APA - HW2

Henry Lee(48097716), Jenna Li(47138355), Shengkai Zhang(48097924), Wenjun Li(48097742), Yifan Cheng(44230268)

2019/11/19

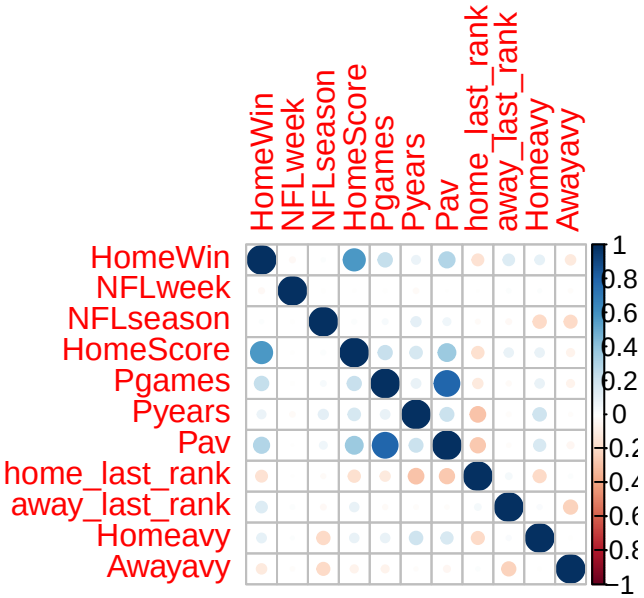
Introduction

In this report, we will showcase a dataset of all the regular season plays from the 2009 - 2016 NFL seasons and build a model to predict the home win probability for the next season. The data consists of 356,769 play-by-play observations across over 2,000 games in 7 years, and contains information about game situation, date information, players involved, results, etc.

The Data

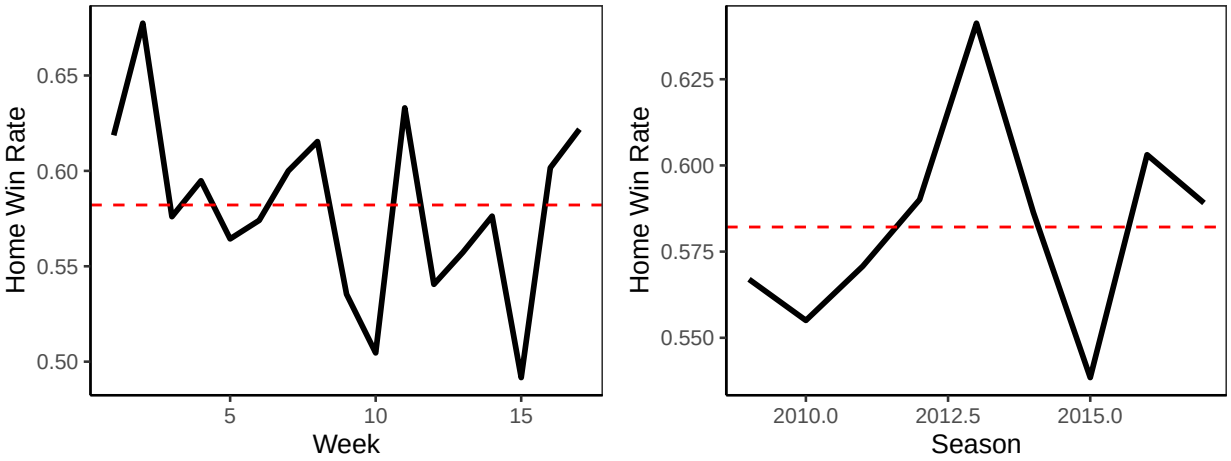
The raw data is able to provide us with information from both play-by-play and gameby- game levels, such as home/away scores, weeks, seasons and detailed plays technical analysis. Using existing data, we are able to derive to other potentially useful metrics, such as, team rankings (ranked by the number of wins in previous season), the number of games quarterback has played in the season, the career length of all team members in a team and the career length of the quarterback. In order not to over-complicate the model but still consider the impact of each individual play to the home win probability, we decide to use the existing player value column in the dataset, as the performance analysis is embedded in the metric.

As preliminary exploratory data analysis, we test the correlations between Home Win Probability and all the other variables discussed above in order to roughly understand the significance of the potential model variables. Note that the home team score, the number of games quarterback has played in the season ("Pgames") and value of players ("Pav") stand out, while away team ranking and home team average career length ("Homeavy") seem to be marginally correlating with the predictive variable. This table becomes our starting point and these variables will be considered when testing our predictive model.

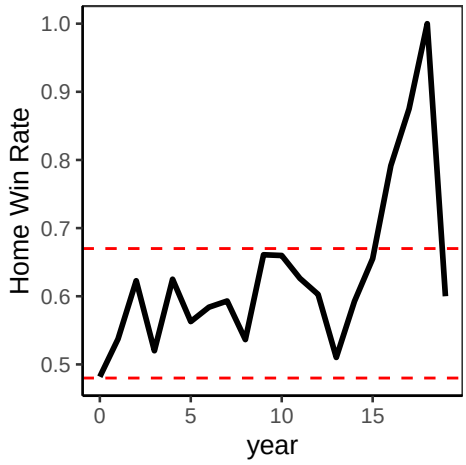


While the variables that stand out in the correlation table certainly make sense, we would like to make sure that other useful variables are not left out without reason and ensure that our first-round candidates are valid with proper justification. Thus, for the next step, we try to visualize if certain pattern exists in the variables studied.

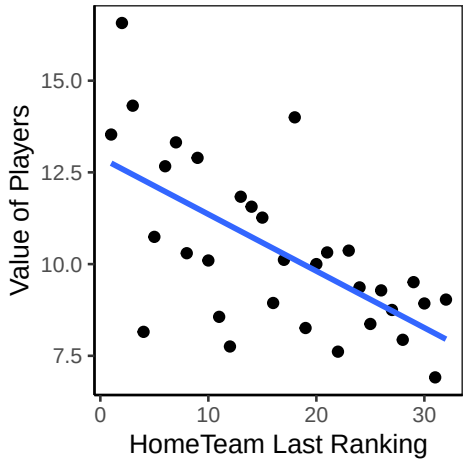
According to the correlation table, week and season (the variables that represent the year and how late we are in the season) seem to show no correlation at all. The two graphs below show their relationships with home win probability. At a first glance, one would notice that there might exist certain patterns, however, because of home advantage, the long-term average home win probability is quite significantly higher than a coin flip (56.875%, marked by the reference line). Therefore, the relationship looks no more than a normal fluctuation, thus confirms the non-correlation.



Intuitively, one might think that the career length of the quarterback would play a role in the probability of winning, and the graph below certainly seems to confirm such assumption. The shocking revelation that when the quarterback has been in the business for 19 years, the team is certainly going to win, draws our attention. Then we found that not surprisingly, that point up on the sky represents Tom Brady. His legendary 19 years has made him an outlier in our dataset. The rest of the data points are within a normal variation from the average home win probability (56.875%).



Lastly, since the value of players ("Pav") is shown significantly correlated with our predictive variable, we would like to test it from another angle - its relationship with the team's previous year ranking. The graph below has shown a clear downward trend for ranking when value drops. This has confirmed the statistical significance of the variable.



The Model

Based on exploration of the data and iterated development informed by regression diagnostics, the final model estimates home win probability in a logistic regression model using the following variables:

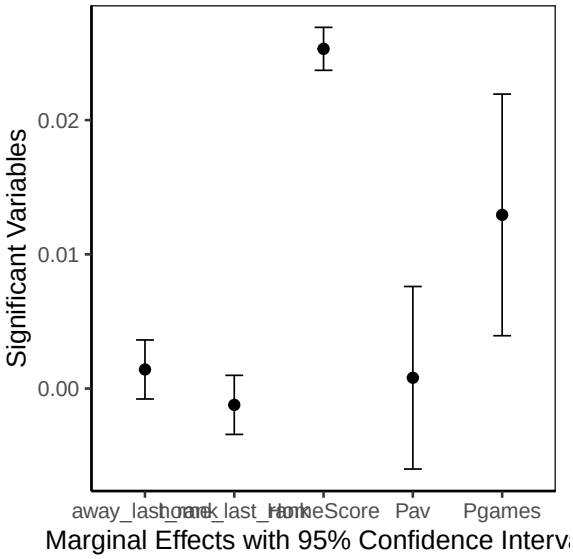
1.Home team; 2.Away team; 3.Home team scores; 4.Number of games quarterback has played in a season 5.Value of quarterback in a team; 6.Home team ranking in the previous year; 7.Away team ranking in the previous year

It is to be noted that the weather conditions, temperature and audience turnout were also thought to be included in our model. However, we believe the factors affect the away team the same way they affect the home team, so we decide to remove these factors from considerations, for the factors above might become confounding variables which bring undesirable effects to the model.

Initially we have tested a model that considers all the possible variables as well as all the players' names. Although the model is able to yield the highest accuracy among all our models (82%), the AIC, the statistical estimator of the quality of the model does not seem to confirm the model's strength. Then we dropped all the play names for the next model, as we believe the divergence between high quality and poor quality is caused by overfitting. After removal of the player names, the quality of the model as measured by AIC has improved significantly while accuracy only drops marginally. (81%). However, we believe that the model might have used more variables than it should have, as too many variables might mean higher cost and longer time for data gathering and cleansing, plus, they might increase the chance that intertwined uncertainties among variables hurt the model. Therefore, at the end, we only included the variables that have shown up significant in the correlation table plus the rankings that we derived from the raw data. While the number of variables have halved, the quality and accuracy of the model almost stay identical to the previous model. When used to predict the game results for the next year (2017), the model achieves an 80% accuracy.

The overall predictive power of this model is strong. The result of the 2017 Super Bowl has been successfully predicted, as Brady-led New England Patriots defeat Atlanta Falcons; plus, the model gives a whopping 98% chance of winning if Patriots encounter Browns. Sorry, Cleveland fans, data science is science.

The following is the marginal effect plot:



Insights

Home Advantage
Home-field advantage is significant in professional football, which is noticeably larger than coin flipping over the long term.

Momentum over Monetary Value
While the value of the quarterback plays a role in the chance of bringing home trophy, the correlation is relatively marginal. However, the number of games the quarterback has played has a quite significant impact on the win probability. In other words, maintaining the momentum of the quarterback is a priority of the team, and injury might damage the chance of winning.

Mean Reversion
Intuitively, one would think that previous years' top performers would extend their streak. While strong teams usually remain robust, the pattern is far from being iron clad. Remember that one time when the Giants who after a strong 11-win season in 2016, Big Blue plummeted to 3-13 in 2017? And the Jaguars went from the worst team in 2016 to having a 10-point lead in the AFC Championship Game against the Patriots in 2017? All that glitters is not gold; every dog has its day.