Predicting NFL Outcomes with Logistic Regression

In this project we are working with data found at fivethirtyeight.com. We chose a data set that involved NFL game outcomes. After finding this data we wanted to ensure that the records were accurate so we individually checked a few of the scores from the data set with the actual score from that NFL game. We found that the data set was indeed accurate so we proceeded to begin our research. Our goal for this project was to use our NFL data to predict what team would win and the score of the home team.

The reason we chose to explore this data set is because to many people, coming up with an idea about what team will win a football game is by and large an intuitive process. By conducting this research, we had the hope that we may be able to predict who will win a football game from a mathematical standpoint. To some people, having this as a tool would be extremely valuable since many people participate in fantasy football leagues, betting, etc.

From our source, we actually were given two csv files. Surveyors from fivethirtyeight collected this data. The first file consisted of NFL records from 1920-2021 with 30 variables in total. The second file consisted of NFL records for the current season (2020-2021) with a total of 30 variables. For our overall data we have 16,810 observations. For our research we limited ourselves to 9 independent variables all consisting of pre-game statistic that involve an elo rating of either a team or quarterback. An elo rating is a score/value that represents the relative skill of the quarterback or team. These variables consisted of;

Elo1\_pre: Home team elo rating before the game

Elo2\_pre: Away team elo rating before the game

QBelo1\_pre: Home teams Quarterback elo rating adjusted before the game

QBelo2\_pre: Away teams Quarterback elo rating adjusted before the game

Qb1\_value\_pre: Home, starting quarterbacks unadjusted elo rating before the game

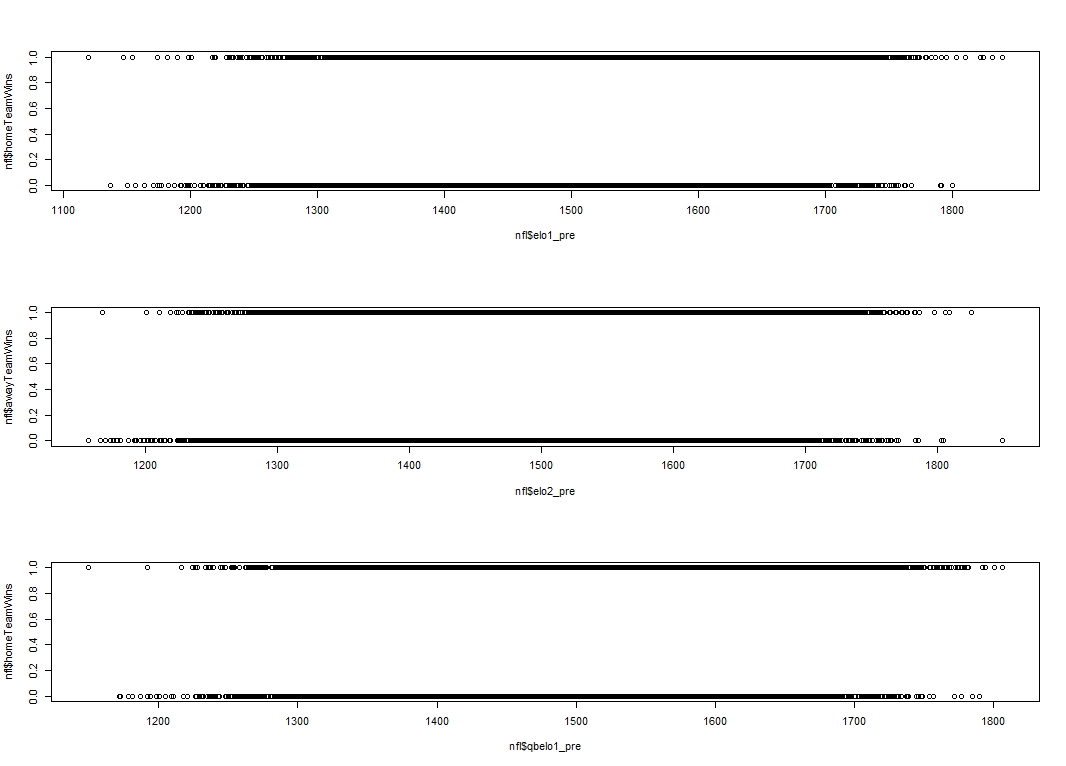
Qb2\_value\_pre: Away, starting quarterbacks unadjusted elo rating before the game

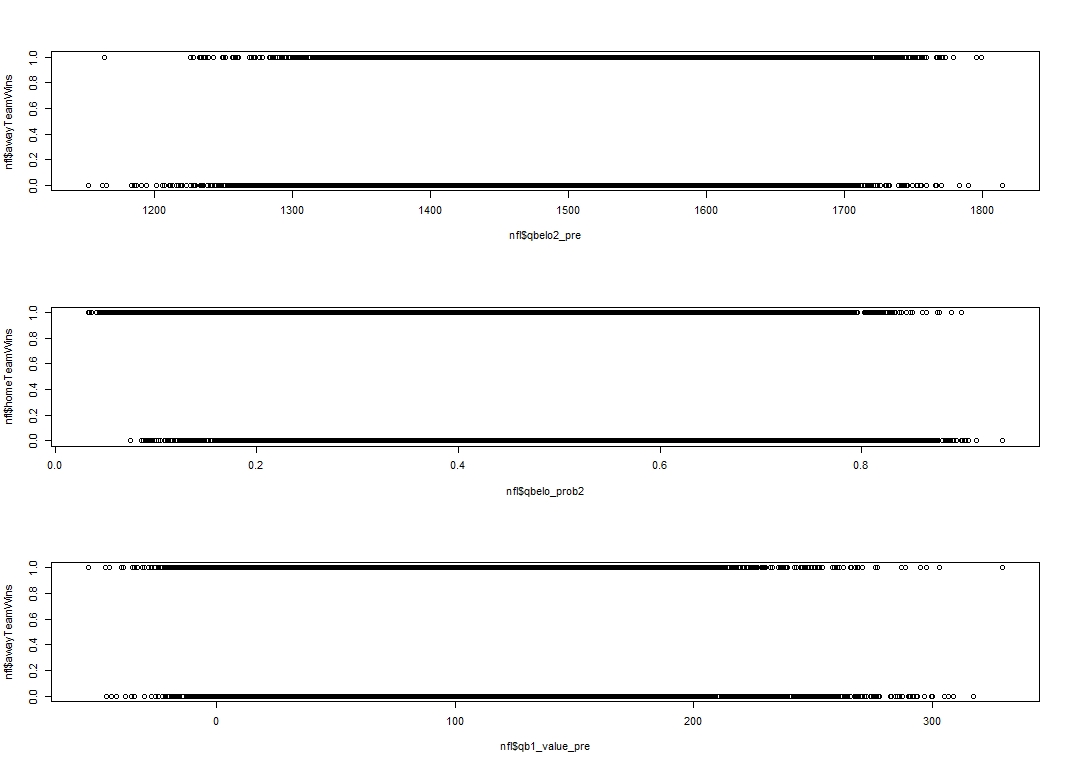
Qbelo-prob2: Away teams’ probability of winning according to QB adjust elo rating

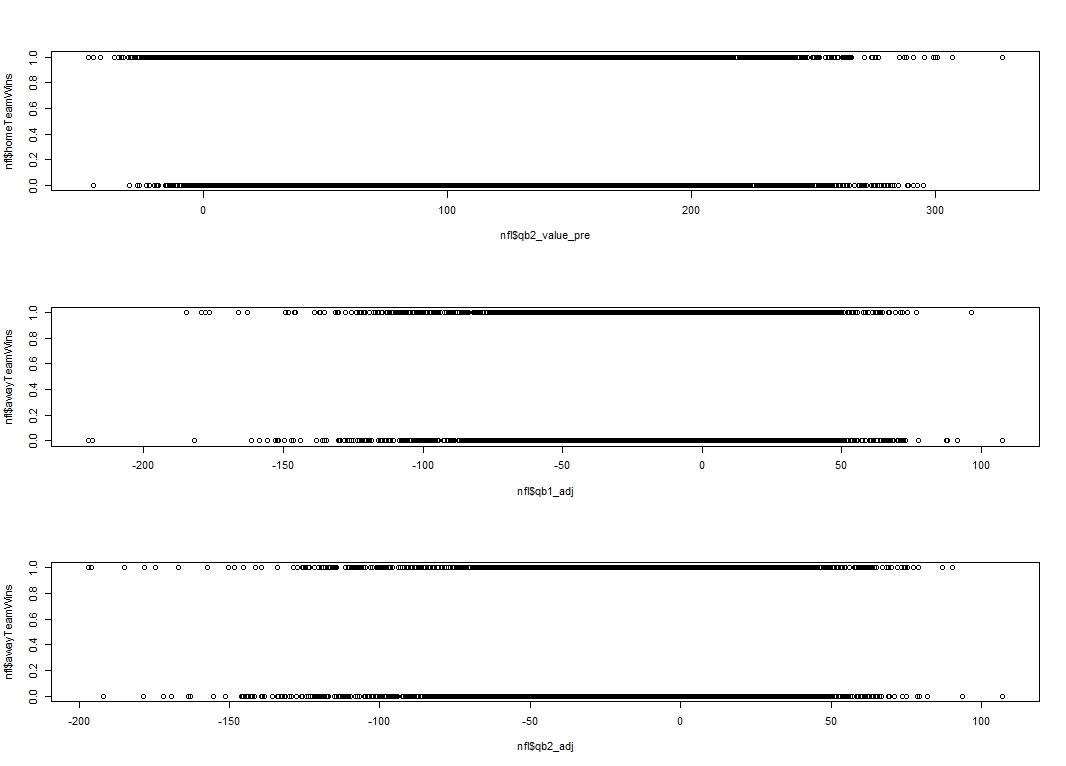
Qb1\_adj: Home team quarterbacks adjusted elo rating before the game

Qb2\_adj: Away team quarterbacks adjusted elo rating before the game

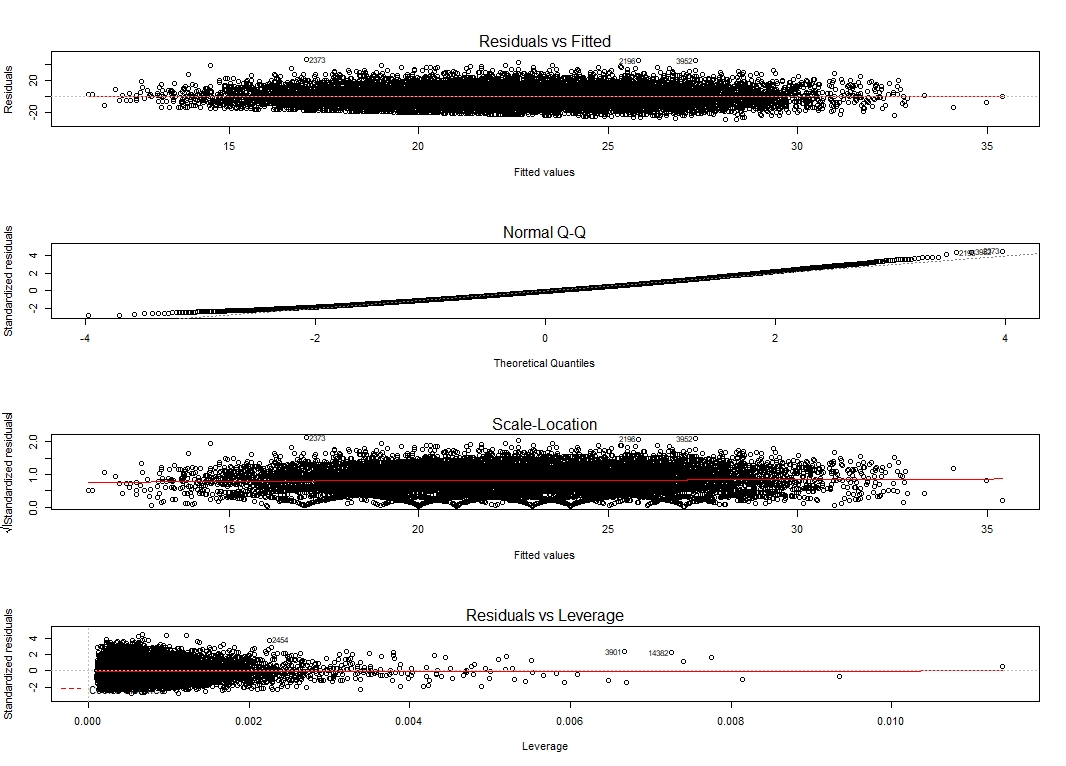
To begin our analysis, we divided our data into a training and a test set. The training set consisted of the years 1920-2019 which left us the data from the last season (2020-2021) for our test set. Our first step for our research was to do some exploratory data analysis to find out which predictors would be most useful for our model/s. To find our best predictors, we plotted each individual predictor variable against an indicator variable we created (wins). For each of these plots we found patterns that indicated that there was at least some relationship between the nine indicators listed above and wins. These plots have been included below:







After finding that these are our best predictor variables to use, we created a linear regression model and a logistic regression model. For our linear regression model, we wanted to predict the score of the home team. For this model, none of the variables used involved any post game statistics. Seeing as how we are trying to predict the score of the home team it would not be fair to use any post game stats. With that being said, all of the nine predictor variables did include an elo rating for either the team or the quarterback. After fitting our linear regression model, we assessed the assumptions that must be met for the linear model to work, and found that the residuals were normally distributed, the data was homoscedastic and not skewed. Furthermore, there didn’t appear to be any outliers or highly influential points present in the data. Our diagnostic plots for our linear model can be found below.



To test the accuracy of our linear model we looked at the RMSE (residual mean squared error). We got an RMSE value of about 10 which means that our linear model can predict the score of the home team within about 10 points on average. When we looked at the summary of our linear model, we found that Qbelo1\_pre was moderately significant with a p-value of .00379, and both Qb1\_value\_pre and Qb2\_value\_pre were highly significant with p-values of less than 2e-16 and equal to 6.81e-06 respectively.

Using the same predictor variables that we found previously, we created a logistic regression model. For our logistic regression model, we wanted to predict whether or not the home team would win. For this model, none of the variables used involved any post game statistics. Again, all of the nine predictor variables did include an elo rating for either the team or the quarterback. After fitting our logistic regression model, we calculated the sensitivity of our model to be 80.5% and the specificity of our model to be 57.5%. This means that when predicting wins for the home team our model was correct about 80.5% of the time. However, when predicting if the home team lost, the model was correct about 57.5% of the time. Below is a table that represents the confusion matrix that we used to calculate sensitivity and specificity.

|  |  |  |
| --- | --- | --- |
| Predictions vs. Actual Outcomes | 0 | 1 |
| 0 | 77 | 26 |
| 1 | 57 | 107 |

By using variables that included elo values for either teams or quarterbacks, we found that we could predict both the score of the home team and the outcome of the game fairly well. Our Linear model predictions were on average within 10 points of the actual score of the home team for any given game. Our logistic model predictions were accurate roughly 80.5% of the time when predicting if the home team wins and 57.5% of the time when predicting if the home team lost. Looking at our first five observations that are pasted below, we can see that both these models’ accuracy holds to be true.

* Game 9/10/20: KC vs HOU

prediction: KC wins Score = 29 outcome: KC wins Score = 34

* Game 9/13/20: BUF vs NYJ

prediction: BUF wins Score = 26 outcome: BUF wins Score = 27

* Game 9/13/20: DET vs CHI

prediction: DET loses Score = 24 outcome: DET loses Score = 23

* Game 9/13/20: WSH vs PHI

prediction: WSH loses Score = 20 outcome: WSH wins Score = 27

* Game 9/13/20: MIN vs GB

prediction: MIN loses Score = 24 outcome: MIN loses Score = 34

Overall, we found that it was in fact possible to create models that can predict NFL game outcomes. To improve our model, it would be interesting to get data that involves influential players at key positions. What we mean by this is it could be beneficial to have data from rookie players that maybe have a larger impact on the game than one might think and our current data reflects.