NFL Data Analytics For A New Era

Mitchell Collins, Gus Moir,
Jonah Muresan, Daniel Savage

California Polytechnic State University, San Luis Obispo

Introduction

With the increasing amount of data kept on American Football in the modern era, we decided to use the massive quantity of data available to test our predictive procedures to compare the outcome of any given game in our data versus the likelihood of that outcome. Not only is NFL data widely available in large quantities, but NFL analytics has an increasing demand for accuracy. While there are a fair amount of predictive models for wins by team already, there are few that look to assess their reliability of accuracy. We will also stem off of our predictive model to look at optimal fourth down strategy and how current coaching decisions should change in tendency.

Our goal for this analysis is to create a predictive model that will output a percentage for a team to win for any given game. We look to develop an accurate adaptive model that can be used to predict wins for future seasons to come. We will add this adaptation by using previous performance in games as the sole indicator for upcoming outcome. These variables and the regression model will then allow us to compare our predicted wins and win probability with the actual win percentile for the hypothetically forecasted year. This will, in turn, allow us to predict for future games one game ahead of time.

We will also take a look into the current trends for fourth down decisions per team. By using a model similar to what we used to asses win percentage, we can also assess optimal fourth down strategy based on situational factors for a team. This will account for the in game situation as well as the team's ability to execute that outcome.

This analysis will provide insight into using logistic models and their usefulness at comparing teams and their ability to play against each other. This is applicable for both a team's analytical side for their upcoming games as well as for better coaching decisions on fourth down.

Data Selection

We gathered our data from Advanced Football Analytics [1]. Knowing what sort of data we were looking for, this seemed like one of the best resources at the time. The site produced consistent CSV files on every play that occurred in the NFL for the seasons 2002 to 2012. Although we would have liked more data to create a larger sample, this format was rare with the specific variables and observations it contained, so we gathered the data and merged it into a CSV file containing over 450,000 plays of NFL data. The raw data file contained twelve original variables that outlined basic descriptive statistics about the game such as a unique game ID, which team has the ball, and current score for the game. One observation in the base data set represents one play or "down". The combination of these observations makes up the total of all plays for all teams for all games for the 2002 to 2012 seasons. The most interesting of these variables is a one sentence description of what happened in that play. This lists any important event on the play such as a touchdown, fumble, or penalty. Because this description has a standard template to describe the plays that is fairly constant throughout all seasons in our data we parsed the description to find indicator variables, cumulative statistics for each play, and in depth statistics per game for each team so we can evaluate their effectiveness at any point in the season.

Data Cleaning/Validation

Even though the data was quite flexible, it was by no means easy to handle. Throughout the course of our research we found more and more problems within our data. We ran into instances where plays were missing, plays were out of order, descriptions of what occurred in the play were incorrect and inconsistent, and scoring plays not being counted in the given score variable. To begin our work of the data, we created indicators to describe what happened in a play based off of the descriptions. This allowed us to easily see and analyze exactly what happened in a play. Therefore, if the description of a play was askew, then our data became incorrect because of the faulty reporting. The first step in the cleaning process was to find which observations contained these mistakes. These were found by searching for extreme situations within our data set and validating them to insure they weren't erroneously reported. One example of this was finding any play that resulted in a loss of 20 yards or more and cross referencing the given description of that play with what the other given variables for that observation reported. Once we identified as much

inaccurate data as possible, the cleaning began. To identify how a play should be correctly recorded, the original play must be looked up through a third party source [2] and then a change must occur within the program. These changes range from rewriting the entire description to changing the field position of a team and over 15 other variables in at least 300 observations. Then the validation code was run again to check if these extremities were properly taken care of.

Problem

Looking at the base data set, we can see it is easy to tell overarching trends for any given game. By observation alone you can see who is winning, who has the ball, and the time remaining in the game. The first idea we decided to address was to see how well a predictive model could use this raw data to predict win percentage. By cleaning the data and adding our own more descriptive and advanced statistics, we can use this data to interpret how well a team is doing and how well they will do in a future game. We then wanted to narrow our data down from every play to specific events to see if we could extrapolate based on a number of situational factors what the optimal decision for a fourth down situation is. These problems provide unique insight on a team's current and expected performance and suggests, based on team tendency and situation, what their optimal play call is for fourth down.

Analysis

To start our predictive model we created more advanced and in depth statistics that we thought would have a high correlation with whether the team won a given game. After parsing the description for each play and extracting the information, we created 191 other variables. These variables are either typical stats used to compare teams in the NFL like third down conversion percentage, or variables we conceived based on the necessity for a comparative statistic for that trait or tendency of a team.

Game Prediction Model

The next step was filtering these variables to figure out what would fit in our model to predict win probability. To do this we took all our variables and created a logistic regression, looking to maximize our R-squared value and minimize the MSE. [1] Our Logistic regression model is denoted as

$$\hat{p} = \frac{e^(b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p)}{1 + e^(b_0 + b_1X_1 + b_2X_2 + ... + b_pX_p)}$$

We ran a stepwise logistic regression with all of our 153 plausible variables testing for a significance at a level of 5 percent. We ended with a formula that had 9 significant variables, predicted the correct outcome of the game 64 percent of the time, and had an R-squared of 12.4 percent. With a more in depth look at our prediction, we encountered a false negative situation (predicting a loss when it was a win) 21 percent of the time and a false positive (predicting a win when it was a loss) 15 percent of the time.

When we predicted a game with a percent chance of winning between 40 and 60 percent, the percent chance of correctly determining the outcome was lower than when our model predicted an extreme chance of winning (80 percent or higher) or losing (20 percent or lower) *Figure 7*. This makes intuitive sense as when we predict the game to be close, the game will likely be close and could go either way.

Fourth Down Analysis

Once we knew that our predictive model was useful in helping us predict the event of a game, we turned to our next analysis goal in evaluating what should occur on fourth down based on optimal percentages. To begin this we created a scale for the three types of plays that could occur (a field goal is kicked, a punt occurs or a team goes for it on fourth down), and assessed the usefulness of predicting each one in the

[1] Mean Squared Error

current situation. For a punt, we determined the field position the punt would give the other team, for a field goal, we created a model that gave a percentage for how likely the kick would be good, and for a fourth down, the probability of converting on fourth down. We predicted punt distance using a basic linear regression, accounting for both yard line on the play and punter ability. We accounted for field goals and fourth down conversion using a logistic regression. By taking the probability of the offensive team scoring the next touchdown or field goal, the defensive team scoring the next touchdown or field goal, or neither team scoring before the end of play and multiplying those probabilities by their expected point value, we are able to create an expected point value for the next scoring play. We calculated the probabilities of each of the scoring possibilities using yard line, minutes left in half, and half (only looking at first and 10 plays because they are the only plays we need for the model). We then use this expected value to determine the probability of winning the game depending on the different outcomes. We used the following accepted formula from pro-football reference to calculate the probability in the game:

$$1 - \Phi(\frac{.5 - (OffScore - DefScore + (Line(\frac{TimeRmn}{60})) + ExpValue)}{13.45/\sqrt{(\frac{60}{TimeRmn})}})$$

- OffScore - Offensive score
- DefScore - Defensive score
- Line - Vegas line proxy created by percent chance of winning as determined by previous model
- ExpValue - Expected points scored on current drive
- TimeRmn - Time remaining on game clock

Our findings for this are that coaches in the NFL are generally too conservative on fourth down. For example, our model recommends that coaches go for it on fourth and 8 when the game is tied late in the fourth quarter and they are driving down the field. However, coaches tend to either punt and give up the ball or attempt the unfavorable field goal and fail. The mistake NFL coaches clearly make the most is not going for it. When punting or kicking a field goal is the recommended fourth down decision, coaches correctly make the decision around 80% of the time. However, when the recommended decision is to go for it, coaches only make the decision to go for it 20% of the time. [3] [4]

Closing Comments

Based on the best subsets method for picking the most significant variables for correctly predicting wins, we observed that the number of third down trips the opposing offense attempts, number of Redzone trips the offense gets, the number of touchdowns the team gets not on offense, the number of field goal attempts for the team, the number of Redzone trips the opposing team gets, the number of field goal attempts the opposing team gets, the number of sacks the opposing team gets, points allowed for the home team minus points allowed for the away team, total points scored by the home team minus total points scored by the away team were the most significant. It is important to note that while we had basic statistics such as average passing yards per game and average rushing yards per game, these more simplistic variables did not make the best subset which means that an increase in yards does not significantly translate into an increase in win percentage. As a result, the more advanced statistics that we created are more accurate indicators of wins. This is because these more sophisticated stats build off the more simplistic ones so they not only incorporate the associated increase from the basic variables but also benefit from factors such as the spot of the ball.

Generalization

For our analysis we created basic NFL statistics and from those, came up with either original or less popular statistics that we thought would have a significant effect on win percentage. Future studies could recreate

popular advanced NFL statistics such as QBR [2] or or DVOA [3] and then use them in a similar predictive model to see if these popular statistics are as significant at predicting wins as typically assumed. In addition off-field factors could be added to see if they affect a team's win percentage. We could look at elements such as attendance per home game for a team or revenue per season.

Based off of our optimal fourth down model future studies could look at specifically what variables make a team successful or unsuccessful on a fourth down situation. This could help teams know what traits to improve in order to improve in fourth down situations optimally. In addition we could look at the effects of if teams in the NFL were to use optimal probability to call plays on fourth down. For example, if teams did try to convert more on fourth down ,which is not typical by today's standards, would the league adjust to this change and thus the optimal strategy would then change again? Would teams be less caught off guard by going for it on fourth down so the chance of success or conversion is significantly decreased?

Conclusions

Looking at the predictive wins logistic model we can see which of our 153 plausible variables were the most significant predictors of a team's chance of winning any given game. Furthermore, looking at the odds ratio plot in *Figure 9* we can easily see, through the use of a Wald's adjusted confidence interval, whether those nine significant variables were positively or negatively associated with a team's percent chance of winning. Looking at the plot of predicted proportion of success vs. actual proportion of success given the data set we can see how close our model is at predicting a team to win or lose. In addition we can look at the predicted success rate vs. the proportion of correctly categorized games. This means that given the model predicts a team to win a game with an 80 percent chance of success, we can see that for this category of success we correctly predicted the result 82.2 percent of the time with our current model.

For the fourth down kick model, we received results that we may have expected. As has been found with other research into fourth down decision making, NFL coaches fail to make optimal fourth down decisions more than half the time. Our results showed that most coaches only made the correct call between 60 and 70 percent of the time. As seen in *Figure 3*, the coaches tend to make more mistakes when making decisions in between the 30 and 70 yard lines, as well as at the goal line. Additionally, they tended to make more mistakes when they have less yards to go.

During our studies we learned more about how analytics can affect the sports world and the importance of clean data to perform a well rounded study. Even though we may not all be sports analysts, we will surely carry the lessons we learned into our next ventures as aspiring statisticians.

[2] Quarterback Rating
[3] Defense-adjusted Value Over Average

Appendices

When using the predictive best subsets logistic regression to predict wins for a given team we ended up with nine significant factors. The follow list gives the full name of these variables.

1. Average amount of third down plays the opposing offense attempts
2. Average amount of times the opposing team is sacked
3. Average amount of touchdowns the team gets while not on offense
4. Average amount of Redzone trips the offense gets
5. Average amount of Redzone trips the opposing team gets
6. Average amount of field goal attempts for the team
7. Average amount of field goal attempts the opposing team gets
8. Points allowed for the home team minus points allowed for the away team
9. Total points scored by the home team minus total points scored by the away team

While creating our predictive model we needed to take the 16 games prior to the specified date or season. As a result we used a recursive macro to count and store these games so that we could extract the necessary data and statistics from the games. Below is a portion of code that shows the recursive part of our macro, which succeeds in merging the mean of the previous sixteen observations with the current observation.

Figure 1

```
"%if &rownum < 5582 %then %do;
        %let rownum = &rownum+1;
        %compile(&rownum);
    %end;
    %else %do;
        data sas.set1;
            set final;
        run;
        proc print data = final (obs = 500); run;
    %end;
%mend compile;"
    }%
}
```

Figure 1 immediately follows the macro's initial Proc SQL call, which specifies the monotonic function that allows us to identify and grab the mean of the previous sixteen weeks. We then combined this data with the actual result of the game in order to get a variable that checks to see if we predicted the game correctly. This macro call may seem unneccesary but it simplified the code from having many pointless retain statements.

Figure 2

```
data NFLdata2;
set NFLData1;
by  off gameid;
if first.off=1 then do;spot=1; curr=1;
    weeknum=1;lastdate=.;season=input(substr(gameid,1,4),8.);
    gamenum=1;curr=curr+1;week=1;spot=spot+1;
    date=mdy(substr(gameid,5,2),substr(gameid,7,2),substr(gameid,1,4));
end;
else do;
    if spot=17 then spot=1;
    lastdate=date;
    date = mdy(substr(gameid,5,2),substr(gameid,7,2),substr(gameid,1,4));
    if weeknum=>18 then weeknum=1;
    else do;
        if date>lastdate+7 then weeknum=weeknum+2;
        else weeknum=weeknum+1;
    end;
    if weeknum>=19 then weeknum=18;
        week=spot;
        gamenum=curr;
    curr=curr+1;
    spot=spot+1;
    if substr(gameid,5,2)='01' then season=input(substr(gameid,1,4),8.)-1;
    else season=input(substr(gameid,1,4),8.);
end;
    retain curr spot date weeknum;
    drop curr spot week date lastdate;
run;
```

Figure 2 counts the total number of games for a team, keeps track of unique game ID, and the week of the game while accounting for bye weeks and varying length weeks. It also creates SAS® dates to validate the variables for further use. This code was useful for identifying which games occurred in a specific season for a team. Since some games occur in January but are still considered part of the same season, we needed to evaluate which clusters of sixteen games are in the same season so we could then create cumulative statistics for that season.

Figure 3

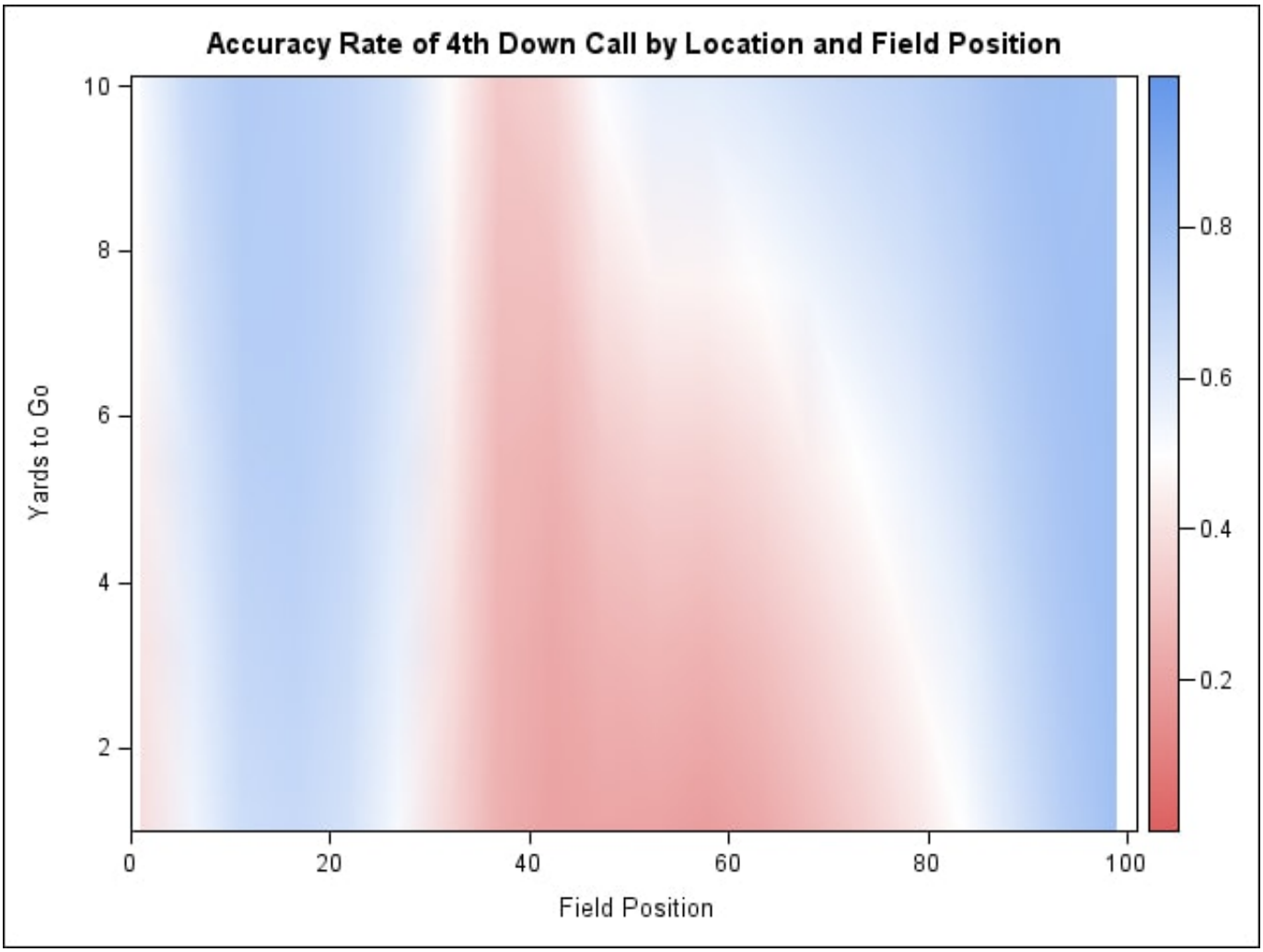

Figure 3 shows a heat map of the average coach in the NFL's ability to make the optimal call on fourth down based on expected value. From this we can see that tending towards the middle of the field more coaches tend to make calls that do not match up with optimal fourth down strategy. For example, with two yards to go and on the opponents forty yard line only about fifteen percent of coaches play according to the optimal strategy. Comparatively, when teams are facing longer distance yards to go on fourth down (eight to ten yards to go), they are more likely to make the more efficient play (as deemed by our model).

Figure 4

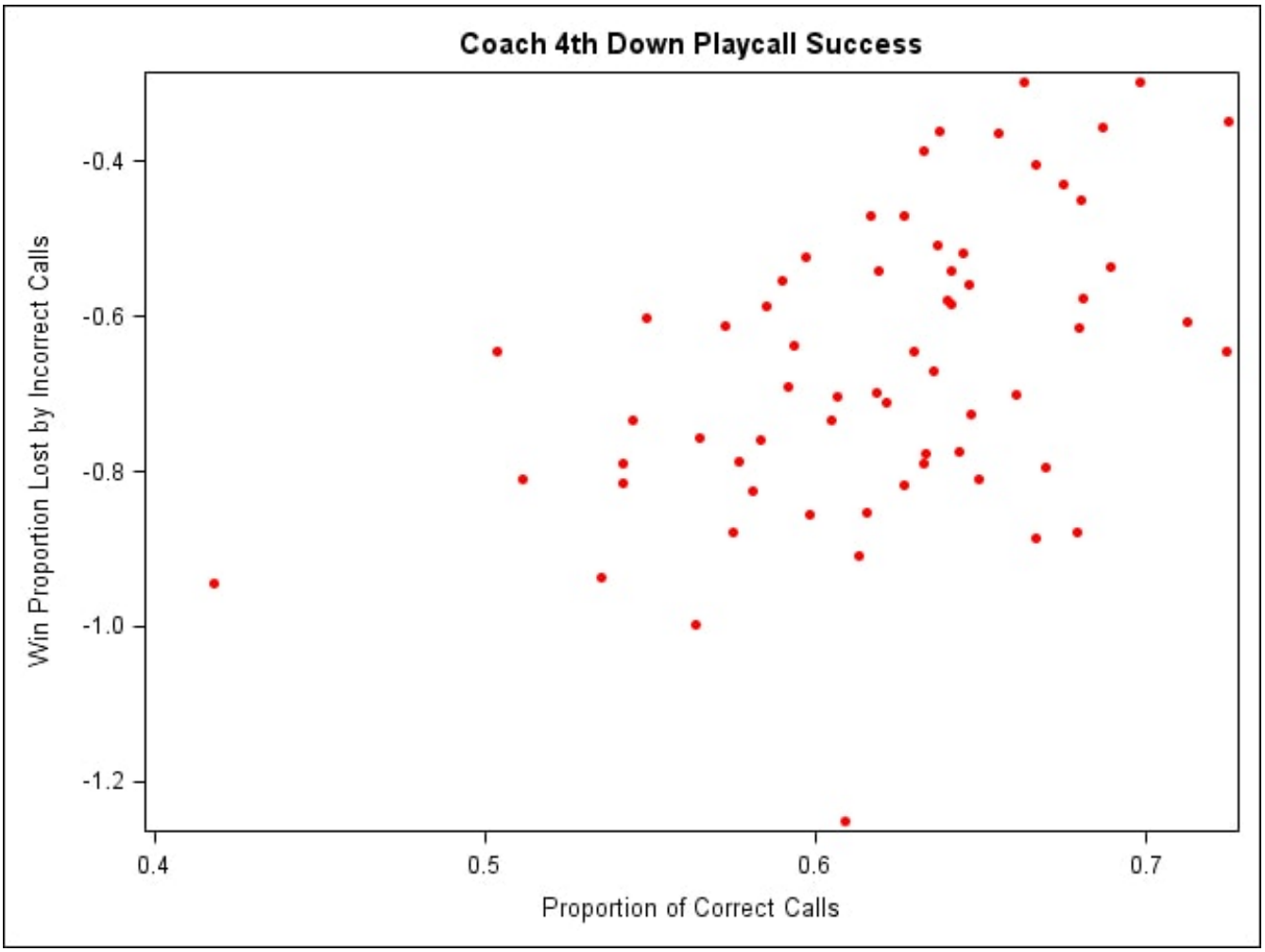

Figure 4 shows how the proportion of correct calls is correlated with win proportion lost over two seasons in the NFL (2011 & 2012 seasons). As seen, the higher correct proportion of calls is related to a lower win proportion loss. The teams closer to the bottom generally resulted in worse records, making poor decisions on fourth down can be seen to have affected this, proving that our optimal fourth down model helps teams to a winning record. Unfortunately, in the case of the dot at the bottom of the y-axis, the 2012 Oakland Raiders made incorrect calls on fourth down in very significant situations over the course of the season. Although they had a higher correct proportion than some other teams, this led them to just a four win season and management changes in the upcoming seasons.

Figure 5

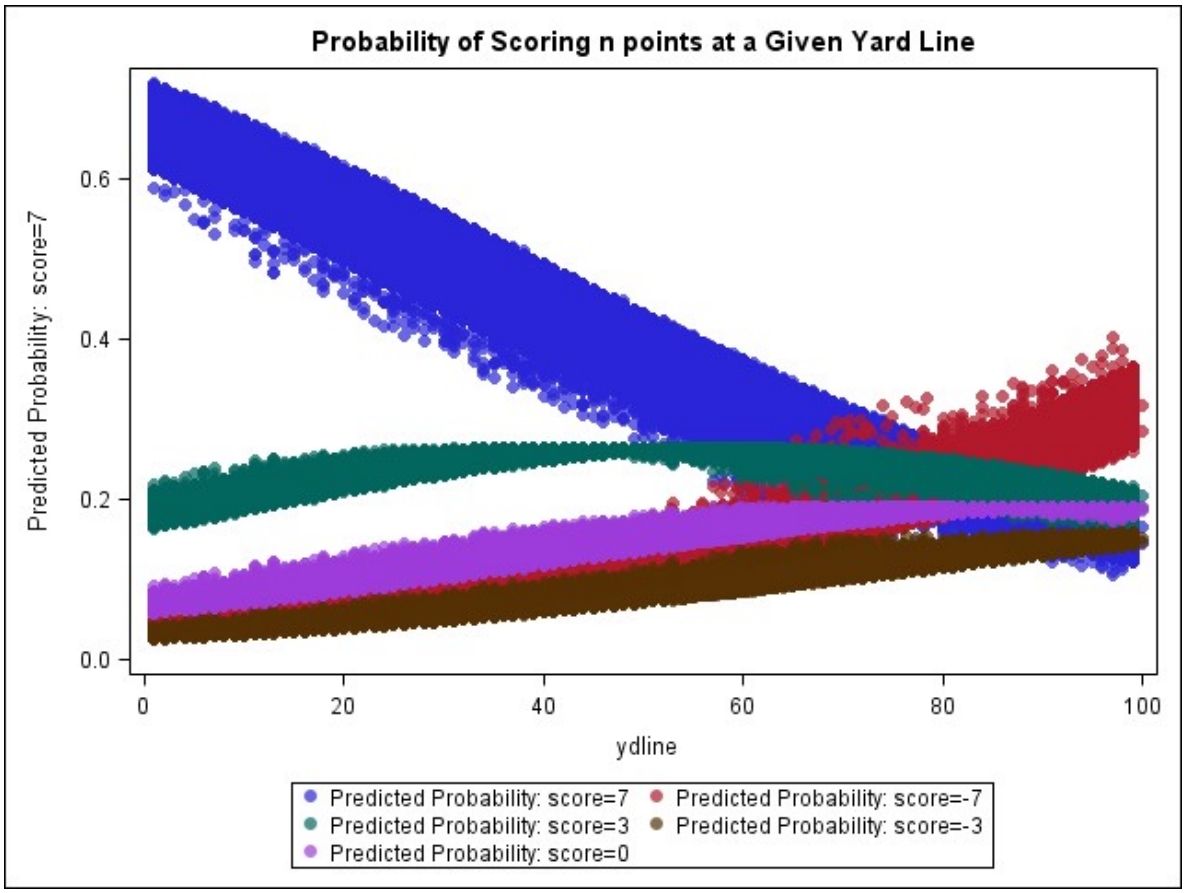

Figure 5 shows how the probability of scoring changes with yard line at a given first down. The number of points scored is represented by the different colors in the graph. It could be you scoring the touchdown or field goal, the opponent scoring a touchdown or field goal, or neither. The change in probability for scoring either a field goal or touchdown while possessing the ball increases as you approach the other teams end zone. This graph shows that when starting with the ball inside your own twenty yard line the opposing team is more likely to score a touchdown than the offensive team.

Figure 6

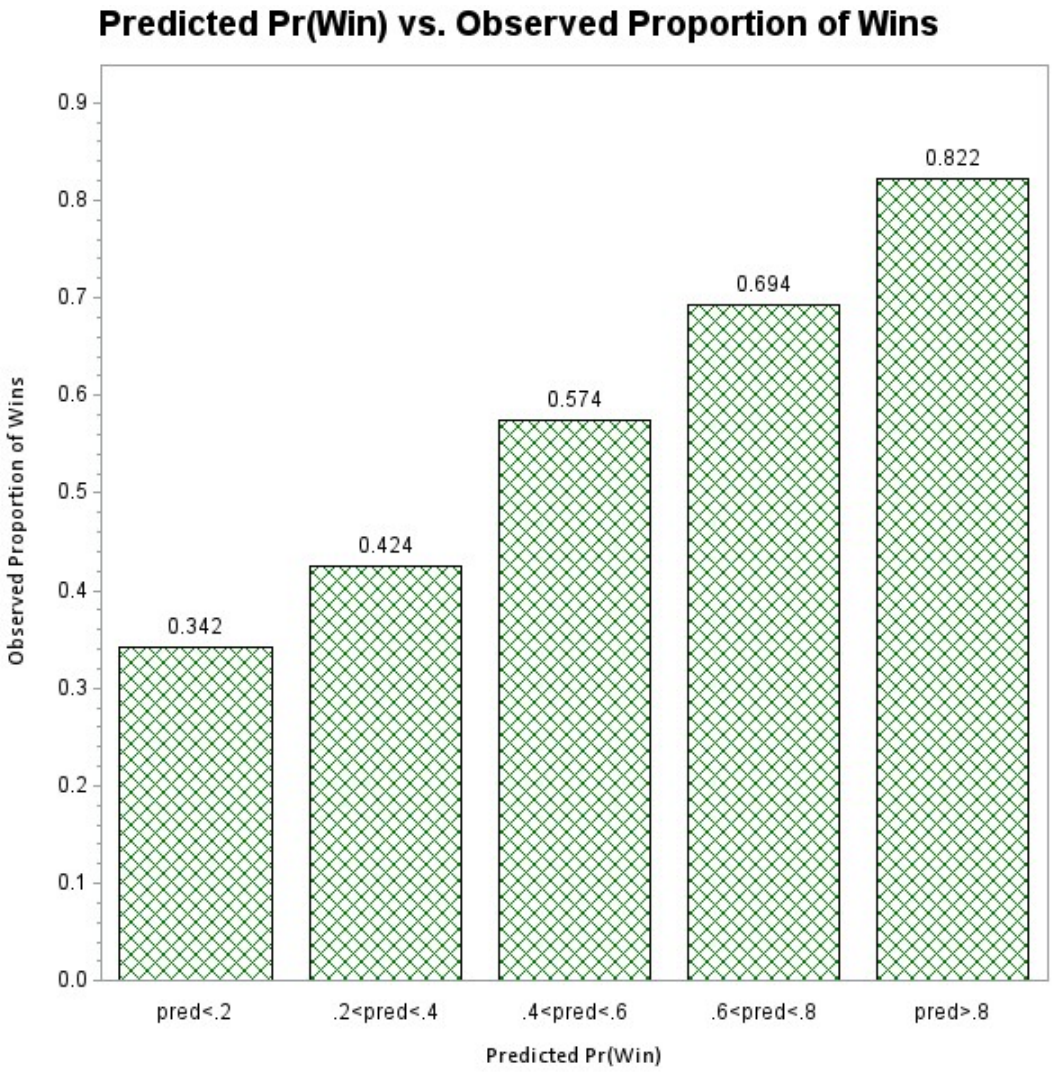

Figure 6 shows our predicted win probability in five equal bins and the actual observed proportion of wins for the home team. This shows an upward trend in actual win probability as our predicted model's probability increases. We can observe that for games that the home team was predicted to win between 60% to 80% of the time they had an actual average win probability of 69.4%.

Figure 7

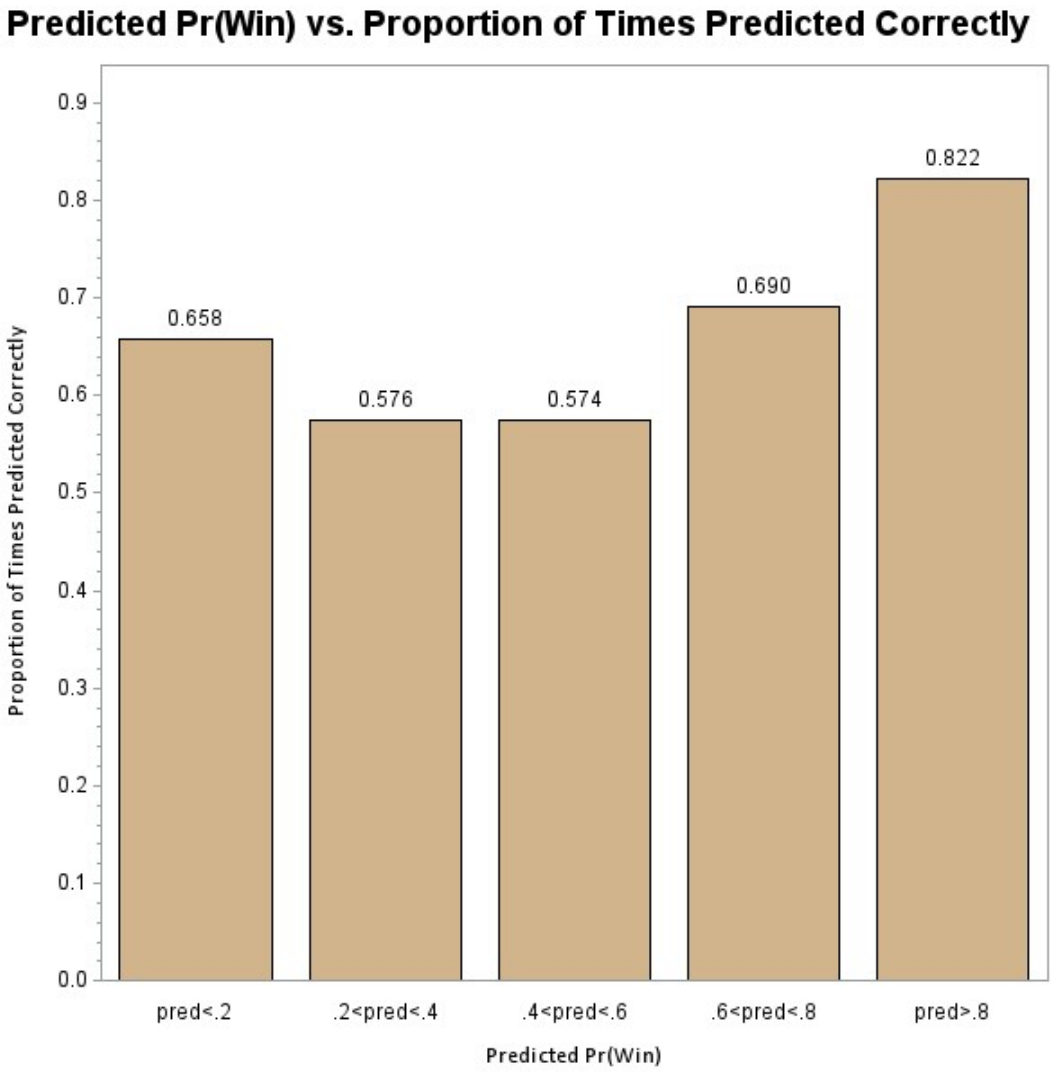

Figure 7 shows our predicted win probability and how many times we correctly predicted either a win or a loss for a game that falls within one of the five bins. This shows that our model does a better job of predicting games that it deems much more likely for one team to win than the other. As the teams predicted win probability gets closer to 50 percent the model has a lower chance of correctly predicting the result.

Figure 8

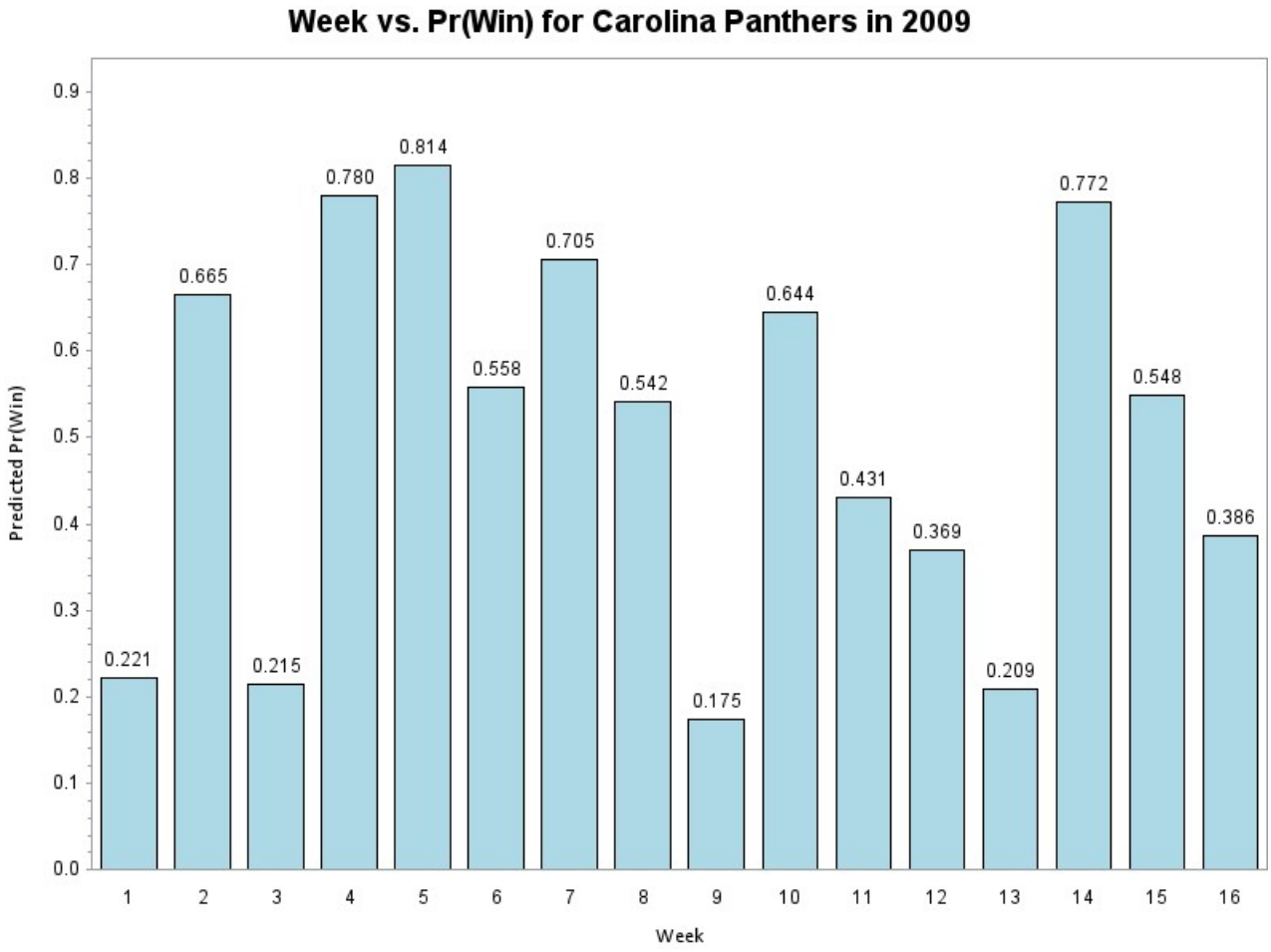

Figure 8 shows our predicted probabilities of The Carolina Panthers winning every game in their 2009 season. We see from the graph that we predicted they lose seven games. In actuality using [2] we know that they had eight losses. While this at first glance appears to be a fairly successful prediction of their season, our model predicted the panthers would lose games they actually won and predicted they would win games they actually lost. Our model predicted they would win or lose a given game this season correctly nine times making its accuracy for the season nine out of sixteen. This is why we must be careful predicting a whole season instead of a single game.

Figure 9

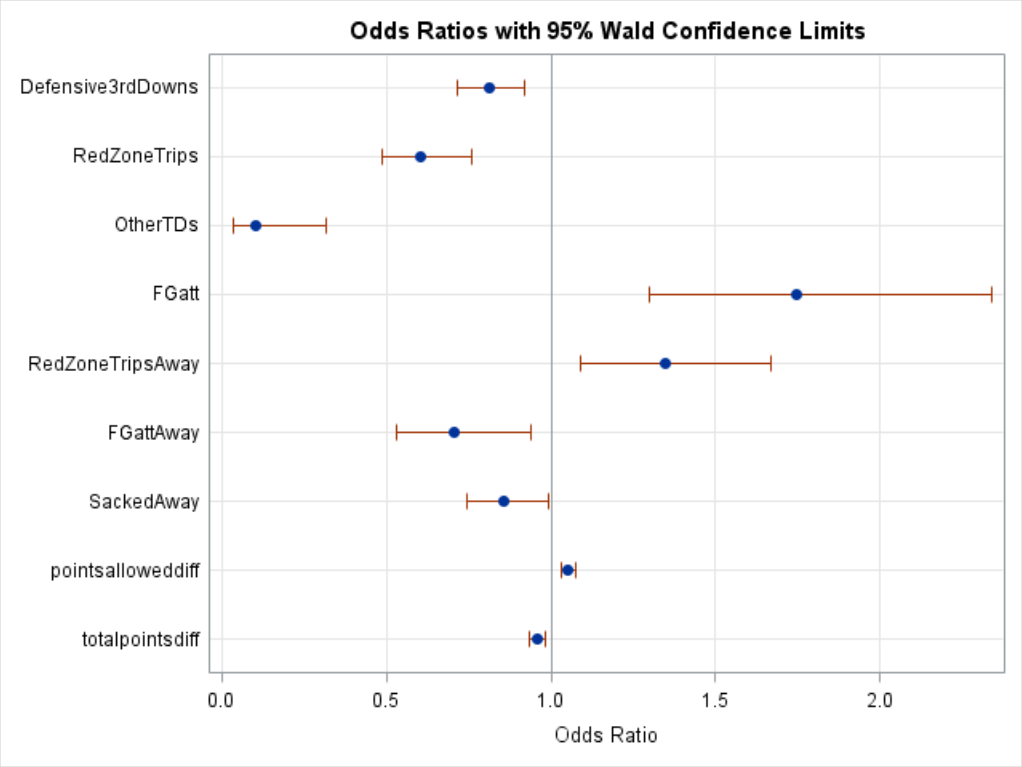

Figure 9 shows that since none of the 95 percent Wald confidence limits include zero, each of the given variables is a significant predictor of win, in the presence of the other variables. By viewing if the Odds Ratio is greater or less than one we can see if that variable has a positive or negative effect on a teams win probability.

References

[1] Brian Burke. Play-By-Play Data. http://archive.advancedfootballanalytics.com/2010/04/play-by-play-data.html, 2017. Online.

[2] Mike Kania. Pro Football Reference. http://www.pro-football-reference.com/, 2017. Online.

[3] Nate Silver. FiveThirtyEight. https://fivethirtyeight.com/, 2017. Online.

[4] The New York Times. NYT 4th Down Bot. http://nyt4thdownbot.com/, 2017. Online.