ISyE 6416 – Basic Statistical Methods - Fall 2015
Bonus Project: "Big" Data Analytics
Proposal

Team Member Name: Benjamin Peters

Project Title: Clustering of NFL Quarterbacks Based on Performance Metrics Correlated with the Probability of Winning a Game

Proposal

Problem Statement:

The objectives of this project are threefold:

1) Determine the effect of quarterback performance on the probability of winning an NFL game.
2) Based on relevant performance metrics, cluster the players into groups. The groups should reflect tiers. For example, the best quarterbacks would be in tier 1 while the worst would be in the lower tiers.
3) Based on which tier a quarterback belongs to, the amount of money they should be paid can be assessed.

Background:

The premier professional American football league is the National Football League (NFL). Over the past decade, there appears to have been a large emphasis in the quarterback position. For example, NFL quarterbacks are often assigned a win-loss record similar to that of baseball pitchers or hockey goaltenders. The quarterback position is the only position in the NFL to be assigned a win-loss record. The emphasis on the quarterback position is understandable. Consider Figure 1:

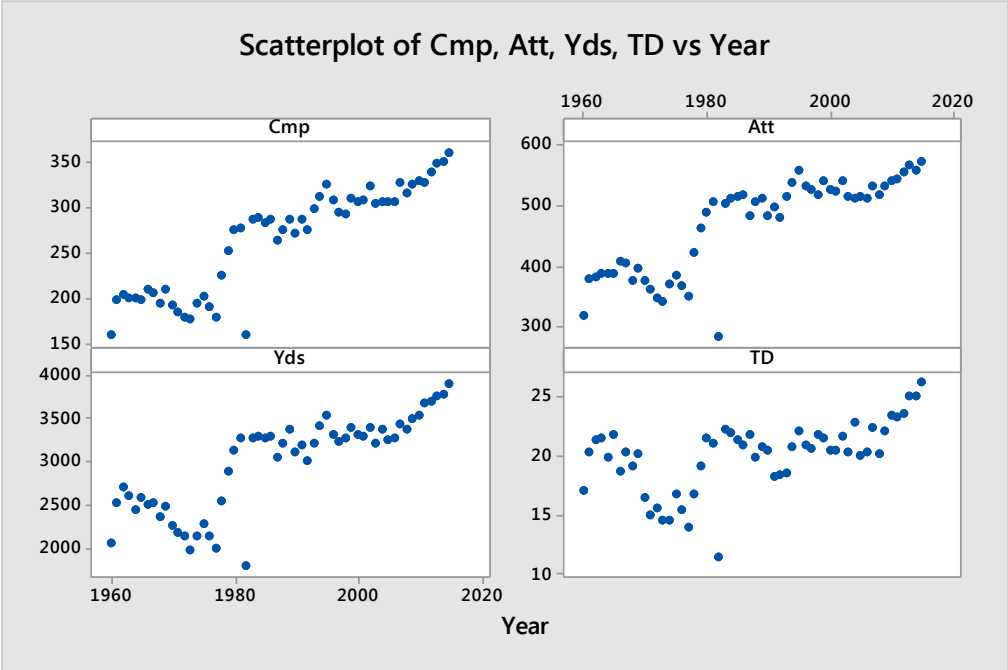

Figure 1: Team Yearly Average

Figure 1 shows that over the years there has been a general trend in the yearly passing attempts, completions, yards and touchdowns per team (It should be noted that the 1970s have a reputation for being a time period when the running backs were most prominent. As a result, teams passed much less in this decade than in any other decade since 1960).

With the increasing trend in passing, it makes sense that quarterback would be the highest paid position in the league. In the NFL, quarterbacks average over $3.8 million. With that much money being paid to quarterbacks, there is heavy responsibility on a team's quarterback to perform. The goal of this study is to find quarterback performance metrics (pass completion percentage, yards, touchdowns, interceptions, etc.) that are indicators of a team's probability of winning a game. A clustering method can be used to categorize the NFL's quarterbacks into groups based on the significant indicators. It is expected that players such as Tom Brady and Ben Rothlesberger would be clustered together as they are known to post favorable values for the mentioned metrics. The poorer quarterbacks would be clustered together as well. Using these clusters, predictions can be made on how much a quarterback can earn on his next contract. The results of this study could be used in negotiations between a player's agent and NFL teams to determine the value of the player. The next section discusses data collection and modeling framework.

Data Collection and Modeling Framework

The data for this project can be collected from Pro-Football-Reference.com. This website contains a plethora of data ranging from single game statistics, season statistics, career statistics, and game results. An NFL season consists of 256 games. Each team plays once a week for 17 weeks (with a week off). In order to model the probability of winning based on quarterback performance, quarterback performance is collected for each game of the season. The response is whether a team wins or loses coded as 0 and 1 respectively. Therefore, the regression problem can be modeled as a logistic regression where the win/loss result is the response variable and the performance metrics are the predictor variables. The exponential of the coefficients of the regression model indicate how the odds of winning change due to a unit change in the predictor.

In order to select the appropriate model, an L_1-norm penalty can be applied to the regression. This is called the LASSO penalty and drops the values of the insignificant predictors. In addition, the L_2-norm penalty can be applied. Elastic net is a tool which uses both penalties. While the LASSO penalty shrinks coefficients of insignificant predictors down to zero, the elastic net adds the ridge penalty (L_2-norm) in order to select amongst correlated predictors.

Expected Results

It may be difficult to find significant predictors amongst the basic statistics provided in a box score. For example, the raw numbers are not always indicative of the result of the game. For example, throwing for many yards may be considered favorable. However, there are games where a team may trail their opponent by a large margin. As a result, the team will get desperate and start throwing the ball more frequently to try to make up the difference in the time they have left. As a result, passing yards tend to inflate although the chance of winning may already be negligible. The final raw passing yards may show a quarterback throwing for a very high 500 yards, but the team lost by over 10 points. Therefore, some statistics may

need to be created to better represent the story of the game (It should be noted that the situation mentioned also leads to a higher chance at turnovers since the quarterback has to make riskier passes.) It is also expected that some players that may have mediocre records (.45-.55 career win percentage) may post performance metrics that are on the level of upper tier players. This is due to the many variables that go into a football game. If this case occurs, it is most likely the result of their team's weakness in another aspect of the game or that they are prone to making mistakes at inopportune times in the game (Turnovers late in close games).

Conclusion

This project can provide an objective view of the NFL's hierarchy of quarterbacks. A player that is looking to get a new contract could use the results of this study to determine how much he is worth. There is a dependency of these contracts on time that needs to be considered. Therefore, an additional part of this project can include modeling the contracts of NFL quarterbacks as an ARIMA model. The trend can be accounted for and be used to adjust the amount of money a player is expected to receive given which group of quarterbacks his stats are comparable to.