Applied Predictive Analytics NFL Game Outcome Prediction

Team Lombardi

12/1/2021

Prepared by:

- Collin Guidry (48501146), Ashley Jurak (48501172), Jiten Mistry (48666610), Priyanka Raj (48682704), Kevin Wolf (47616989)

Introduction:

- This report is meant to explain the relationship between the probability that a team will win (at home) and various NFL team cumulative season statistics, such as the difference between the total number of interceptions and fumbles between the home and away teams. By identifying significant indicators of success, a team can improve its strategy and areas of focus to improve outcomes. Ultimately, we hope fans will use this information to make informed betting decisions when wagering on the game.

Data Preparation

- The NFL data used in this analysis is presented at the game-level, representing a matchup between two teams, as well as at the play-level. By aggregating data on the play-level, the total yards or interceptions a team produces can be calculated for each game. Once the play-level statistics are aggregated to the game-level, the cumulative sum throughout the season is calculated for each game. For example, the total interceptions from all previous games in a season can be totaled for both the home and away teams. The difference of these cumulative season statistics is then calculated between the home and away teams to create a variable of comparison for each matchup. For the first week of the season, totals from the previous season are referenced. This approach, of calculating the differences of cumulative statistics between teams, is necessary for capturing information that is only available prior to a matchup and comparing both teams.

Assumtions when predicting a home team's win

- Teams have a 56% chance of winning when playing at home. (n=2,936)
- On average, home teams win by a 2 point spread.
- On average, odd-makers predict home teams will win a 2.2 point spead line.

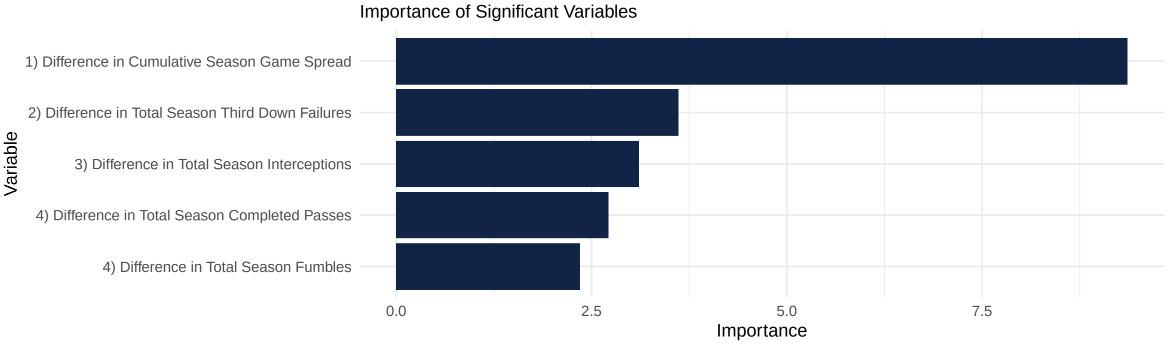
Model Variable Selection

- The **spread line**, a game spread prediction generated by oddsmakers, is the single best predictor when creating the highest accuracy model, yet it does not explain how individual team performance metrics contribute to wins. As the spread line does not allow us any practical interpretations, we chose to base our model on season performance metrics instead. These insights help us better understand which variables might have an explanatory effect on home win probability.
- Once we gained this understanding and computed correlations, highly correlated variables were selected as candidates to begin building a model with the highest level of predictive accuracy.

Endogeneity of Variables When considering the variables to be used in our model, the oddsmakers' spread line was by far the most significant in predicting win probability. When the spread line is included in the model in addition to season statistics, illogical relationships are introduced due to endogeneity and multicollinearity. Because the spread line is generated with season statistics, such as total season touchdowns, the two metrics are correlated with each other. Total touchdowns scored in a season is significantly and positively correlated with win probability. When used in a model with spread line, a team's total touchdowns is significantly and negatively correlated with win probability, which defies logic. Spread line was removed to address the issue of endogeneity it created when used with other variables. Variables were only included if they were both statistically significant and contributed a logical relationship to the model.

Variables of Importance The following combination variables were selected for the model, in order of highest importance.

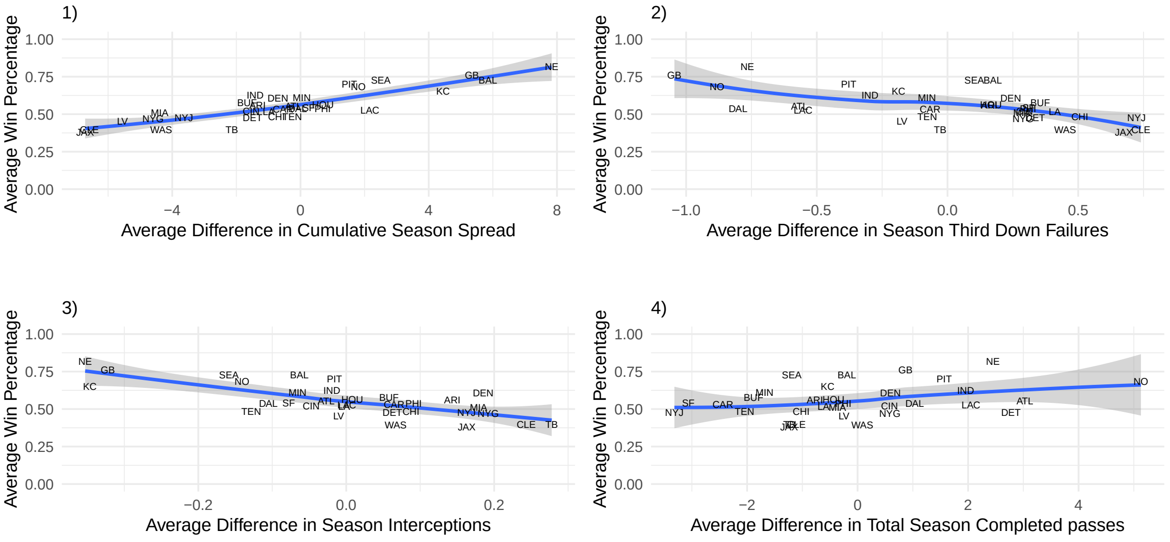
1) Difference in both teams' **cumulative game spread** thus far in the season (CGS)
2) Difference in both teams' **total third down failures** thus far in the season (TTDF)
3) Difference in both teams' **total interceptions** thus far in the season (TI)
4) Difference in both teams' **total completed passes** thus far in the season (TCP)
5) Difference in both teams' **total fumbles** thus far in the season (TF)



Each variable's contribution towards predicting home win probability is quantified in the figure above.

Variable Relationships

- How do these variables correlate with a team's percentage of home wins?

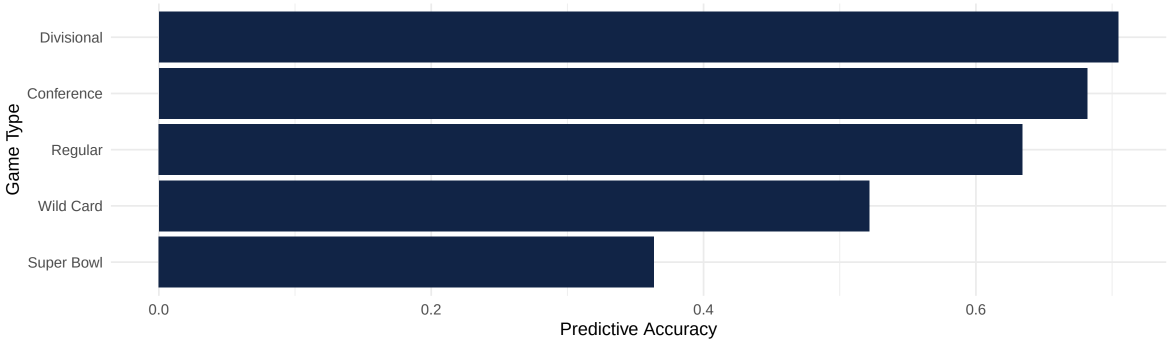


Each NFL team is plotted in the figures above. Teams towards the top of each plot have higher home win percentages.

1) The **Difference in Cumulative Spread** graph (Figure 1) shows the totaled points the each has won or lost by throughout the season compared to its opponents. This relationship indicates that teams with a higher CGS than that of their opponents are more likely to win home games.
2) The **Difference in Third Down Failures** graph (Figure 2) shows the total third down failures each team has compared to its opponent. This relationship indicates that teams with fewer TTDFs than the opposing team are more likely to win home games.
3) The **Difference in Interceptions** graph (Figure 3) shows the total interceptions each team has compared to its opponent. This relationship indicates that teams with fewer total interceptions then the opposing team are more likely to win home games.
4) The **Difference in Completed Passes** graph (Figure 4) shows the total completed passes each team has compared to its opponent. This relationship indicates that teams with a greater number of completed passes than the opposing team are more likely to win home games.

Logistic Regression Model

Model Accuracy and Limitations The model can accurately predict game outcomes 63% of the time, which is 12.5% more accurate than assuming the home team will win each game. NFL game outcomes are not easily predicted by high-level statistics, especially as for teams of similar skill are matched. As shown in the figure below, We do not recommended using this model for super bowl or wildcard games.



Model Use Cases - How could the model be used?

1) **To quantify the relationship between a team's statistics and chances of winning.**

- For example, if our cumulative spread record is X points greater than our opponent, how much does this increase our odds of winning?
- **Assumptions**:
  - To estimate the effects this variable alone, we must assume all other differences in season performance (between us and the opponent) are treated as the average.
  - As this situation is hypothetical, we assume the opponent's records is unknown and treated as arbitrary, and that our prediction is based solely on how much better than the opponent our team's record we will be leading up to the game.
  - We must assume this game is played at home. If it were an away game, the model would need to be used and interpreted differently.

Cumulative Season Spread Points Greater Than Opponent	Win Probability at Home	Win Probaility Increase
-1	0.57	0.00
20	0.76	0.19
41	0.89	0.32

- In this table, we assess the probability of winning a home game when increasing the difference in cumulative spread (between us and our opponent) by three touchdowns (+21 points).
- The first row represents the base case, where the home team has a 57% probability of winning by default.
- The second row's prediction indicates that a 21 point greater cumulative spread than the opponent will result in a win probability of 76% (+19% from the base).
- The third row's prediction indicates that a 42 point greater cumulative spread than the opponent will result in a win probability of 89% (+32% from the base).
  - Note: The incremental 21 point spread change from the second to third row only increased win probability by 13%, rather than 19%. The incremental impact on win probability diminishes.

2) **To quantify what level of performance a team requires to win a matchup.**

- For example, if we desire to have an 80% probability of beating an opponent later in the season, how many fewer season interceptions than the opponent should we strive for to achieve this?
- **Assumptions**:
  - The act of reducing interceptions alone is not responsible for a game's outcome, rather the collective effort of improving factors of the team to reduce interceptions will improve outcomes.
  - Same assumptions as the previous use case.

Fewer Season Interceptions than Opponent	Win Probability at Home	Win Probaility Increase
0	0.57	0.00
6	0.83	0.26

- The prediction in the table above indicates that in order to achieve an 80% win probability, our team must have 6 fewer total season interceptions than its opponent.

Other Variables to be Considered to Negate Bias

- Our experiment is designed to compare the difference in how two teams performed on the field before a game, to predict the outcome of a game. What is not accounted for is the difficulty of previous matchups these teams endured. If team 1 faced weaker or stronger teams than team 2, the season statistics will not reflect the effort required to these win games leading up to the matchup. An ELO rating system would be the appropriate tool to negate this, by adjusting a zero-sum ranking index for each team following a matchup. If a poorly ranked team upsets a highly ranked team, their ranks will be greatly impacted as opposed to two teams of similar skill. Though effective, the use of this model would add complexity to the interpretation of our analysis and falls outside of our scope.

Conclusion Ultimately, the most useful statistics for fans to look at to determine win probablility are the cumulative difference in spread, third down failures, interceptions, and completed passes between the home and away teams on a cummulative basis throughout the season. The model is useful to help quantify a team's chances of winning and the level of performance likely required to win a game. Although not included in this model, it would be interesting to consider the level of difficulty of each matchup to create a model of higher accuracy. The accuracy of our model is 63%, which we believe to be the strongest model based off the data provided.