Homework2 Logistic Regression Analysis on NFL Play Data

Xuanhui Liao 44592188, Yunchan Sun 47737525, Becca Wernick 47818736, Taowan Yang 44510821, Xiaowen Zhang 47801472

Nov. 18 2018

Project Overview

In this project, we performed logistic regression analysis on the NFL Play Data to predict whether the home team will win a game. Our dataset contains 102 variables over 362,447 plays over the course of 8 seasons.

We decided to aggregate our data to game level as we are aiming to make predictions for each entire game. As the raw data combines statistics for the Home and Away teams in the same column, one of the most important elements of our data preparation was to identify the pertinent statistics, and then split those columns into two variables, one for the Home team and one for the Away team. We also made new variables for each type of play, for Home and Away team, for example, home_Type_Pass and home_Type_Run, so that we can aggregate the number of each type of play for each team. Additionally, we created calculated variables, like the pass completion ratio, so that we could evaluate the effect of accuracy of passers on winning a game. For the purpose of modeling, we chose the statistics that we felt had an impact on the game performance. For example, the statistic, "ydsnet", which displays the net number of yards gained on a play, was split into "home_ydsnet" and "away_ydsnet" and then evaluated at the game level. Also, we excluded all probabilities since we did not want to base our prediction on others' prediction. After data cleaning, we were analyzing 2,048 games across 65 variables.

Exploration of the Raw Data

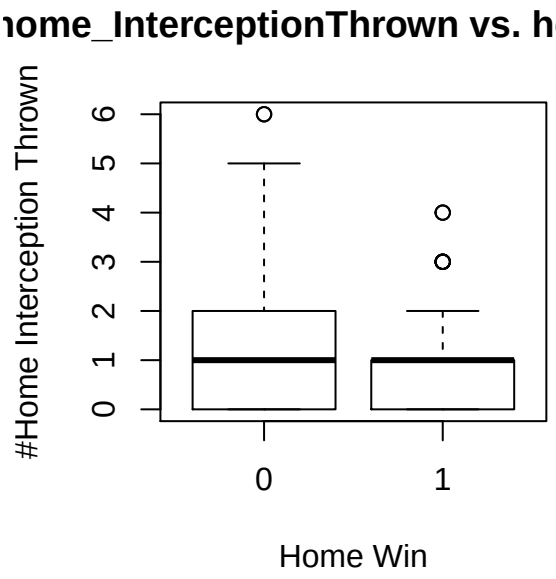
After exploring all variables, we discovered that the following four numerical variables in box plots have more distingiushed distributions between whether home team won or not.

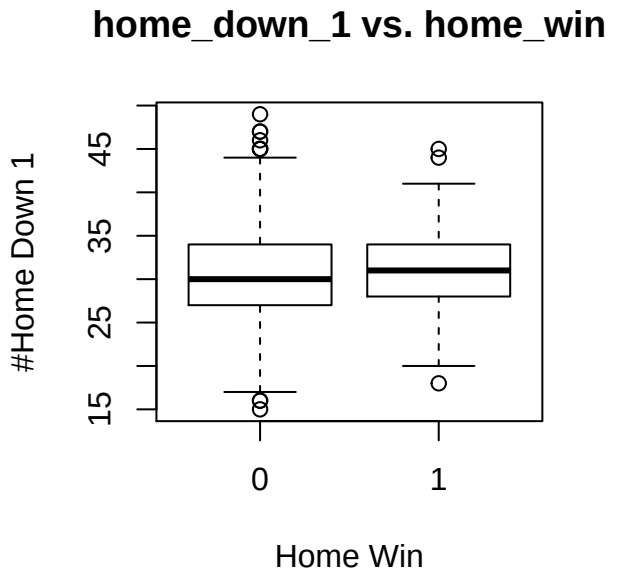
Regarding home_InterceptionThrown, when the throw of the home team is inctercepted less frequently, which means the home team is good at offense, it is more likely to win the game.

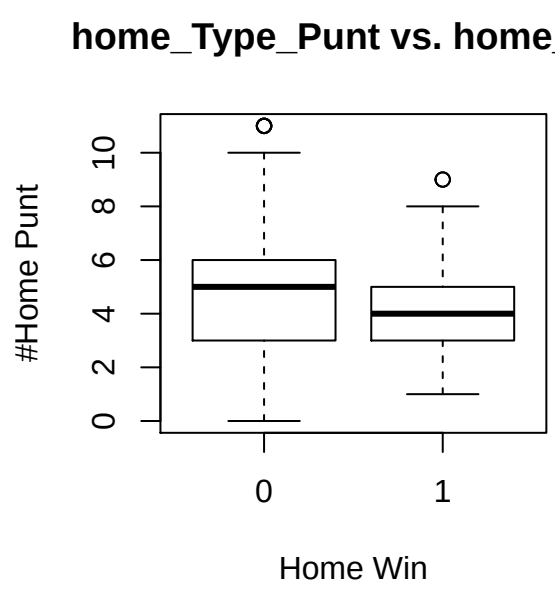
Regarding home_down_1, in the games home team won, the home team performed more stable and had more total number of first downs.

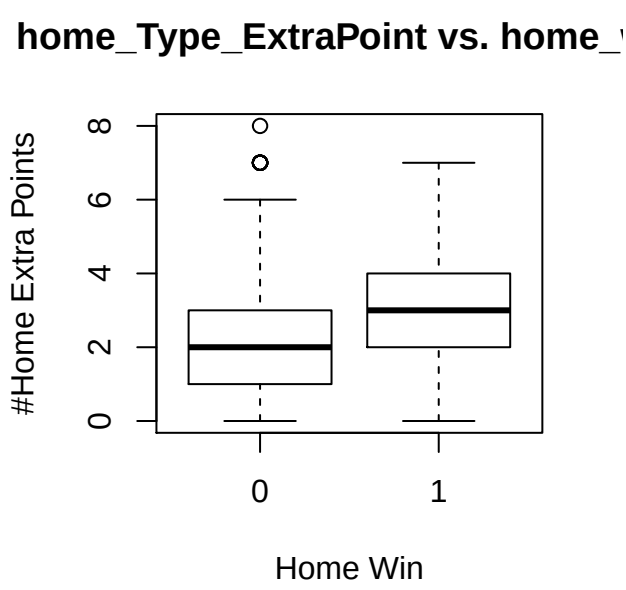
Regarding Home_type_punt, the distribution of home team win has lower average with less variability. In the games home team won, home team punted less on average and have more stable performance. Therefore, punting is an indicator of mediocre offensive performance and is an important indicator.

Regarding home_type_ExtraPoint, distribution of home team win has higher average and higher maximum. The more extra points home team got, the more likely it would win, which is not terribly surprising but still and important marker.

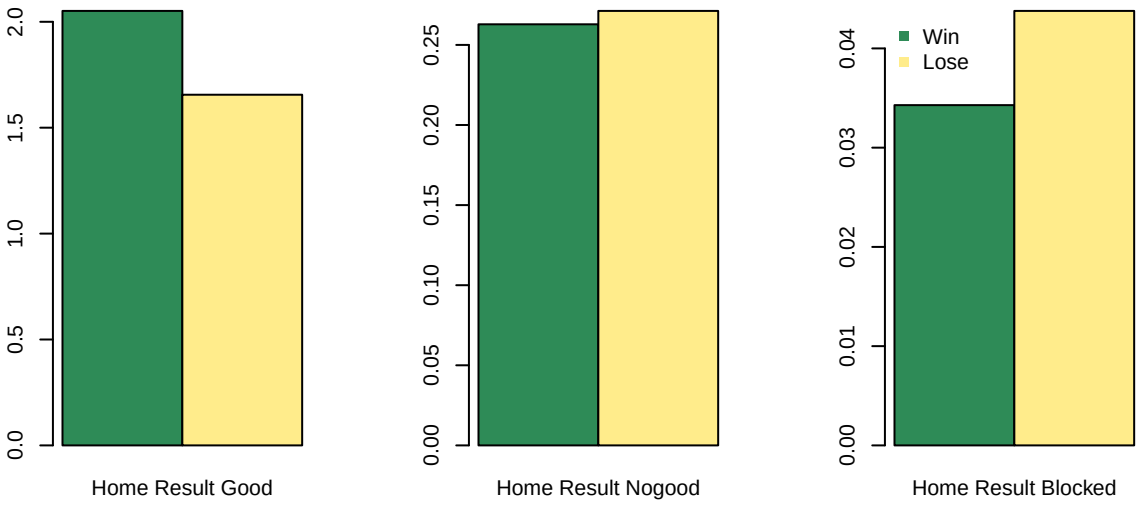








From the barcharts below, "Field Goal Result" shows a significant impact on the result of the game (if home team wins or not). If the home team has more "Good" than "Blocked" field goals, it is more likely to win the game; while "Nogood" outcome does not have a huge impact as missed field goals are a rarity in the NFL, especially at the time of our data.



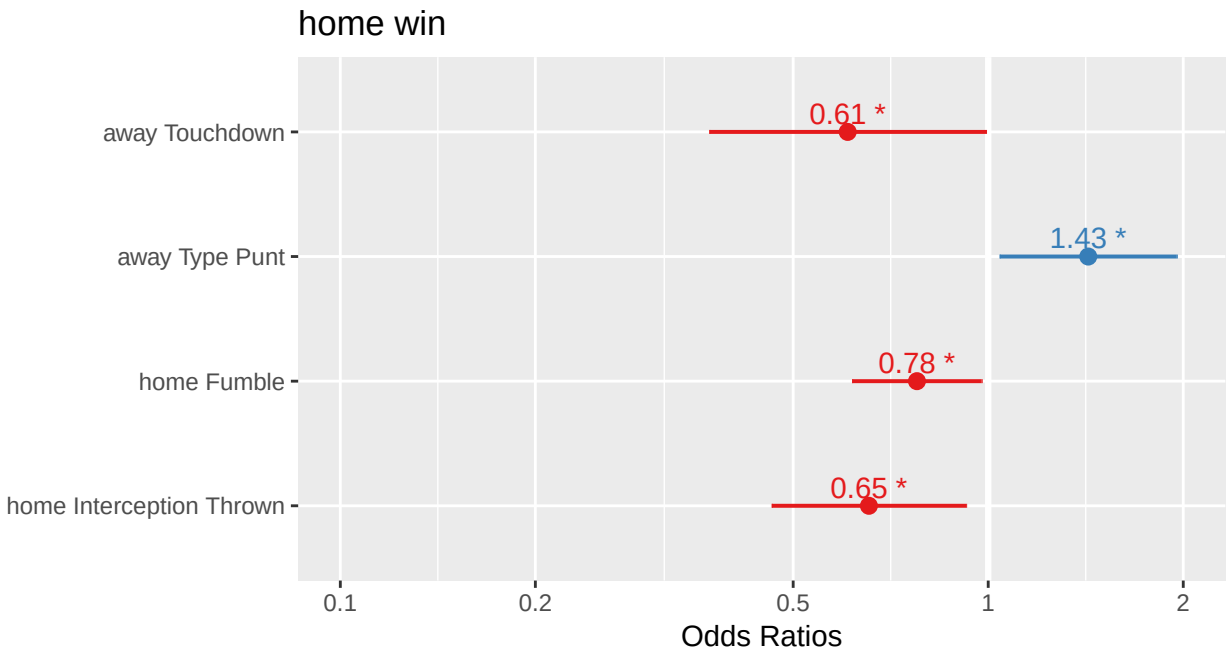
Model Building

We randomly split our data into two sets: 70% of our data is used to train the model and the remaining 30% for validation purposes. We applied the logistic regression model to predict whether the Home team will win. We then validated our model by performing Pseudo-R2, predicted-vs-actual, and ROC curve analysis as diagnostic tests. Our final model passed all the diagnostic tests and was able to predict with 90% accuracy whether or not the Home team will win a game. Our model's AUC (area under curve) score is 84.18%, which indicates that our model is performing relatively accurately.

We performed the AME (Average Marginal Effect) Analysis and found that:

- The chance of home team winning decreases by 2.6 percentage points with one unit increase of awaydown_1
- The chance of home team winning increases by 2.8 percentage points with one unit increase of away_PlayAttempted

We also generated Odds Ratios Plot for selected influencial variables(home_InterceptionThrown, away_Touchdown, home_Fumble, away_Type_Punt).



As shown above, home_PlayAttempted,home_InterceptionThrown,away_Touchdown and home_Fumble have an odd ratio below 1 which are 0.23, 0.65, 0.61, and 0.78 respectively, while away_Type_Punt, away_PlayAttempted are above 1. Therefore, we concluded that:

- The odds of homewin with a home_InterceptionThrown is 35% less than that without home_InterceptionThrown
- The odds of homewin with an away_Touchdown is 39% less than that without away_Touchdown
- The odds of homewin with a home_Fumble is 22% less than that without home_Fumble
- The odds of homewin with an away_Type_Punt is 43% more than that without away_Type_Punt

There are certain elements that our data does not take into account that could have an effect on the outcome of several variables. One of the most notable is weather. Fumbles, punts, dropped passes, and running speed during rushes and passes are just a few elements that could be affected by the weather. In a perfect world, we would have identical weather conditions across the board for all games. As this is an impossibility (even cold or rainy conditions vary by location and day), the weather will impact players' ability without directly impacting the score. Further, player injuries (and disciplinary actions) are not taken into account. A star player or multiple linemen being injured would likely have a great impact on the entire game, from plays called to yards gained to the final score. Again, if we were given our choice, all players would remain uninjured (and eligible for games). The referee crew could also have an impact on the outcome of individual variables as well as the game as a whole. Individual judgement calls are only, by rule, reviewed on certain plays. Personal judgement and interrelationships of the crew as a whole can have a negative impact on the game. In our perfect experiment, the referees would have machine-like precision and all calls would be reviewed in an objective manner. As it stands, even calls that are reviewed have an element of subjectivity.