Applied Predictive Analytics - NFL Home Game Outcome Analysis

Alex Grill - February 2023

Introduction

It is well understood that competition is intense and each win holds significant value in the National Football League (NFL). The objective of this study is to determine factors that have significant impact on a home team's win probability. Through analysis of game data from the past two decades and the use of logistic regression, our findings will provide an understanding of the key determinants in NFL home team success. We hope this will be of value to teams and fans alike.

NFL Data Background and Preparation

To carry out the objectives of the study, the first step was to gather a firm grasp of all available information. The data set, nflfastR, contains NFL play-by-play data back to 1999 with over a million recorded plays and 372 variables per play. These variables range from specific play measures to entire game attributes. After this, we put the data through a rigorous preparation process to ensure accuracy and consistency for modeling and analysis. Our analysis includes only regular season games from 2002 to 2022 in order to account for the most recent NFL expansion and coinciding division reorganization.

We then funneled the play-by-play data down to the game level, leaving us with 4,937 games and 47 variables per game. At the game level, we aggregated 10 performance statistics: yards per carry, yards per attempt, yards per first down, third down conversion rate, and red zone scoring percentage – both for the home team's offense, and allowed by the home team's defense. We also created 3 margin variables – turnover, punt, and sack margin – which consider the home team's performance on both ends of the ball in each category. Along with this, we created variables to determine if the game was an AFC, NFC, or divisional match up. Other variables were created to determine specific game conditions such as temperature, wind speed, whether the game was played on a grass or turf field, and whether the game was played indoors or outdoors. We also created more nuanced variables such as if the home team was a Vegas spread underdog, if the home team received the first half kick-off, and if the home team scored two or more touchdowns. In the end, we eliminated some variables due to high levels of correlation and model trial-and-error.

Finalized Variable List Used in Modeling

- Home team's **Yards Per Carry** - both for the offense, and allowed by the defense
- Home team's **Yards Per Attempt** - both for the offense, and allowed by the defense
- Home team's **Yards Per First Down** - both for the offense, and allowed by the defense
- Home team's **Third Down Conversion Rate** - both for the offense, and allowed by the defense
- Home team's **Red Zone Scoring Percentage** - both for the offense, and allowed by the defense
- Home team's **Turnover Margin** - the difference in takeaways on defense, and turnovers on offense
- Home team's **Punt Margin** - the difference in forced punts on defense, and punts on offense
- Home team's **Sack Margin** - the difference in forced sacks on defense, and sacks on offense

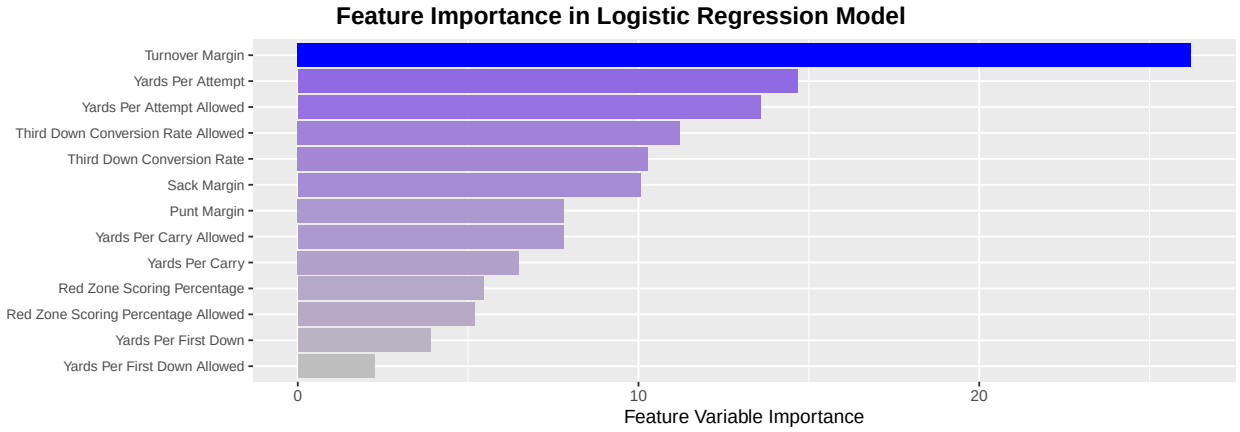
After testing numerous variable combinations, we decided that the optimal model consisted of only the 13 aggregated home team statistics. We tried hundreds of variations – binary only, continuous only, binary/continuous combinations – before coming to this conclusion. Sticking with the aggregated statistics

leaves an even split of positive and negative correlation indicators, and also alleviates the issues we saw with variable significance discrepancies in other iterations.

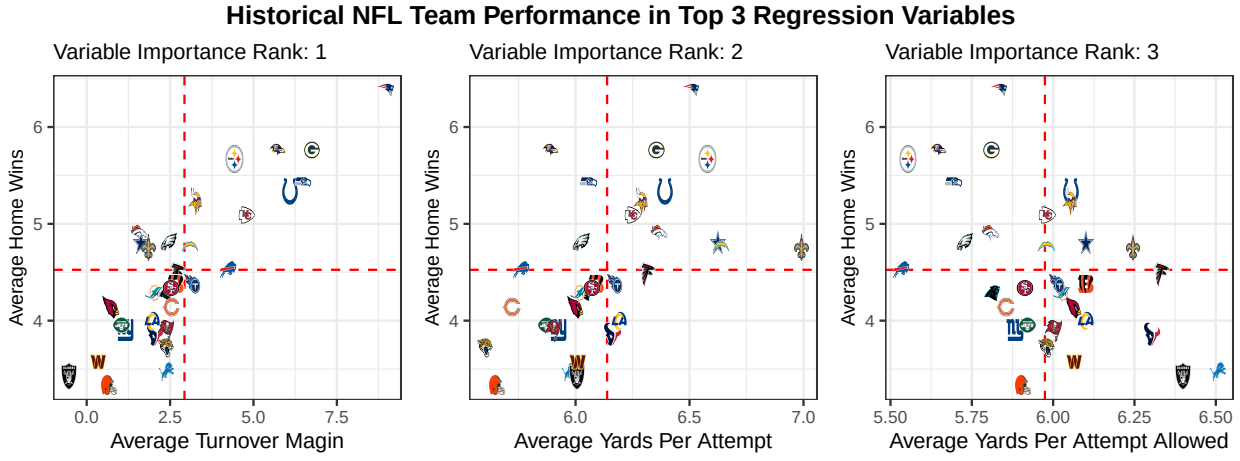
Logistic Regression Model

The created model is able to predict the outcome in games over the past 21 seasons with an 85% accuracy rate. In terms of its ability to predict home team wins, out of the 2,777 home team wins in the sample, the model correctly predicted 2,441 wins, or nearly 88% of the occurrences. In predicting home team losses, out of the 2,160 home team losses in the sample, the model correctly predicted 1,773 losses, or nearly 82% of the occurrences. With these numbers in mind, the next step in understanding the model output was to determine which variables are strongly influencing these predictions.

Feature Variable Importance: In the image below, all model variables have been ranked in relation to their overall impact on home team win probability.



Historical NFL Team Performance: One may wonder, "How has each NFL team performed in these specific metrics over the past 21 seasons?"



The graphs above show all 32 NFL teams plotted individually. The placement of the team logo holds value both vertically and horizontally. In all graphs, the vertical placement of the logo signifies the team's

per-season home win average from 2002 to 2022. The horizontal placement signifies the team's per-season average in each of the model's 3 most important variable metrics. The dashed red lines indicate the NFL average.

- In the **Variable Importance Rank: 1** graph, there is a positive linear relationship between a team's home wins per season, and their turnover margin. The teams with significantly more takeaways than turnovers have had more wins per year over the past 21 years.
- In the **Variable Importance Rank: 2** graph, there is a positive linear relationship between a team's home wins per season, and their quarterback's average yards per attempt. The teams with quarterbacks who threw efficiently have had more wins per year over the past 21 years.
- In the **Variable Importance Rank: 3** graph, there is a negative linear relationship between a team's home wins per season, and their defense's average yards per attempt allowed. The teams whose defense held opposing quarterbacks to inefficient passing attempts had more wins per year over the past 21 years.

Situational Model Application

This model could be brought into play in a range of real-world scenarios. Namely, it helps quantify the impact of specific performance based metrics on home team win probability. The following examples show the model being used to answer questions from two perspectives: the NFL super fan, and an NFL defensive coordinator.

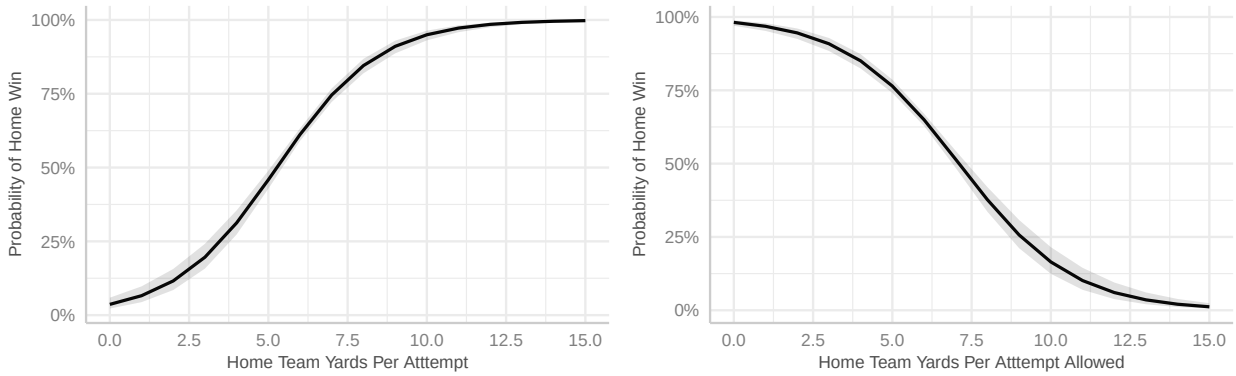
- **Note:** In order to effectively analyze the marginal effects of a single model variable on home team win probability, we must hold all other variables at a constant. We have chosen each variable's 21-year mean as the constant.

NFL Super Fan: "How likely am I to see a Bengals win at the next home game I attend if Joe Burrow throws 3 interceptions? What if Trey Hendrickson gets 2 strip sacks?"

Home Team Turnover Margin	Probability of Home Win	Change in Win Probaility
-3	0.091	0.000
-2	0.191	0.100
-1	0.358	0.267
0	0.569	0.478
1	0.757	0.666
2	0.881	0.790
3	0.946	0.855

The relationship between a home team's **Turnover Margin** and their win probability is the strongest in terms of predictive impact in our model. As the table above shows, when the home team's turnover margin is -3 (meaning they lost possession of the ball three more times than they took it away), the probability of them winning the game is only 9.1%. Conversely when the home team's turnover margin is 3 (meaning they took the ball away from their opponent three more times than they lost it), the probability of them winning the game is significantly higher at 94.6%. Furthermore, win probability increases by 79% when comparing turnover margins of -3 and 2. This information would be useful for a fan of a team to know prior to attending their next home game. It provides them with a clearer understanding of the effect turnovers and takeaways play in the outcome of their favorite team's game.

NFL Defensive Coordinator: "This season our defense has leaned heavily on man coverage designs, and our cornerbacks are getting beat on outside go routes for explosive plays far too often. Would switching scheme to eliminate the deep ball give us a better chance of winning at home this week?"



As shown by the graphs above, a higher **Yards Per Attempt** and a lower **Yards Per Attempt Allowed** lead to major increases in home win probability. For instance, a home team upping their yards per attempt from 2.5 to 7.5 yards and lessening their yards per attempt allowed from 7.5 to 2.5 yards results in 64.8% and 48.4% increases in their chance of a home win, respectively. Understanding how both of these metrics impact home win probability could help NFL coaches optimize their scheme during pre-game preparation. A defensive coordinator could elect to adopt more deep zone coverage plays; allowing for shorter passes like slants and screens, but limiting deep ball opportunities. In theory, this would lower the average yards per attempt and greatly improve win probability.

Model Limitation and Endogeneity

This model has a few limitations that can affect how accurate our feature variables are. One of the issues is endogeneity, which is when a variable is tied to the error term and gives biased and inconsistent estimates. For example, things like stadium volume level, weather conditions, and game attendance could influence the game's outcome, but this model doesn't take those into account. However, when more variables are taken into account, conflicting variables and multicollinearity can become a problem.

The model is also a bit broad, as it doesn't consider specific team match ups. This means the model can't be used predict the outcome of a particular game featuring specific opponents – such as the Cowboys versus the Eagles – as the data is aggregated only to show home and away team metrics. It's better used for getting a general idea of how home team performance statistics have affected game outcomes over the past 21 years.

Finally, the model doesn't take into account factors that are difficult to quantify such as hurt players, or unique circumstances that may help motivate a team to win. Player injuries are a big deal since the absence of a player could change the game's outcome and the variables we chose to study. If we had access to player injury data we could create a new variable that weighs how important the absence was, thus resulting in less variable bias. Beyond this, unique circumstances such as the 2022 Damar Hamlin injury and the September 11 attacks could play a role in a home team win. Though a rare occurrence, it could be argued that these traumatic situations may have impacted a team's motivation to win and the following outcome.

In summary, while this model provides significant information and interpretive results, it is subject to limitations that may compromise its situational use.

Conclusion

In conclusion, this report aimed to estimate home team win probability in NFL games over a 21 year span using 13 different performance based variables. Our model predicted with 85% accuracy the outcome of these past games. The variables determined to have the strongest impact on home team win probability were turnover margin, yards per attempt, and yards per attempt allowed. Future improvements to the model would include creating and testing more nuanced variables such as injury impact. The model could also be refocused to the play level, which would open the doors to deeper scheme and scenario analysis.