*Final Report*

***SMB-A Capstone Project Report***

***Predicting Margin of Victory in the NFL***

*by Dylan Wilkerson*

**Introduction**

Sports betting took sports fans by storm when the Supreme Court lifted the nation-wide ban on athletics-based wagering on May 14, 2018. Since that date, billions of dollars have been wagered on sporting events in markets that go from the NFL, to cricket, and even extending to Russian Ping Pong. While the field of sports betting is growing exponentially, there is also a segment of sports media growing congruently with it. All of this mirrored by the boom of big data and informed decision making is the perfect recipe for accurate predictive modeling precise line making for sportsbooks to offer their customers.

In this scenario, the NFL has decided to add another week to the NFL season. As a part of Fanduel’s analytics team, we are hoping to create accurate lines to maximize profits for week 18 of the NFL season. Previously, Fanduel has used a team rating system to rank teams and determine the margin of victory. The goal of this venture is to stray away from using an OLS model, where we solve for team rating by minimizing SSE, provided with the scores from the season’s previous games. Instead, we want to explore and implement a predictive modeling system where we can utilize season long statistics and use these to predict a margin of victory for our given matchups.

The goal of this project is to project and promote accurate lines, that provide two sides of a game that are appealing to our customers. In an attempt to maximize profit, achieving an even amount of money on both sides of a game is the goal that we are attempting to achieve. We want to discover which stats are the most important (independent variables) when predicting a margin of victory (dependent variables), And use these variables to formulate a model that will be used for predicting the margin of victory for a given matchup. These matchups were determined randomly in R, and consist of one AFC team playing one NFC team. Eight NFC teams will be the, “Home” team while 8 AFC teams will be the, “Home” team as well. These home teams will be randomly selected in R as well. Once these matchups were created, I used previous game data and a function in Excel to create a rating system that would be used as a reference to the output my predictive model would generate. The usage of this Excel feature is highlighted in the next section of the report.

**Data**

The primary dataset that I used for this project was from stathead.com. I specifically extracted game by game data that provided me with 417 observations, which contains every NFL game from this season recorded two times. Each observation contains a reference team and an opposing team. All of the statistics provided in each observation is recorded regarding the reference team. There is also an observation for the same game where the opposing team in the first observation is now the reference team, and vice versa. The stats in this second observation for the game represent the new reference team. For instance, in a game between KC and TB, the first observation would have KC as the team and TB as the opposing team. Points forced is 21 and points allowed is 14. In the second observation, TB is the team and KC is the opponent. Points forced is 14 and points allowed is 21.

The data contains many variables that range from time of possession to penalty yards. The appendix will feature a list of all of the variables and what they represent, however I will quickly highlight the ones that were specifically used in the analysis: PassAtt (Pass attempts), TotYds (Total yards), X3DConv (Third down conversions), X4DConv (Fourth down conversions), QBR (Quarterback rating), DPly (Defensive Plays), DY.P (Defensive yards per play), and PA (Points allowed). This data did not contain any missing values, which made data massaging and cleaning simple.

I also used various outside resources for descriptive pieces. I used the “teamcolors” library in R to extract the NFL logos they provided. I then used these logos for various graphs. I also used offensive and defensive stats over the course of the years to visually represent relationships that exist between a team’s skill and various stats provided by the aforementioned datasets. These datasets were accessed from profootballreference.com.

Excel provides a very useful function called, “Solver” that I utilized to create a team rankings system that is mentioned in Professor Draper’s Sports Analytics Class (BUSMGT 7334). The purpose of this team ratings system was to provide a reference to the values of margin of victory that I was going to be predicting. When given the scores of previous NFL games from the season, this method uses the actual margin of victory and predicted margin of victory to produce the error term. This error term is squared. Once all of the error terms are squared for all games during the season, they are summed. This results in the SSE. The goal of the solver function is to minimize the SSE by changing values of independent variables. The output provides with values that are then plugged in to this formula, which is provided from the book “*Mathletics: How Gamblers, Managers, and Sports Enthusiasts Use Mathematics in Baseball, Basketball, and Football”*:

***MOV= Visiting team rating – Home team rating + Home field advantage***

which is then used to predict the margin of victory for the next matchup, given the teams that will be competing. These ratings average out to zero, so the y intercept represents the home field advantage. When I completed this process, I noticed that home field advantage for the 2020 NFL season was very close to zero, only measuring up to approximately .02. For a normal season, this would be very unusual and be a cause for re-evaluation. However, when you consider that most stadiums were hosting fans or only hosting a limited capacity due to COVID-19, I decided to continue my evaluation without paying this anomaly too much mind. However, when calculating the margin of victory, I will use the offensive rating for the home team and the defensive rating for the home team. More on this in the predictive portion of the report.

**Analysis**

This project consists of multiple parts of analysis, that all comes together to provide insight on solving the issue at hand. The first part of this is the descriptive analysis. Visualizing the data helps us as data analysts identify any trends, and gather information to make optimal decisions. **Figure 1.1** details the grasp that each states have on the sports gambling scene. As you can see, Nevada and New Jersey boast the most robust handle on sports betting, with Pennsylvania following closely behind. The reason Nevada has a big handle on sports gambling is self-explanatory, but why does New Jersey have the biggest handle? This is because New Jersey hosts many mobile sportsbooks where you can bet form your phone, rather than being required to be in a brick and mortar casino. Atlantic City also being in NJ can be credited to the state’s handle on sports wagers. This shows us that the work line-makers do is important and can have serious ramifications if an offering is miscalculated or mispriced.

The subsequent 3 figures were created to help with the process of variable selection for the models I will be creating. I used Excel to create a baseline ratings system of each team, and using this rating for each team as a variable, I was able to regress select variables to team rating and see if something specifically stood out. Even though this rating is what I am trying to recreate with my analysis, the rating still provides a good general overview of a team’s performance. The insight gained by using these plots can help us further as we try to predict a game’s margin of victory using game stats.

As you can see in **Figure 1.2**, when an offensive team gives up penalty yards it does not seem that it effects a team’s rating or performance very much at all. The regression line boasts a slope of -.0009. This is not strong at all, and tells us that using offensive penalty yards against would not be a good variable to use when trying to predict margin of victory. **Figure 1.3 and Figure 1.4** tell us a different story. 1.3 shows us that the more turnovers a team allows, the lower their rating will be. While this seems like an obvious trend, it is important to note to what degree this trend takes place. With a slope of -1.167, this tells us that with every extra turnover a team commits on offense, you can expect their rating to drop by approximately 1.167 points. In **Figure 1.4** we take a look at a team’s rating and how it is affected by how many yards they allow per play on defense. This is another case that tells us there could be some meaningful context with regards to how yards allowed pertains to the margin of victory of a game and the performance of a certain team. This regression line has a slope of -7.719, telling us that when a team allows one more yard per play on average, we could see their rating drop by 7.719 points. Over the course of a season, seeing a one yard per play increase ends up being a lot of yards, and this could be why we see such a decrease in team rating as yards allowed per play increases. Once we visualize these trends and have an idea what variables are meaningful and which ones are not, we can move on to our model testing and selection process.