PREDICTING THE OVER/UNDER FOR NFL GAMES

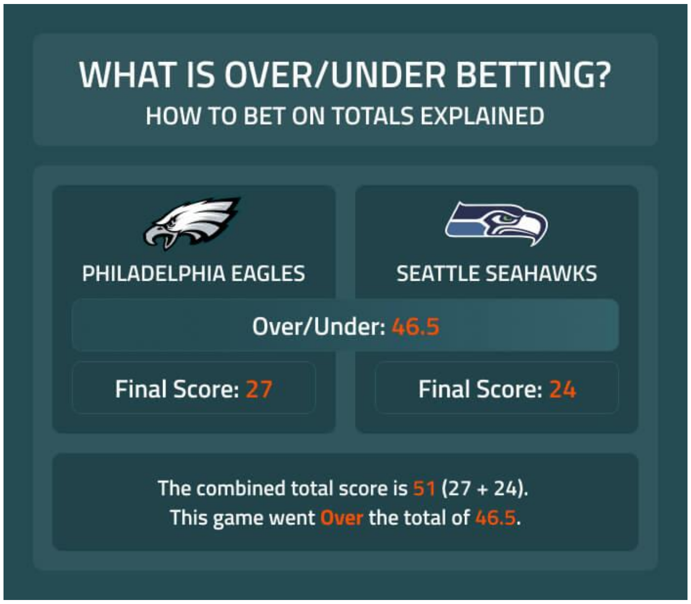
Mark Fier

STAT 692 5/2/22

Introduction

It’s a beautiful Fall Sunday and you’re in a casino with your loved ones doing what every true American does, betting on NFL games. You’re not sure what or how much to bet because you’re not exactly a gambler but you’re trying to spice things up in your life. You decide to bet the over on the Cowboys vs. Texans game which is set at 60.5 because you think to yourself “I know what I’m doing right?”. You look at the final score and the total is 45. Now you’re crying in the lobby of a casino because you realized you know nothing about sports betting, and you could’ve flushed $100 down the toilet and the outcome would’ve been the same. Wouldn’t it be great if there was a tool you could’ve used to predict the over/under of an NFL game?

Hi, Billy Mays here…. just kidding. Today we’re going to discuss how we can apply statistics to sports just like every other professional team or free-lance data scientist is trying to do but with a more degenerate twist. We are going to use statistics to predict the over/under on NFL games. We will specifically be looking at the over/under with regards to the total points scored between two teams in a single game. The diagram below is a great visual explaining the concept of the over/under bet.



WHAT IS OVER/UNDER BETTING?

HOW TO BET ON TOTALS EXPLAINED

The combined total score is 51 (27 + 24).
This game went Over the total of 46.5.

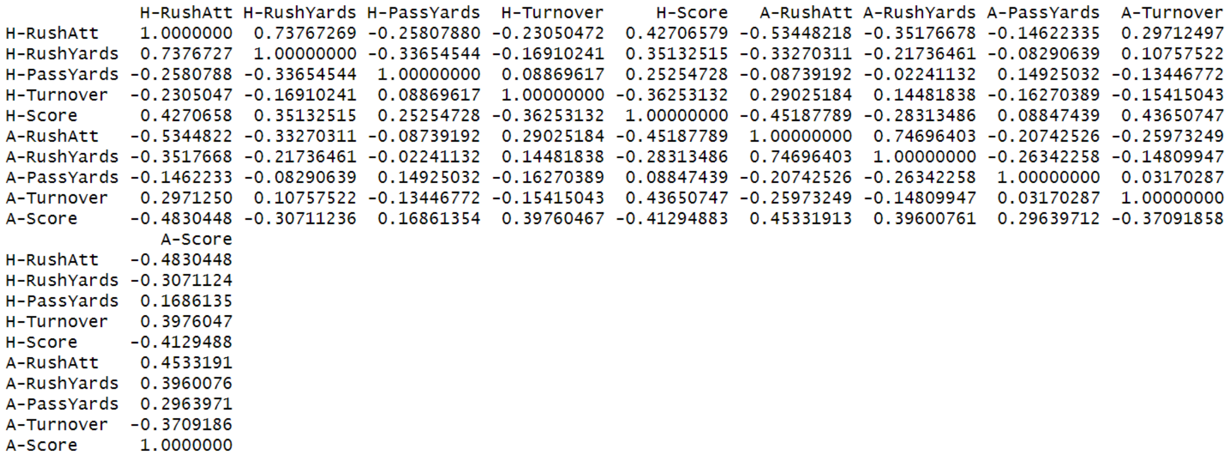
This study will be very beneficial to people who are involved in sports betting whether it’s on a regular basis or just occasionally. The models we will use in the study will hopefully become a tool used by people who have little sports knowledge and need help with their decisions as well as people with a lot of sports knowledge, but their intuition is failing them. Helping people become more profitable through statistics will be the key to making sure the House doesn’t win this time.

Pre-Processing/Analysis

We will use NFL game stats from 2010-2019 to fit a model that can predict the over/under. Our training data will be composed of games from the 2010-2018 season and our test set will be games from the 2019 season. We excluded games that were considered a “push” or games where the total score is equal to what the over/under was set at. The training data ended up having 2271 observations and the test data had 255 observations. The analysis and modeling were done in RStudio using a variety of packages such as dplyr, tidyverse, glmnet, caret etc.

The first part of the analysis included grouping the training data into games that hit the over and games the hit the under. There were 1138 games that hit the under and 1133 games that hit the over, so this is a balanced data set. We looked at the correlation between the variables in the data set to see if there were any noticeable relationships.

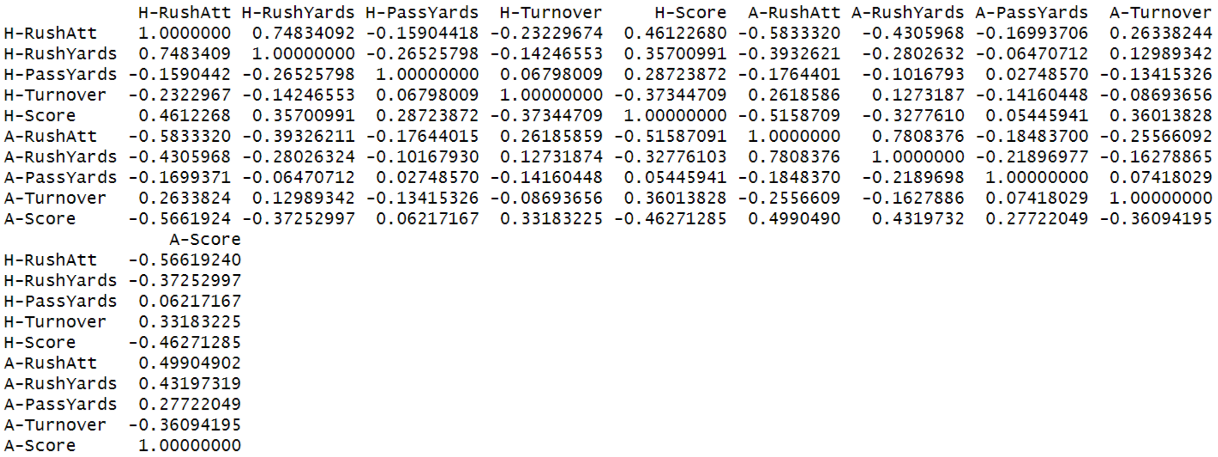
Fig. 1.1 Correlation Between Variables in Games That Hit the Over



```
             H-RushAtt H-RushYards H-PassYards  H-Turnover     H-Score   A-RushAtt A-RushYards A-PassYards  A-Turnover
H-RushAtt    1.0000000  0.73767269 -0.25807880 -0.23050472  0.42706579 -0.53448218 -0.35176678 -0.14622335  0.29712497
H-RushYards  0.7376727  1.00000000 -0.33654544 -0.16910241  0.35132515 -0.33270311 -0.21736461 -0.08290639  0.10757522
H-PassYards -0.2580788 -0.33654544  1.00000000  0.08869617  0.25254728 -0.08739192 -0.02241132  0.14925032 -0.13446772
H-Turnover  -0.2305047 -0.16910241  0.08869617  1.00000000 -0.36253132  0.29025184  0.14481838 -0.16270389 -0.15415043
H-Score      0.4270658  0.35132515  0.25254728 -0.36253132  1.00000000 -0.45187789 -0.28313486  0.08847439  0.43650747
A-RushAtt   -0.5344822 -0.33270311 -0.08739192  0.29025184 -0.45187789  1.00000000  0.74696403 -0.20742526 -0.25973249
A-RushYards -0.3517668 -0.21736461 -0.02241132  0.14481838 -0.28313486  0.74696403  1.00000000 -0.26342258 -0.14809947
A-PassYards -0.1462233 -0.08290639  0.14925032 -0.16270389  0.08847439 -0.20742526 -0.26342258  1.00000000  0.03170287
A-Turnover   0.2971250  0.10757522 -0.13446772 -0.15415043  0.43650747 -0.25973249 -0.14809947  0.03170287  1.00000000
A-Score     -0.4830448 -0.30711236  0.16861354  0.39760467 -0.41294883  0.45331913  0.39600761  0.29639712 -0.37091858
               A-Score
H-RushAtt   -0.4830448
H-RushYards -0.3071124
H-PassYards  0.1686135
H-Turnover   0.3976047
H-Score     -0.4129488
A-RushAtt    0.4533191
A-RushYards  0.3960076
A-PassYards  0.2963971
A-Turnover  -0.3709186
A-Score      1.0000000
```

There weren't any significant correlations other than rushing attempts and rushing yards for both the home and away team. The same can be said when looking at the correlation for the games that hit the under.

Fig. 1.2 Correlation Between Variables in Games That Hit the Under



```
             H-RushAtt H-RushYards H-PassYards  H-Turnover     H-Score  A-RushAtt A-RushYards A-PassYards  A-Turnover
H-RushAtt    1.0000000  0.74834092 -0.15904418 -0.23229674  0.46122680 -0.5833320  -0.4305968 -0.16993706  0.26338244
H-RushYards  0.7483409  1.00000000 -0.26525798 -0.14246553  0.35700991 -0.3932621  -0.2802632 -0.06470712  0.12989342
H-PassYards -0.1590442 -0.26525798  1.00000000  0.06798009  0.28723872 -0.1764401  -0.1016793  0.02748570 -0.13415326
H-Turnover  -0.2322967 -0.14246553  0.06798009  1.00000000 -0.37344709  0.2618586   0.1273187 -0.14160448 -0.08693656
H-Score      0.4612268  0.35700991  0.28723872 -0.37344709  1.00000000 -0.5158709  -0.3277610  0.05445941  0.36013828
A-RushAtt   -0.5833320 -0.39326211 -0.17644015  0.26185859 -0.51587091  1.0000000   0.7808376 -0.18483700 -0.25566092
A-RushYards -0.4305968 -0.28026324 -0.10167930  0.12731874 -0.32776103  0.7808376   1.0000000 -0.21896977 -0.16278865
A-PassYards -0.1699371 -0.06470712  0.02748570 -0.14160448  0.05445941 -0.1848370  -0.2189698  1.00000000  0.07418029
A-Turnover   0.2633824  0.12989342 -0.13415326 -0.08693656  0.36013828 -0.2556609  -0.1627886  0.07418029  1.00000000
A-Score     -0.5661924 -0.37252997  0.06217167  0.33183225 -0.46271285  0.4990490   0.4319732  0.27722049 -0.36094195
                A-Score
H-RushAtt   -0.56619240
H-RushYards -0.37252997
H-PassYards  0.06217167
H-Turnover   0.33183225
H-Score     -0.46271285
A-RushAtt    0.49904902
A-RushYards  0.43197319
A-PassYards  0.27722049
A-Turnover  -0.36094195
A-Score      1.00000000
```

After grouping the games by over and under, we grouped by home and away teams to see if there was a pattern of teams consistently being a part of games that hit the over or the under.

Fig. 1.3 Home Teams That Hit the Over

Saints	43
Patriots	43
Cowboys	40
Lions	40
Raiders	40

Fig. 1.4 Away Teams That Hit the Over

Eagles	43
Colts	42
Panthers	40
Packers	40
Buccaneers	40

Fig. 1.4 Home Teams That Hit the Under

Chiefs	47
Colts	45
Browns	42
Cardinals	39
Falcons	39

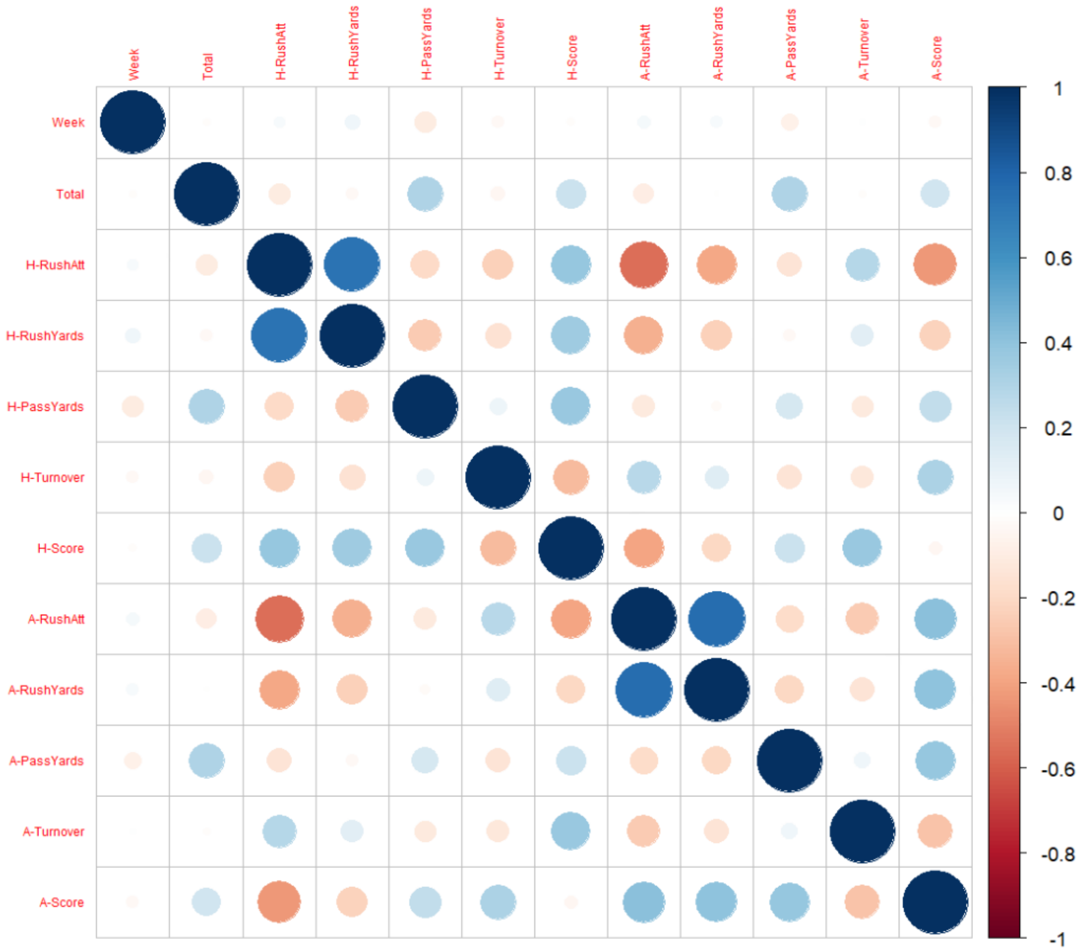
Fig. 1.5 Away Teams That Hit the Under

Steelers	43
Cowboys	42
Dolphins	42
Bengals	39
Vikings	38

The number next to the team's name represents the total number of games each team was a part of from 2010-2018 that hit the over or under. Although we see which home and away teams consistently hit the over or under, there doesn't seem to be a clear pattern of certain teams falling into multiple categories other than the Cowboys and Colts.

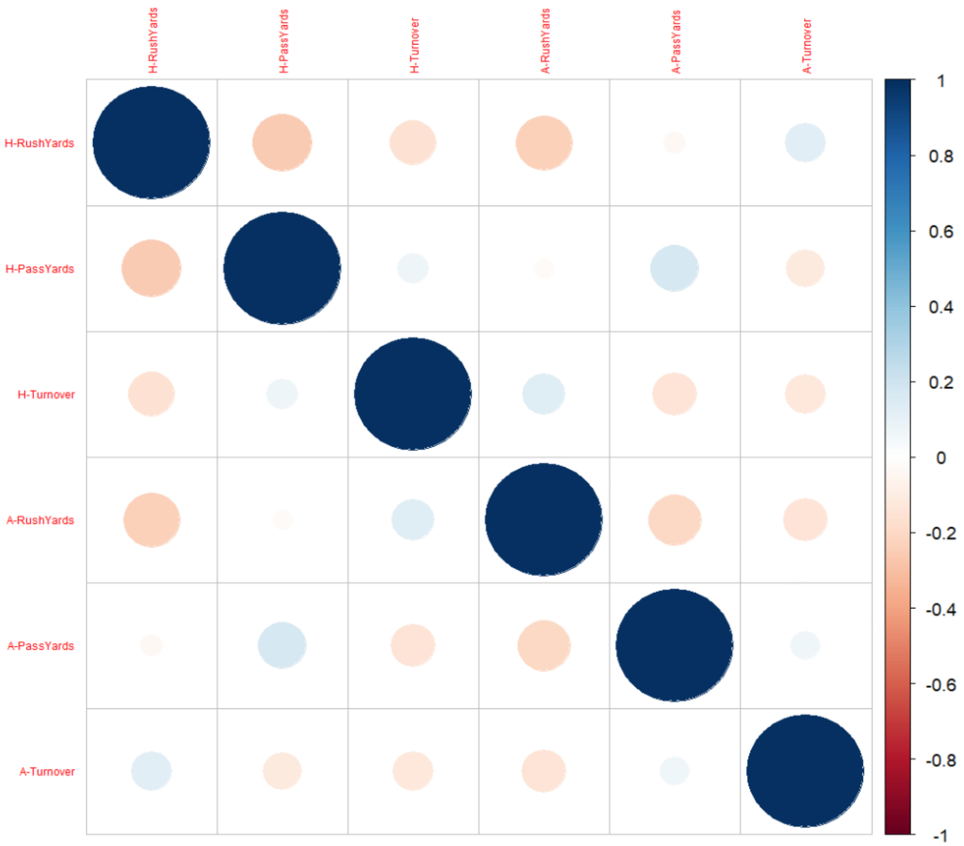
With regards to pre-processing, we removed variables such as team name, the week number of the game and the original total which the over/under was set at because we strictly wanted to use game stats in our model. When looking at the correlation between the predictors, we had to remove a couple features from the model such as rushing attempts and score because of multicollinearity concerns. We used this plot to determine which predictors should be removed based on correlation.

Fig. 1.6 Correlation Plot Between the Predictors



In the end, we ended with 6 total predictors: rushing yards, passing yards and turnovers for the home and away teams.

Fig. 1.7 Final Predictor Correlation Plot



Linear Regression Model

The purpose of fitting a linear regression model was to see if we could predict the number of points the home and away team would score. We could then add the predicted scores together to make our decision instead of fitting a model that strictly classifies the outcome of a game as over or under. Before fitting the model, we had to separate the home and away stats into their own training and test sets. We also looked for significant interactions between the three predictors used for the home and away models.

Fig 2.1 P-Values for Interaction Terms

- RushYards * PassYards (Away) = 0.332
- RushYards * PassYards (Home) = 0.387
- RushYards * Turnovers (Away) = 0.158
- RushYards * Turnovers (Home) = 0.825
- PassYards * Turnovers (Away) = 0.266
- PassYards * Turnovers (Home) = 0.213

There are no significant interactions between the predictors because all the p-values are greater than 0.05 so we will not consider them in our final model. After fitting the model, we see the performance is poor, specifically, when we look at the mean squared error and the r-squared value.

Fig 2.2 Linear Model Output

```
Home Model

  •  -0.787 + 0.0902(RushYards) + 0.0695(PassYards) – 2.282(Turnovers)
  •  R-squared = 0.45
  •  MSE = 52.003

Away Model

  •  -1.453 + 0.0915(RushYards) + 0.0639(PassYards) – 1.866(Turnovers)
  •  R-squared = 0.435
  •  MSE = 58.45
```

The first bullet point for each output includes the intercept and coefficients for the predictors in the model. The r-squared for both models are less than 0.5 meaning less than 50% of the variation can be explained by the predictors for both models. The MSE for both models shows that both of their predictions are off by approximately 7.2 points and 7.6 points meaning both models are off by a full touchdown on average when predicting the number of points scored by the home and away team.

When looking at the plot of the residuals, we see the variance is not constant. There is a decreasing pattern occurring in the middle of the plot suggesting that a linear model is not appropriate for the data. This could be because the offensive statistics we are using are discrete variables and we need to include some continuous stats that have a stronger relationship with points scored.

Fig 2.3

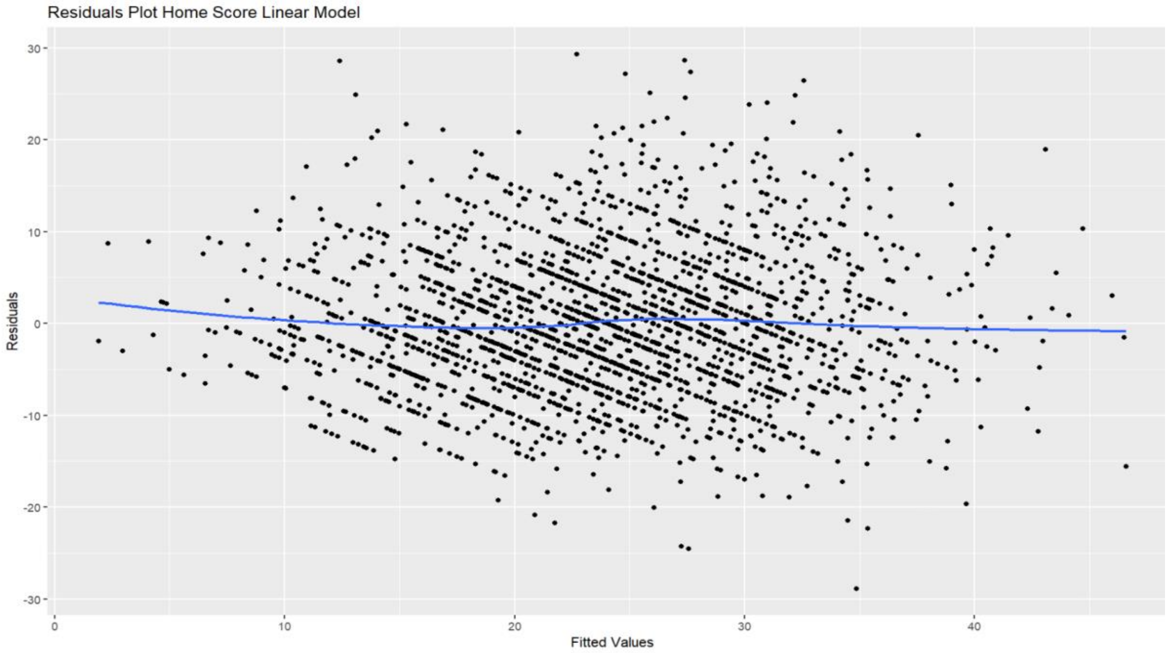
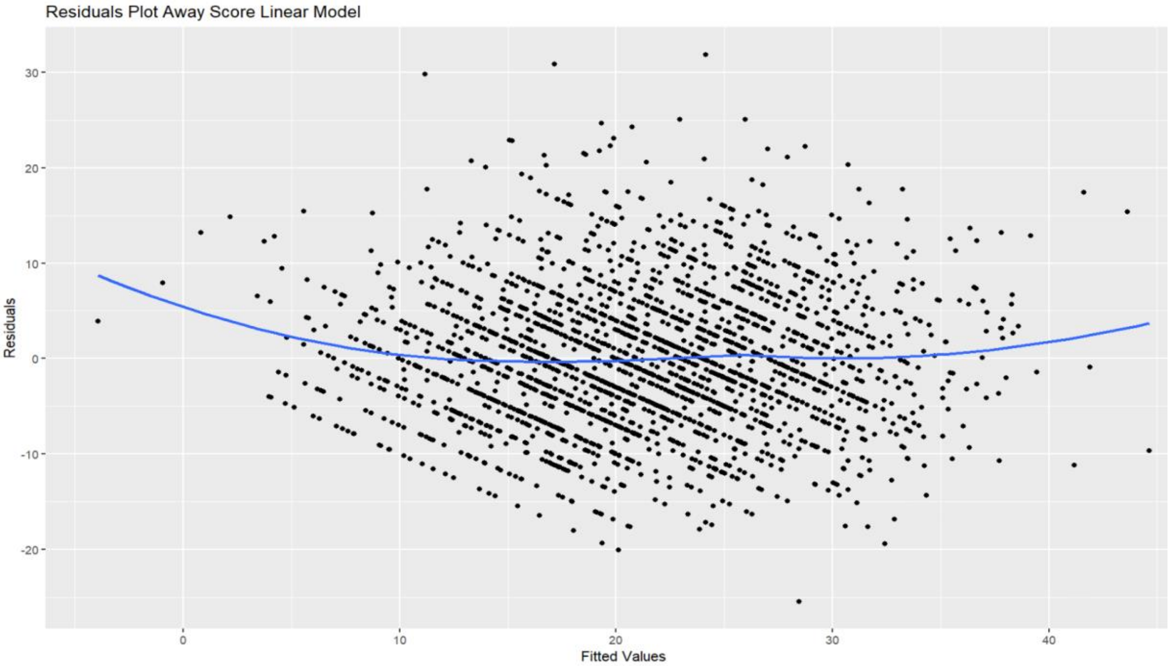


Fig 2.4



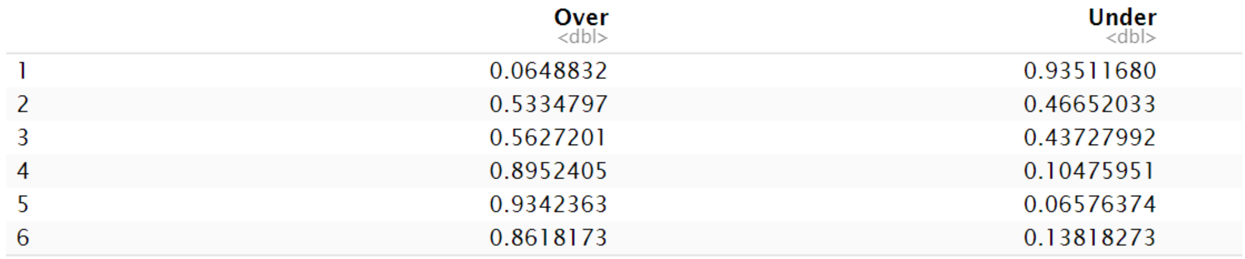
Going forward, we should investigate other offensive stats that have a stronger correlation with points scored if we want to answer our main question of interest using linear regression.

Logistic Regression Model

The purpose of fitting a logistic regression model is to see if we can accurately classify the outcome of a game as over or under. We specified ‘Over’ as our variable of interest for classification and applied 10-fold cross validation to prevent the model from overfitting.

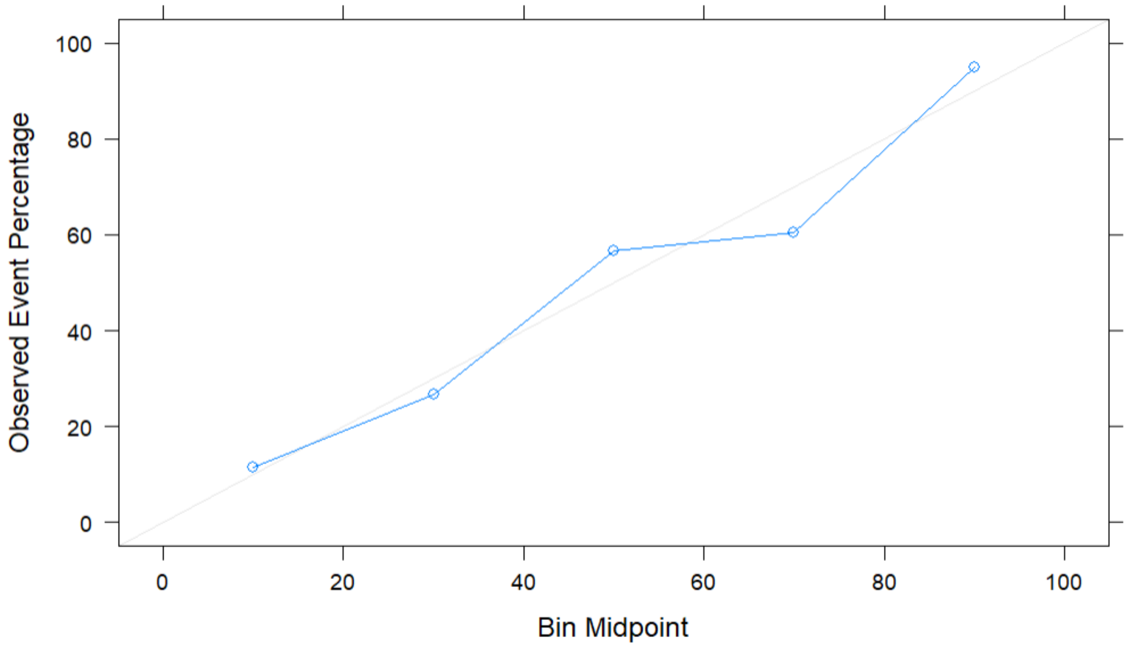
We can look at some of the predicted probabilities and the calibration plot to get a sense of how the model is performing. Looking at the first few predictions, it seems our model predicted more games to be over than under. The calibration plot supports this fact as it seems there is slight overestimation of our variable of interest.

Fig 3.1 Predicted Probabilities



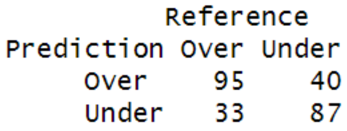
	Over	Under
	<dbl>	<dbl>
1	0.0648832	0.93511680
2	0.5334797	0.46652033
3	0.5627201	0.43727992
4	0.8952405	0.10475951
5	0.9342363	0.06576374
6	0.8618173	0.13818273

Fig 3.2 Calibration Plot



We used a few metrics to measure model performance, but our main metric of interest will be accuracy. This is because our original question is to see if we can accurately predict the over/under of an NFL game. According to our results, our model has an accuracy of 0.714, meaning our model can accurately classify a game as over or under 71.4% of the time. One issue with the model is the specificity, or true negative rate, is lower than we’d like it to be. This is a recurring theme of the model as we saw previously, the model tends to predict the over as the outcome for most games. The confusion matrix shows the predictions made after fitting the model on the test set. If we want to improve model accuracy, we might have to use the same approach we discussed for our linear regression model and find better offensive stats to use in the logistic regression model.

Fig 3.3 Confusion Matrix

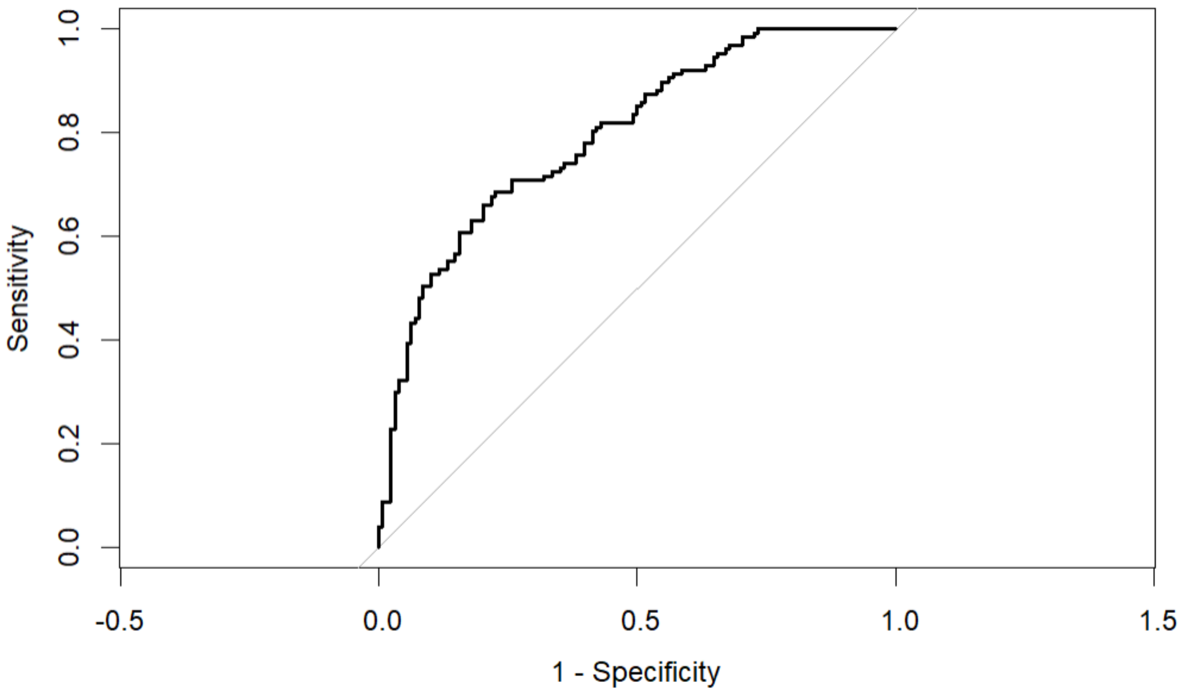


```
          Reference
Prediction Over Under
     Over    95    40
     Under   33    87
```

Fig 3.4 Logistic Model Output

- Model = 8.554 – 0.0132(H-RushYards) – 0.0112(H-PassYards) - 0.0754(H-Turnover) - 0.0123(A-RushYards) - 0.0102(A-PassYards) - 0.172(A-Turnover)
- Accuracy = 0.714
- Sensitivity = 0.742
- Specificity = 0.685
- AUC = 0.795

Fig 3.5 ROC Curve



The ROC curve is visualizing model performance at different thresholds of true negative rate (x-axis) and true positive rate (y-axis). We can use the area under the curve (AUC) as a metric for model performance when referring to the ROC curve. We see in figure 3.3 our model has an

AUC of 0.795 meaning our model is a fair classifier when distinguishing between over and under.

Conclusion

When comparing the 2 models, it’s obvious that logistic regression was superior. However, it would be interesting to dive further into predicting the number of points a team will score as an alternative to classifying the over/under. This could lead to a better model depending on the predictors that are used. The logistic regression model may not be as accurate as we’d like, but this first iteration is an exciting step forward. If we interpret the results practically, a model that can accurately predict over/under bets 71.4% of the time is still very profitable. This can be a reliable tool for someone who may not have great intuition when it comes to sports betting and needs help making decisions.

In sports gambling, there’s no such thing as a safe bet. Even when all the math checks out, the best player can get hurt on the first play and it changes the whole game. Although we’ll never make a perfect model, we have shown that logistic regression is a step in the right direction. Statistics will continue to bridge the gap between traditional sports intuition and scientific research when it comes to making decisions in sports gambling. Hopefully, we can continue to improve on this research and help others in their future sports betting endeavors.