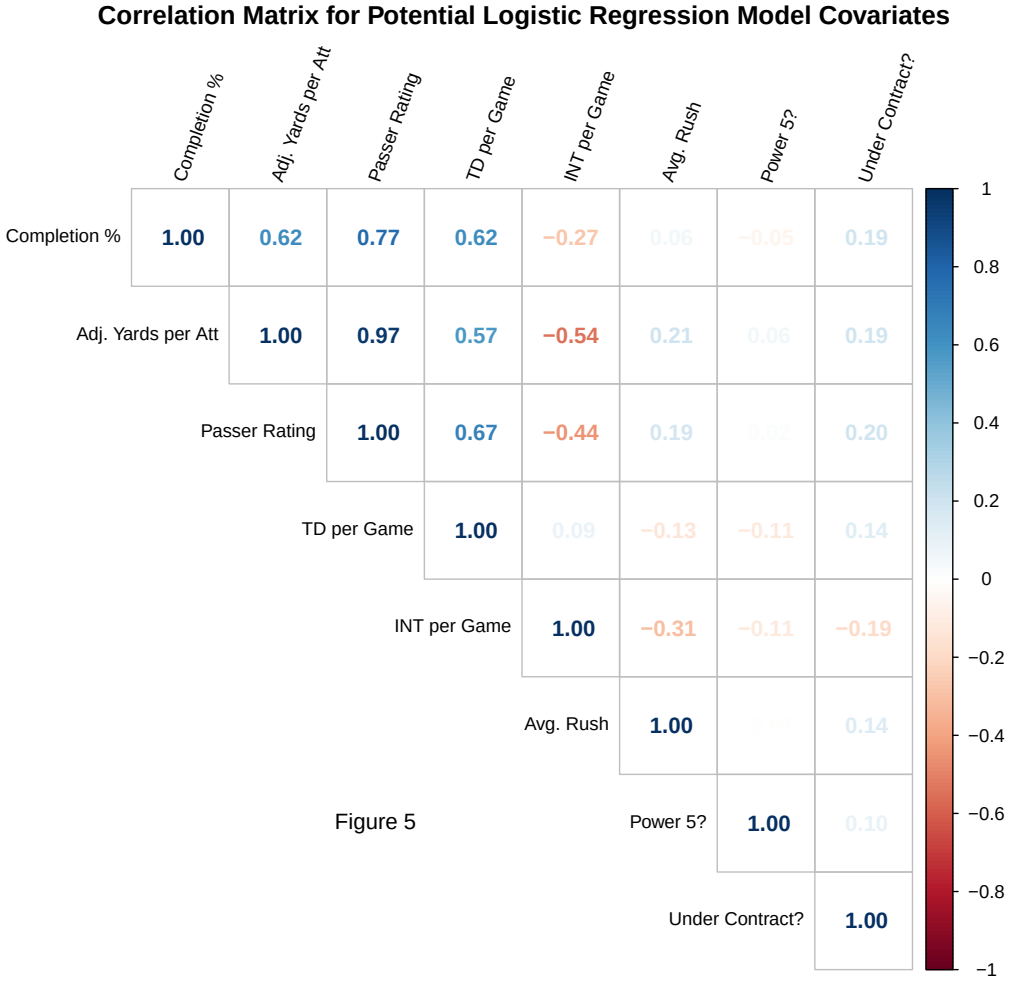


Figure 5

In the plot above, the numbers represent the Pearson correlation coefficients between two variables. The closer the values are to 1 or -1, the stronger the correlation between two variables. The purpose of this plot is to see which variables are most highly correlated with each other. Correlated predictors is a problem in logistic regression because it leads to multicollinearity, which inflates the variance of model parameters, making them unstable and highly sensitive to training data. Additionally, multicollinearity leads to p-values that are higher than they may seem. In other words, when multicollineairity is present in high amounts, results that may *appear* to be statistically significant may not *actually* be statistically significant because p-values are higher than reported.

In the plot above, the highest correlation between two predictors is 0.97, which is the Pearson correlation coefficient between adjusted yards per attempt and passer rating. There is also a high correlation coefficient (0.77) between completion percentage and passer rating. The *positive* correlation coefficient between completion percentage and adjusted yards per attempt is surprising because, as noted earlier, quarterbacks can pad their completion percentage through short, safe passes. Therefore, one would expect *higher* completion percentages to be associated with *lower* yards per attempt, and vice versa. To remedy the problem of multicollinearity, I decided to remove predictors that have a bivariate Pearson correlation coefficient of 0.7 or above, as suggested by *Using Multivariate Statistics, 7th Edition* (Tabachnick and Fidell (2010)). Therefore, I will remove Passer Rating from my pool of predictors.

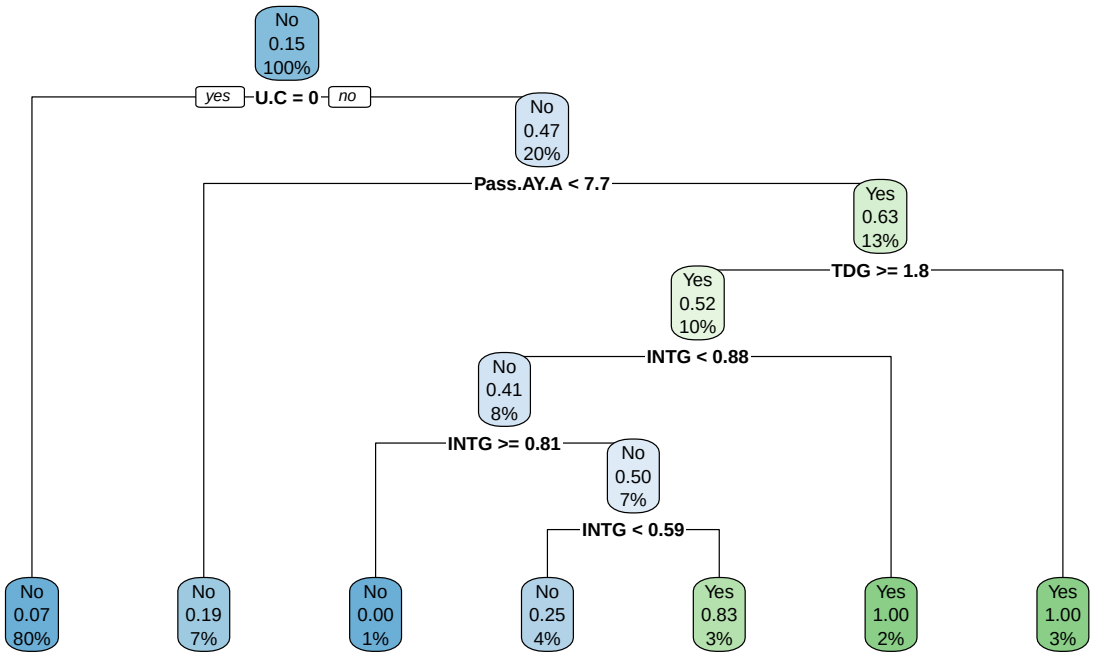
Discuss ways to select features when you have variables that are correlated such as PCA, LASSO regression

2.2.2 Classification Tree

I also decided to create a classification tree to further guide my variable selection process. Classification trees are very useful tools for variable selection, as they rely on little assumptions (especially compared to logistic regression), can give insight into the order of importance of predictor variables, and are useful for examining potential interaction effects. Similar to how an interaction effect allows one to interpret the effect of one predictor on the response based on the value of another predictor, with classification trees, the interpretation of one node is often dependent on the value of another node.

Nonetheless, classifications trees have drawbacks as well. A classification tree itself can be very sensitive to hyper-parameter tuning, meaning that its results may not be as stable and useful as those from logistic regression. Additionally, decision trees are fit using a greedy algorithm, recursive binary splitting. Recursive binary splitting is "greedy" because at each step of the process of building the tree, the "best" split is made at that particular step, rather than looking ahead and making a split that will lead to a better overall predictions. Since we are using a decision tree to explore potential relationships/variables for our logistic regression model, this tendency may mislead us into ignoring the macro-level trends and correlations. Another notable downside of decision trees is that they can easily be overfit to one's training data, particularly when using many predictor variables without a type of model selection criterion. To account for this problem, we decided to use a post-pruning method called cost complexity pruning. At a high level, cost complexity pruning obtains a sequence of trees that are penalized for the number of nodes they have via a tuning parameter. The optimal value for this tuning parameter is then obtained using cross-validation, and consequently the optimal tree is obtained using this value.

I fit the classification tree using the `rpart()` function from the `rpart` package and pruned the tree using the `prune()` function from the same package.



Add interpretations for classification tree Beautify Classification tree

2.3 Model Specification

$$\log\left(\frac{P(\text{Pro Bowl}_i = 1)}{1 - P(\text{Pro Bowl}_i = 1)}\right) = \beta_0 + \beta_1 \text{Average Adjusted Passing Yards per Game}_i + \beta_2 \text{Average Touchdown Passes per Game}_i +$$

$$\beta_3 \text{Average Interceptions Game}_i + \beta_4 I(\text{Under Contract}_i = \text{Yes})$$

2.4 Model Diagnostics

The three conditions/assumptions that should hold for logistic regression are as follows:

1. The log-odds have a linear relationship with the predictors.
2. The data were obtained from a random process.
3. The observations should be independent from each other.

3 Results

Table 1: Logistic Regression Model Output

Variable	Coefficient	95% Confidence Interval	P-Value
Intercept	-7.999	-13.099, -3.33	**0.001**
Adjusted Passing Yards per Game	0.677	0.15, 1.235	**0.013**
Touchdowns per Game	-0.834	-1.974, 0.247	0.137
Interceptions per Game	1.932	-0.654, 4.533	0.141
Under Conract = Yes	2.536	1.638, 3.504	**<0.001**

4 Discussion

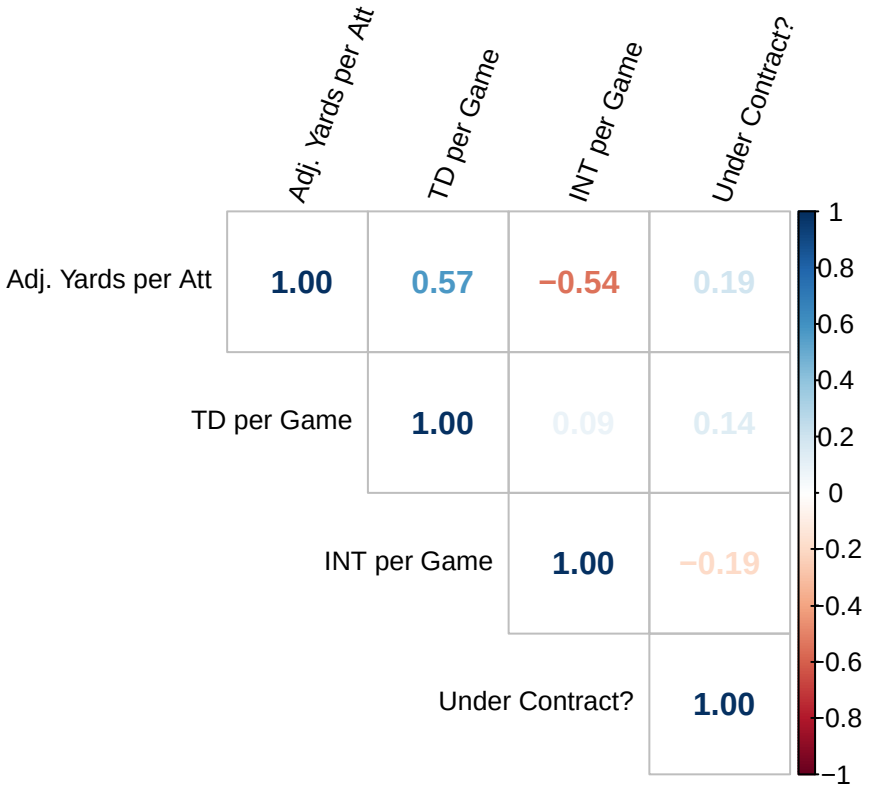
4.1 Conclusion

4.2 Limitations and Future Work

4.3 Summary

Appendix

A1 New Correlation Matrix



A2 Alternative 1 - Modeling Number of Years on NFL Payroll

Variable	Coefficient	95% Confidence Interval	P-Value
Intercept	3.288	-3.004, 9.581	0.304
Adjusted Passing Yards per Game	0.109	-0.644, 0.862	0.776
Touchdowns per Game	-0.617	-2.086, 0.851	0.408
Interceptions per Game	1.283	-1.937, 4.504	0.433
Under Conract = Yes	3.313	1.858, 4.768	**0.000**

A3 Alternative 2 - Modeling Games Started

Variable	Coefficient	95% Confidence Interval	P-Value
Intercept	8.648	-67.702, 84.998	0.824
Adjusted Passing Yards per Game	2.240	-6.897, 11.376	0.629
Touchdowns per Game	-1.966	-19.785, 15.854	0.828
Interceptions per Game	-1.281	-40.357, 37.794	0.949
Under Conract = Yes	49.343	31.688, 66.997	**0.000**

A4 Logistic Regression Variance Inflation Factors (VIF)

Variable	VIF
Adjusted Passing Yards per Game	2.747579
TD per Game	2.054434
INT per Game	2.164123
Under Contract = Yes	1.155023

References

"Average Playing Career Length in the National Football League (in Years)." n.d. https://www.statista.com/statistics/240102/average-player-career-length-in-the-national-football-league/.

Fujita, Scott. 2022. "Football Positions: Ranking the Most Important Positions in Football [Full Details]."

Hermsmeyer, Josh. 2019. "The NFL Is Drafting Quarterbacks All Wrong." https://fivethirtyeight.com/features/the-nfl-is-drafting-quarterbacks-all-wrong/.

Kozlowski, Joe. 2020. "What Is an NFL Passer Rating and How Is It Calculated?" https://www.sportscasting.com/what-is-an-nfl-passer-rating-and-how-is-it-calculated/.

Mahoney, Joe. 2021. "Run, Quarterback, Run!" https://www.milehighreport.com/2021/9/1/22650269/dual-threat-quarterbacks-nfl-history.

Miller, Matt. 2013. "How Do Scouts Break down NFL Quarterback Prospects?" https://bleacherreport.com/articles/1632018-how-do-scouts-break-down-nfl-quarterback-prospects.

Mojica, Agustin. 2020. "How Long Is the Average NFL Career?" https://www.sportscasting.com/how-long-is-the-average-nfl-career/.

"NFL Positional Payrolls." n.d. https://www.spotrac.com/nfl/positional/.

Tabachnick, Barbara G., and Linda S. Fidell. 2010. *Using Multivariate Statistics, 7th Edition.* London, United Kingdom: Pearson.