Laken Rivet

Sports Performance Analytics

February 4, 2023

NFL Win Probability Model

Part One

For this individual assignment, I created an NFL win probability model using logistic regression in R. I chose to create a logistic regression model for two reasons. First, of all predictive machine learning models, logistic regression is arguably the most straight forward and the results are easy to interpret. That is, each explanatory variable will be assigned a coefficient that measures the impact on the response variable. Further, R also provides a p-value for each explanatory variable that can be used to gauge the statistical significance of the respective variable's impact on the response variable. Second, there is vast documentation of successful logistic regression models that predict win probabilities of NFL teams utilizing play-by-play data (Grosenbach 2021; Hill 2017). For these reasons, I decided to move forward with a logistic regression model.

The first step of creating my win probability model was data acquisition for model training. Play-by-play data was obtained using the load_pbp function from the nflfastR library. Only play-by-play data from the 2011 to 2021 seasons was acquired to account for recent trends in the league. For example, the "home-field advantage" that previously made a huge impact on game results has been consistently diminishing in recent years and was exacerbated by the 2020 pandemic (Wronka 2022). The play-by-play data was then filtered for regular seasons games as this model is intended for use during the regular season. Similarly, data containing information regarding the result of each game was obtained using the load_schedules function from the nflreadr library. Again, data for only regular season games in the 2011 to 2021 seasons was obtained.

Once both data sets were imported, they were merged to create the final data frame for training and testing the model. The final data frame was then filtered to remove non-plays and rows containing null values. The data frame was then split with 80% of rows being allocated to the training data set and 20% of rows being allocated to the test data set. As the names imply, the model was trained on the training data set and then tested on the test data set. Variable selection took place using the training data set.

Selecting the explanatory variables to include in the model was an iterative process of adding new variables, evaluating summary results, then removing variables, if necessary. As suggested, I first included the quarter, down, yards to a first down, seconds remaining in the game, and yard line. Prior to passing the quarter, down, and response variable to the model, they first were converted to factors and filtered to remove “NA” values. The initial summary of the model calculated p-values of less than 0.05 for all the explanatory variables, suggesting changes in their values are related to changes in the response variable’s values. Thus, they should be kept in the model. I then went through the variables provided in the play-by-play and game result data and selected additional variables that I thought may impact win probability. Two variables that I selected proved to have p-values greater than 0.05 - penalty and rush attempt. Both were binary variables that were indicative of if either of the respective events occurred during the play. Ultimately, their p-values were too high, and they were not included in the model. The remaining binary variables I selected that were included in the final model are tackle for loss, interception, sack, and pass attempt. Finally, the non-binary variables of offensive team time outs remaining and defensive team timeouts remaining were also included in the model. All additional explanatory variables were also converted to factors and filtered to remove “NA” values prior to being passed to the model.

In addition to the logistic regression model, I also created a random forest win probability model. I chose to create the second model because I wanted to see if a more advanced machine learning method would produce results similar to logistic regression. I created the model with the ranger library, and 1000 trees were included. To ensure a fair comparison, I used the same explanatory variables and data sets. A comparison of the resulting win probabilities generated by the logistic regression model and the random forest model can be observed in Figure 1.

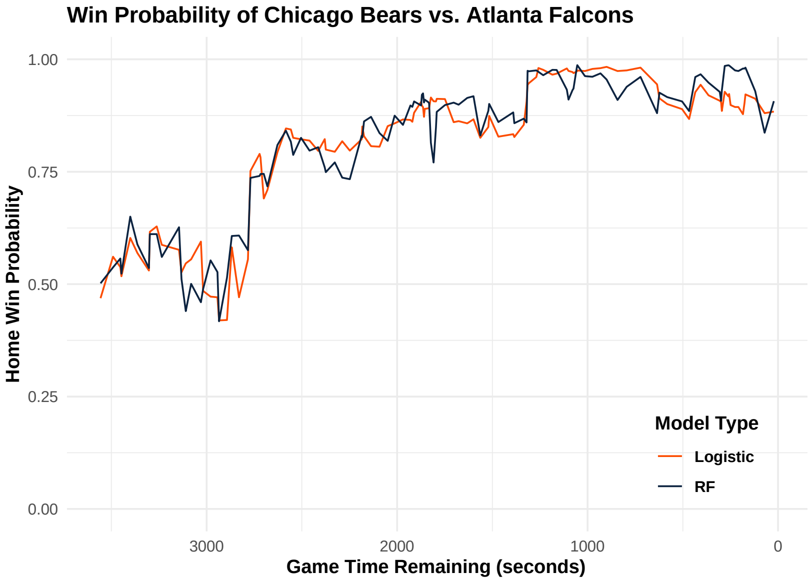

Figure 1. Comparison of win probabilities for the Chicago Bears throughout the September 11, 2011 match-up against the Atlanta Falcons.

The two models follow the same general trends, with the random forest model appearing to be more erratic in certain areas. Overall, I was pleased to see the models performed similarly and moved on to evaluating accuracy.

Accuracy of the models was evaluated visually by comparing the resulting home win probabilities of the models to those calculated by the nflfastR library. Comparison against nflfastR was chosen because their model, based on the gradient boosting machine learning technique, has been extensively tested and calibrated (Baldwin 2021). A comparison of the models I created versus the nflfastR model can be observed in Figure 2.

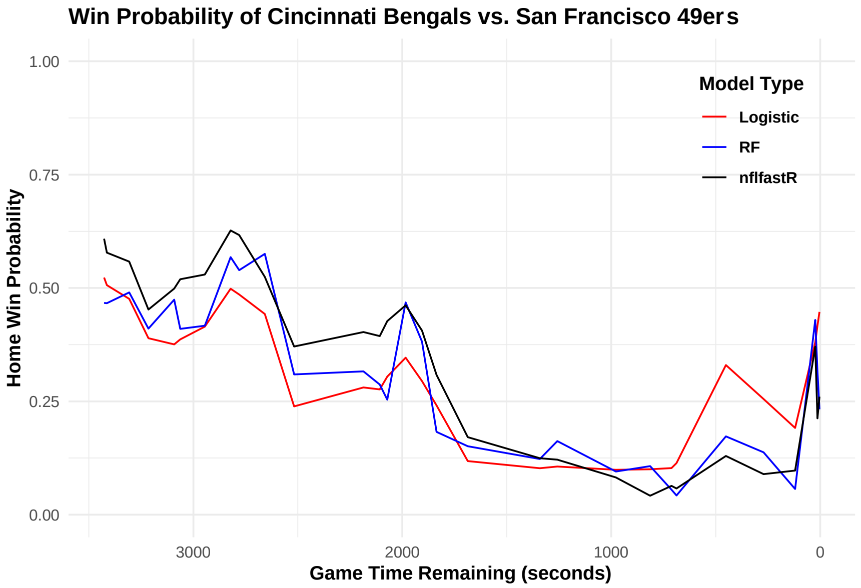

Figure 2. Comparison of win probabilities for the Cincinnati Bengals throughout the December 12, 2021 match-up against the San Francisco 49ers.

All three models follow the same general trends, which is a good indicator that my models can predict win probability decently well. Interestingly, in the first half of the game both of my models under-valued the Bengals' win probability compared to nflfastr but over-valued their win probability towards the end of the game. Of the two models I created, the random forest model performs more like the nflfastr model than the logistic regression model. With these observations in mind, I will perform the analysis in the following questions on the random forest model.

Part Two

To evaluate the impact of individual plays on the win probability model, data from the Washington Commanders versus the Kansas City Chiefs game that took place on October 17, 2021, was passed through the model. The data contained 20 input variables with information regarding the quarter, down, yards to go until a first down, seconds remaining in the game, yard line, score differential, loss of yards, sacks, interceptions, pass attempts, offensive time outs remaining, and defensive time outs remaining for each play. A visualization of the win probability for the Commanders can be observed in Figure 3.

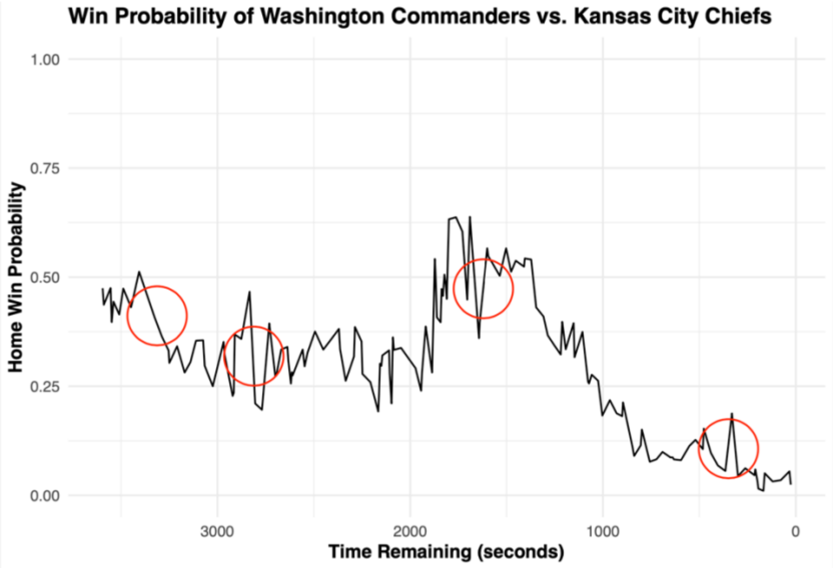

Figure 3. Comparison of win probabilities for the Washington Commanders throughout the October 17, 2021 match-up against the Kansas City Chiefs.

The visualization has four red circles that highlight large shifts in the win probability to examine. Starting from the leftmost circle, we see a significant decrease in the win probability from about 50% to 30%. The plays that correspond to this large decrease are a series of successful pass plays by Kansas City's P. Mahomes for 19 and 27 yards, respectively, that allows the Chiefs to come within 2 yards of the goal line and run in a touchdown. Logically, it makes sense that such a quick advance down the field would have a large impact on the model, especially because pass attempts is one of the input variables.

The next red circle highlights another significant decrease in win probability for the commanders with a drop of about 25% in a single play. Again, this large decrease corresponds to a massive pass play by P. Mahomes for 49 yards. Like we observed in previous play, the model places a lot of weight on successful passing. This is because passing gives the best opportunity to quickly advance towards the goal line and reset the downs. The third demonstrates a period of sharp increase in the win probability for the Commanders from about 40% to about 55%. Interestingly, this shift corresponds to the Chiefs decision to punt on the fourth down that lands on Washington's 44-yard line. Because the model has no indicator for a punt occurrence, it simply sees a punt play as changing position on the field and possession of the ball.

Finally, the last red circle is stray spike in win probability in the midst of constant decrease towards the end of the game. When examining the corresponding data to this play, it's clear this spike is due to an incomplete pass by P. Mahomes on a second down. It's difficult to understand why this particular incomplete pass contributed so much to the win probability, but I'd imagine it's because it was so late in the game and on a second down. The further into the game a play is, the higher stakes come with advancing or not advancing position on the field. However, another incomplete pass occurred three plays later on a second down and did not have as large of an impact. The only difference between the two plays is that the first occurred in Kansas City's zone while the second occurred in Washington's. This suggests that the model attributed less importance to the second because the Chiefs were so far advanced on the field.

Ultimately examining the play-by-play impact on the model has provided a few revelations. First, to no one's surprise, passing plays are crucial in the NFL. This has been demonstrated several times over and is highlighted in chapter 18, "What Makes NFL Teams Win?", of *Mathletics* (Winston, Nestler and Pelechrinis 2022). Second, punting can increase the win probability of the opposing team. Again, this may not be revolutionary, but it's important that the model reacts in the way we'd expect it to, logically. The degree to which a punt impacts the win probability largely has to do with the current state of the game. That is, factors like position on the field, time remaining, down, and quarter come into play. Finally, incomplete passes improve the win probability of the other team. This goes hand-in-hand with the importance of passing and variability of impact based on the state of the game. It's important to note that this model is not perfect, and sometimes unusual changes in the win probability occur. However, by examining how much different types of play impact the win probability, a team can discover key insights to game-time decision making.

Part Three

As with all models, there are limitations that must be addressed for the win probability model. The first limitation stems for the data organization of nflfastR. There are instances in which nflfastR data will not match official data and thus may impact certain aspects of the model. For example, the NFL classifies scrambles as rush attempts where and nflfastR does not (Baldwin n.d.). While such differences are small, they may account for differences in models built with nflfastR. Another limitation, and arguably the most important, is that win probability models will never be able to predict win probabilities perfectly. They are an estimation based on input variables, and like all other estimations, there is a margin of error to be accounted for. A final limitation I'd like to address is computer processing power. Ideally when creating a random forest, you'd be able to increase or decrease the number of trees to find the optimal balance of fitting. However, my personal computer does not have the bandwidth to include more than 1000 trees and thus limits my ability to explore a larger model.

An area of the model that I would like to improve upon in the future is the measure of accuracy. While a visual comparison with nflfastR's model provided valuable insight into which of my models to move forward with, I'd like a more quantifiable metric. In the future I'd want to create a measure of accuracy based on Statsbylopez's article "All win probability models are wrong – Some are useful" where he examines observed win probability versus estimated win probability by quarter (2017). Similarly, I'd like to use a calibration method like Ben Baldwin outlines in "nflfastR EP, WP, CP xYAC, and XPass models" where he can calculate calibration error for a model and then optimize it based on those results.

This model can be applied to any scenario in which a probability is calculated. For example, probability of completing a pass could be calculated using input variables similar to those in the win probability model. There would need to be additional information collected on individual players such as a quarterback's pass completion percentage and a receiver's catch rate. Another application of this model could be in aiding the decision to go for it or punt. This could be done by revisiting past decisions to go for it or punt and examining how those choices impacted win probability. For example, previous game data could be filtered and organized by punts in different quarters. The average impact of these punts on the win probability could then be calculated to provide an estimate of how a punt may impact the current game. The same can be done for decisions to go for it, which could be further categorized by type of play used in the drive. Ultimately, this model is incredibly useful because it allows teams to use past data to aid in future decision-making.

Part Four

In chapter 18 of *Mathletics*, the authors explore what aspects of a team's offense and defense contribute most to winning. They test Bud Goode's 1960 discovery that yards per pass attempt on both offense and defense were the most important factors in predicting a team's success. They do so using a linear regression (without an intercept) that estimate a team's scoring margin (points for – points against) using eight independent variables. The variables include offensive yards per pass attempt (PY/A), defensive yards allowed per pass attempt (DPY/A), offensive yards gained per rush (RY/A), defensive yards allowed per rush (DRY/A), turnovers committed (TO), defensive turnovers (DTO), penalty yards difference (penalty yards committed by team – penalty yards committed by opponents) (PENDIF), and return touchdown difference (return touchdowns – return touchdowns by opponent) (RET TD). The authors also provide a small discussion about whether a good rushing attack actually sets up a good passing game. To prove this mathematically, they create a correlation matrix of the independent variables used in the linear regression.

I’ve compiled the same statistics for the 2022 regular season and ran the same regression to gain insights for the upcoming NFL season. The output from the regression can be seen in Table 1.

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0	#N/A	#N/A	#N/A
PY/A	59.70145948	8.759373329	6.815722683	4.75496E-07
RY/A	20.08845225	17.61189244	1.140618609	0.265283483
PENDIF	-0.044434257	0.072299961	-0.614582039	0.544613868
TO	-6.300702792	1.611478801	-3.909888724	0.000661446
DTO	5.664418592	1.366418881	4.145448129	0.000364622
DPY/A	-55.59149456	11.1497175	-4.985910589	4.30798E-05
DRY/A	-22.87590816	12.47065482	-1.834379068	0.079025712
RET TD	17.54884	9.045792765	1.940000225	0.064218213

Table 1. Linear regression output, without an intercept, estimating point differential for the 2022 NFL season.

Notably, the only variables with a p-value low enough to be considered statistically significant are offensive passing yards per attempt, turnovers, defensive turnovers, and defensive passing yards per attempt. It should be noted that the small sample size of a singular NFL season could be contributing to the lack of significant variables. Regardless, this reinforces the well-known idea mentioned earlier that passing is essential to a team’s success both defensively and offensively. An extra passing yard per attempt represents an additional 59 points while an extra defensive yard per passing attempt represents a loss of 55 points. Additionally, turnovers are costly mistakes. The regression equates each turnover to cost the team 6 points while every turnover we force on the opponent aids us by 5 points. I also created a correlation matrix of the 2022 NFL season variables which can be seen in Table 2.

	PY/A	*RY/A*	*PENDIF*	*TO*	*DTO*	*DPY/A*	*DRY/A*	*RET TD*
PY/A	1							
RY/A	0.23477382	1						
PENDIF	-0.1849726	0.22199602	1					
TO	-0.3846123	-0.0702949	0.25555433	1				
DTO	0.07250494	0.03770526	-0.3685243	-0.0002304	1			
DPY/A	-0.3034208	0.17312359	-0.1912737	-0.0817179	-0.0245352	1		
DRY/A	-0.119329	-0.0314502	-0.209274	-0.1264336	-0.0593241	0.21547932	1	
RET TD	0.18755252	0.35605973	-0.3400335	-0.1336845	0.04962662	0.14000383	0.0980848	1

Table 2. Correlation matrix of 2022 NFL season independent variables.

From the correlation matrix, we are looking at the relationship between each of the variables. Specifically, we want to consider passing yards per attempt and rushing yards per attempt to test the theory that a good rushing attack sets up a good passing game. For the 2022 season it was observed the two variables only had a correlation of about 0.23, which hardly suggests rushing sets up passing.

This analysis provides our organization with some key areas to focus on for the 2023 season. First and foremost, we should prioritize our passing game. This can be done in a multitude of ways. For example, we could instruct scouting staff to focus on identifying the best quarterback and receiver prospects. We could instruct offensive and defensive coaching staff to run more passing plays in practice. Similarly, coaching staff should focus on turnovers. That is, communicating with players the importance of completing passes and preventing fumbles as well as practicing strategies to force turnovers on opponents. Finally, discussions with the front office could be held regarding the proper allocation of resources in regard to the regression results. This means prioritizing spending on players who will contribute the most to significant variables.

References

Baldwin, Ben. 2021. “nflfastR EP, WP, CP xYAC, and XPass models”. Open Source Football, February 4, 2021. https://www.opensourcefootball.com/posts/2020-09-28-nflfastr-ep-wp-and-cp-models/.

Baldwin, Ben. n.d. “A beginner’s guide to nflfastR”. nflfastR. Accessed on February 5, 2022. https://www.nflfastr.com/articles/beginners_guide.html.

Grosenbach, Geoffery. 2021. “An NFL Win Probability Model Using Logistic Regression in R”. TopFunky, January 21, 2021. https://topfunky.com/2021/logistic-regression-nfl-win-probability/.

Hill, Stephen. 2017. “Building a Basic, In-Game Win Probability Model for the NFL”. Medium, December 29, 2017. https://medium.com/@technocat79/building-a-basic-in-game-win-probability-model-for-the-nfl-54600e57fe1c.

Hill, Stephen. 2018. “Enhancing our Basic, In-Game Win Probability Model for the NFL: Random Forests”. Medium, January 23, 2018. https://medium.com/@technocat79/enhancing-our-basic-in-game-win-probability-model-for-the-nfl-random-forests-f9e8bb40583e.

Statsbylopez. 2017. “All win probability models are wrong – Some are useful”. StatsbyLopez, March 8, 2017. https://statsbylopez.com/2017/03/08/all-win-probability-models-are-wrong-some-are-useful/.

Winston, Wayne L., Scott Nestler, and Konstantinos Pelechrinis. 2022. *Mathletics: How Gamblers, Managers, and Fans Use Mathematics in Sports*. New Jersey: Princeton University Press.

Wronka, Shawn. 2022. “What NFL Bettors Need To Know About Home-Field Advantage Before The 2022 Season”. Covers, August 22, 2022. https://www.covers.com/nfl/home-field-advantage.