Estimating Win Probability for NFL Games

John Ruscio and Kevin Brady

The College of New Jersey

Draft Date: 1/20/2021

This manuscript contains analyses completed as part of Kevin Brady's senior honors thesis, completed in the Spring, 2018 semester. The most recent data available when we began this project included the 2016 season. We intend to update the analyses and revise this paper to submit it for publication at some point. For the moment, we provide this draft to help users of our shiny app understand how it works. This app implements our revised PFR model, with one more change: We calculated the uncertainty factor using all data from all 5,324 games played in the 2000 to 2019 seasons and our app uses a value of 13.40022 rather than 13.45 (as in the original PFR model). Comments, questions, or suggestions are warmly welcomed. Please contact ruscio@tcnj.edu.

Abstract

Win probability (WP) models seek to predict the outcome of a game at any moment, allowing us to assess decision making in key situations. For National Football League (NFL) games, WP models range from the relatively simple Pro Football Reference (PFR) model to more complex random forest models. Using play data from all NFL games from 2001-2016, we found both models to be well calibrated (e.g., for all plays with a .75 probability, the team on offense ultimately won almost exactly 75% of the time). We modified the way the PFR model uses game time to improve accuracy and to handle plays in overtime. Our modified PFR model performs slightly better than the random forest model for plays throughout regulation and in overtime.

Estimating Win Probability for NFL Games

Predicting the outcome of any sports contest, such as games in the National Football League (NFL), can be extremely challenging. Fans like to pretend they can identify the defining moments of a game and use this perceived ability to question or challenge coaching decisions of their favorite teams. Coaches like to think they know what to do in key situations in order to maximize their chances of winning the game, or at least to escape blame for a loss. Identifying defining moments occurs in hindsight and can be very subjective. Is there any way to prospectively, and objectively, assess the status of a game in real time? Win probability (WP) models seek to predict the outcome of a game at any moment, given its unique situation, and give us the ability to influence, critique, or congratulate decision making in those defining moments.

Individuals have tried to develop statistical win probability models to predict sport outcomes for over half a century. The earliest models focused on baseball. For example, Lindsey (1961) estimated the win probabilities at the end of each inning, using score (run differential) and inning, while Bennett and Flueck (2005) estimated win probabilities during innings using run differential, number of outs, the half-inning, and the on-base situations.

While most popular in baseball, win probability models have become more common in football. Burke (2014) developed the first NFL win probability model using rather basic variables including score differential, time remaining in the game, down, yards to go for a first down, and field position. The exact model Burke developed is proprietary, now owned by ESPN.

Lock and Nettleton (2014) modified Burke's model by adding in the Vegas line (point spread) to quantify the relative strength of the teams prior to a game, rather than holding pre-game probabilities constant at .50. Using an NFL play database containing data from all games from 2001-2012, Lock and Nettleton also developed a random forest model to estimate win probabilities. Random forests are a machine-learning approach to pattern recognition that requires a very large sample of data and can generate predictions based on complex relationships among many variables (Breiman 2001). Designating seasons 2001-2011 as training data, win probabilities were estimated by creating 500 trees in a random forest (*ntree* = 500) that was allowed to evaluate all possible thresholds on two randomly selected variables at each branching point (*mtry* = 2); each tree was complete when no further splits could be made while still maintaining a minimum of 200 cases per terminal node (*nodesize* = 200). Data from the 2012 season were used to test the accuracy of the model.

Unless otherwise noted, our random forest analyses followed the same procedures described by Lock and Nettleton (2014), with *ntree* = 500, *nodesize* = 200, and *mtry* = 2. Lock and Nettleton arrived at these parameters through a cross-validation strategy detailed in their paper. Likewise, Lock and Nettleton examined the performance of random forests generated using many alternative sets of predictor variables. We work exclusively with the predictors that they ultimately recommended.

Pro Football Reference (PFR), a prominent football archives website, also developed a popular and accessible win probability model (The P-F-R Win Probability Model, n.d,). The PFR model works by constructing a normal curve such that the center represents the current game situation (point differential + expected points for the current drive + Vegas line, discounted as the game unfolds) and the spread represents uncertainty (variability in game outcomes relative to the point spread, discounted as the game unfolds). Specifically the equations for the PFR model's mean and standard deviation are as follows:

$$M_{\mathrm{PFR}} = (pts.off - pts.def) + exp.pts + line\ /\ (60\ /\ min.rem),$$

$SD_{\mathrm{PFR}} = \mathit{uncertainty} \times \mathrm{sqrt}(\mathit{min.rem} / 60)$,

where *pts.off* is the number of points scored by the team on offense, *pts.def* is the number of points scored by the team on defense, *exp.pts* is the expected points for the current drive, *line* is the Vegas line at kickoff (coded here as the amount by which the team on offense is favored to win, negative if they were favored to lose), *min.rem* is the time remaining in the game expressed as minutes (e.g., 60 min at kickoff, 7.5 min midway through the 4th quarter, 0 min at the end of regulation play), and *uncertainty* is the standard deviation of the difference between actual game outcomes (score differential) and Vegas lines (i.e., final game *pts.off* minus *pts.def* minus *line*). PFR uses *uncertainty* = 13.45 based on their analysis of historical game outcomes. Expected points for the current drive is calculated using a series of equations that take into account the down and distance (e.g., 2nd and 2 yields more expected points than 4th and 5) and field position (e.g., beginning a play on your opponent's 25 yard line yields more expected points than beginning a play on your own 25 yard line). Expected points approaches an upper bound of 7 (i.e., under the most favorable conditions, 1st and goal from the 1 yard line, *exp.pts* = 6.971) and has a lower bound of -2.49 (e.g., under extremely unfavorable conditions such as 4th and 10 at your own 1 yard line).

The normal curve constructed using M_{PFR} and SD_{PFR} represents the distribution of expected game outcomes, given the current game situation and the historical level of uncertainty in predicting a final game outcome. Win probability is calculated as the area under the curve above a threshold of 0.5 (which represents favorable game outcomes for the team on offense, or the probability that they will win) plus one-half the area between thresholds of -0.5 and 0.5 (which represents a tied game at the end of regulation, for which PFR assigns an equal likelihood of winning or losing in overtime).

Notice that the PFR model discounts both the Vegas line and the uncertainty factor by time. At kickoff, the Vegas line is an important indication of the teams' relative strengths, but as the game unfolds the score itself is a more revealing indicator of who is likely to win. The Vegas line is discounted in a linear fashion. For example, suppose a team is favored to win by 7 points at kickoff. If they begin the game on offense, the *M* of the PFR model adds 7 points to take this into account. If they begin the 4th quarter on offense, however, the *M* of the PFR model only adds 7 / (60 / 15) = 7 / 4 = 1.75 points. Three quarters of the game have been played, so the Vegas line has been discounted accordingly. The uncertainty factor begins at 13.45, as this is the degree of uncertainty in predicting final game outcomes using the Vegas line at kickoff. However, as the game unfolds, uncertainty is reduced. For example, suppose the team on offense leads by 8 points. If there are only 2 min left to play, one can be much more confident that this team will win than if there were 55 min left to play. Thus, the uncertainty factor is discounted by time, albeit in a nonlinear fashion.

A final note regarding the use of time in the PFR model is that this becomes problematic in overtime. Because the minutes remaining in a game counts down from 60, at kickoff, to 0, at the end of regulation, it is not clear how a reset game clock at the beginning of an overtime period should be handled. Obviously, from a strategic perspective the time on the game clock means something profoundly different in overtime. Moreover, overtime rules are different for regular-season games (which are allowed to end in a tie) and postseason games (which are not allowed to end in a tie), and the overtime rules themselves have changed substantially over time (e.g., beginning in 2012 the game does not end if the first possession in overtime results in a field goal). For all of these reasons, we believe it would be foolish to simply reset the time variable in the PFR model along with the game clock in overtime. Instead, we will not use this model to estimate win probabilities for plays in overtime.

Lock and Nettleton (2014) examined the accuracy of their model and found that it performed very well. In particular, the win probabilities were extremely well calibrated. For example, among all plays on which the team on offense was estimated to have about a .75 probability of winning, these teams ultimately won almost exactly 75% of the time. Research on the accuracy of other win probability models is limited, however. Also relatively unknown is whether the complexity of a random forest model pays off in greater accuracy than that attainable using a much simpler model, such as PFR model.

Using an NFL play database dating back to 2001, acquired from ArmChairAnalysis.com, we evaluate different statistical models of NFL win probability. The first issue we address is the accuracy-complexity trade-off among models. Does the added complexity of a model yield significant gains in accuracy of prediction? If so, is the increase in accuracy enough to warrant the added complexity, or would it be better to use simpler models to get nearly as accurate results? The answer may depend on whether a high-stakes decision is being made, or whether the model is being used by casual fans.

Lock and Nettleton's (2014) random forest model is much more complex than the PFR model. Generating a random forest model requires massive computing power to construct 500 trees, each of which contains a large number of nodes with specific decision thresholds on specific variables that differ for each tree. Using this model to estimate win probability entails providing a specific game situation to this random forest, tracing the data values through each node of a tree to find which terminal node it ends up belonging to, using the proportion of games ultimately won within this node as an estimate of win probability, repeating this process for the other 499 trees, and averaging all 500 estimates of win probability to yield a final value. Generating or using a random forest model requires sophisticated data-analytic software.

The PFR model, in contrast, entails the use of a set of fairly simple equations using data for a particular play. Once expected points are calculated based on down, distance, and field position, the *M* and *SD* of a normal curve are calculated as shown above, and the areas in two regions under the curve are used to estimate win probability. All of this can be done using a simple spreadsheet program.

The first specific aim of this study is to assess whether the added complexity of a random forest model delivers significantly more accurate predictions than the simpler PFR model. Next, we look to evaluate the amount of training data provided to random forest win probability models. The training data is used to set the historical basis for the model, and the more data the more complex and specific the forest can become. Lock and Nettleton (2014) used 2001-2011 as training data, and then used 2012 to test their model. Statistical theory suggests that, all else being equal, using more training data would lead to more accurate predictions, and we will examine the extent to which increasing the amount of training data affects the models' accuracy. However, all else is not equal because there have been significant rule changes in the NFL over time. For example, it is now easier for offenses to score points than it was in 2001 due to changes made to favor the passing game over running the ball, as well as rule changes to increase player safety. Thus, training models using older data may not increase their accuracy in predicting recent outcomes. There may be a limit to how much data is optimal to provide. We will evaluate this by systematically varying the series of training data provided.

After evaluating the accuracy-complexity tradeoff between the models, we plan to improve on one or both of the models. After analyzing the methodology of these WP models, there appear to be areas in which improvements can be made. A major goal of this research is to look for opportunities to improve accuracy of prediction by refining an existing model.

Method

Data Source

We use an NFL play-by-play database that contains all data from the 2000-2016 seasons, acquired from ArmChairAnalysis.com. This is the same database used by Lock and Nettleton (2014), with extra seasons now available. The database includes 4,790 games and 785,509 plays. All "zero down" plays (e.g. kickoffs, extra points, and two-point conversions) were dropped from the sample because the down of a play is important to each model. This brought the sample down to 714,194 plays. The 2000 season was dropped from our sample because of problems with the coding of game time, which is used in all models. This brought the sample down to 4,531 games and 675,629 plays. All tied games were dropped as well, as they do not have a winner and can't be used to evaluate the accuracy of predictions**.** Our final sample had 4,524 games and 674,311 plays. For many analyses, games that included plays in overtime were excluded. These analyses were run with 4,240 games and 669,924 plays in the sample.

Random Forest Model

For each play, the predictor variables provided to the random forest model were game score (points for offense and points for defense), Vegas line, game time (quarter, minutes, and seconds), timeouts remaining for each team, down, distance, and field position. The criterion variable was dichotomous, representing whether or not the team on offense ultimately won the game.

PFR Model

Because it does not require training with data, the PFR model was implemented as described earlier.

Results

We began by comparing the performance of the PFR model to that of the Lock and Nettleton (2014) model to evaluate the accuracy/complexity trade-off. We used the mean squared error of prediction (MSE; win probability vs. a dichotomous variable representing whether the team on offense ultimately won the game) to evaluate the accuracy of a model. The MSE for the PFR model was .15185. Note that until further notice, overtime plays are exluded from analyses.

For the random forest model, we had to determine how to allocate the training data and test data. Lock and Nettleton (2014) used the 2001-2011 seasons as training data, and then tested the model with 2012 data. We included more, fewer, and different years to assess these models across eras, rule changes, and scheme changes in the NFL. For example, we used 2016 as test data for a series of analyses that vary in years of training data beginning with just 2015, then 2014-2015, then 2013-2015, and so forth. Then, we used 2015 as testing data, and worked downward in years with a new series of analyses varying in the amount of training data. We went back as far as seven years of training data. We also did this in the other direction, training models with particular years to test previous seasons--for example, training the model with 2016 and 2015 to make “predictions” for 2014.

Because there was such a large number of random forests to generate, each of which is computing intensive, we used 100 (rather than 500) trees in each random forest. Preliminary analyses showed that MSE tended to stabilize between 50 and 100 trees. As shown in Figure 1, there was a very slight decline in MSE as training years increased. Because the trend had become clear and run time was increasing dramatically, we terminated this series of analyses at 7 years.

To analyze the MSEs obtained for all of these random forests, we performed multiple regression analyses using four predictors: (1) test year, (2) number of years of training data, (3) offset (mean year in training data minus test year), and (4) absolute offset. For example, using 2005 through 2009 as the training year and 2012 as the test year is an offset of -5, because the middle year of the training data is 5 years before the test year. This offset variable is a measure of distance that also takes into account direction. To examine distance free of the influence of direction, we included the absolute offset. This is simply the absolute value of the offset variable, and it indexes the distance between training and test data without regard to which comes first chronologically. By varying the year of test data and the year(s) of training data, we examined their influence on models' predictive accuracy to determine the optimal amount of training data to provide. What we found is that the only variable that significantly influenced model accuracy was the number of years of training data, as already depicted in Figure 1. MSE decreased with each additional year of training data provided, but there were diminishing returns.

Given these results, we estimated a single MSE to best represent the random forest approach based on all available data. This was done using an odd/even split of the games. One random forest was written to make predictions for all plays in even games using a model trained using all plays in odd games, and then a second random forest that did the reverse. This allowed predictions to be made for test data uncontaminated by training data, while using very large samples of training data. The two resulting MSEs were averaged to obtain an overall estimate of model accuracy. The odd/even split returned a MSE of .15188 for the random forest model. Given its complexity, it is noteworthy that this is slightly poorer than the accuracy level of the

PFR model (MSE = .15185). We do not believe this is a substantial difference, but it does seem surprising to find that the added complexity of the random forest does not improve accuracy.

In addition to calculating MSE, we plotted the calibration curve for each prediction model. This entails grouping plays by predicted win probability (e.g., .00 to .05, .05 to .10, …, .95 to 1.00) and calculating the ultimate winning percentage as the proportion of times that the team currently on offense ultimately won the game. To the extent that the points cluster around a diagonal line, the predictions are well calibrated. This would mean, for example, that a win probability of .75 correctly predicts a win for that team exactly 75% of the time. In addition to visually inspecting this curve, calibration can be assessed using a correlation coefficient for estimated and observed win probabilities, with each data point on the calibration curve constituting one case for this correlational analysis. Calibration curves for the random forest and PFR models are shown in Figure 2. In both cases, the correlation between estimated win probabilities and actual win probabilities, both binned as shown in the figure, was $r > .99$.

Despite thee nearly identical MSEs for the random forest and PFR models, as well as the fact that both were very well calibrated, there are nonetheless some potentially important differences in the models' win probability estimates. These can be seen by comparing their MSE as time elapsed over the course of a game. These results only include plays during regulation (i.e., not overtime), as the PFR model cannot be used in overtime. As shown in Figure 3, the PFR model is slightly more accurate early in a game, but soon after halftime and continuing through the end of regulation, the random forest model surpasses the accuracy of the PFR model. Thus, their overall MSE masks some nontrivial differences throughout a game. This observation led to the final part of our study, an attempt to improve the existing models.

For the random forest model, we could not devise any ways to improve its performance. As noted earlier, Lock and Nettleton (2014) already examined many different configuations of variables, as well as different parameters for generating the random forest itself (e.g., values for *mtry* and *nodesize*). We did not pursue this further.

We did, however, attempt to improve the PFR model. Our first approach was to replace the PFR equations for calculating expected points with a random forest model. The PFR model calculates expected points through a series of linear equations. Intuitively, this seemed like it could present problems in prediction given the potential nonlinearities and interactions among many variables. As an alternative, we generated a random forest model that predicted the observed points ultimately scored on a drive given the down, distance, and field position of the current play. Unfortunately, using the results of the random forest rather than the linear equations did not improve the overall performance of the PFR model. We also tried to use time remaining (to halftime or to the end of regulation, based on when a drive took place), an important variable that the PFR equations ignore, in various ways, to no avail. Though it may be possible to calculated expected points more effectively, we were unable to devise a model that worked better than the linear equations of the PFR model.

Next, we attempted to refine the way time is discounted in the PFR model, particularly for the spread of the normal curve. Time is discounted using functions of time remaining in regulation, and as that approaches zero the uncertainty may be discounted too steeply. Win probabilities jump dramatically, and erratically, near the end of games. If time remaining literally equals zero, the PFR model cannot be used. For example, a game cannot end on a defensive penalty, so on rare occasions there will be one or more untimed plays at the end of a game. This would causd a division by 0 error in the PFR model. Beyond that point, as noted earlier, it’s not obvious how to adapt their model for use in overtime play, and we believe it would be foolish to simply reset, time remaining along with the game clock, in overtime. Instead, in many respects all overtime plays take place in a context that's much like the final minutes, or even seconds, of regulation play: The game could end on any given play. This suggests that it might make sense to have a constant level of uncertainty that carries over from the end of regulation into overtime. With that inuition in mind, we set out to improve and extend the PFR model.

First, to deal with the way the model discounts time as it counts down to zero, and thereby stabilize predictions somewhat in order to improve accuracy late in games, we set a minimum time threshold. By systematically varying the minimum time remaining, we tried to find the "sweet spot" for predictions. Rather than allowing the time remaining to continue all the way down to 0, our modified PFR model sets a floor at 450 seconds remaining. Compare the upper-left and upper-right graphs in Figure 4 to see how this changes the time discounting function. This modified PFR model, which only allowed time to run down to 450 seconds, exhibited MSE = .1509. This indicates some improvement over the original PFR model (with MSE = .15185).

Next, we extended all analyses to include overtime play. For the modified PFR model, we set time equal to 450 seconds for all overtime plays, and this yielded MSE = .1512. This compares favorably with the MSE = .1522 obtained for the random forest model when evaluated using all plays, including those in overtime.

A few additional refinements were made to the PFR model to discount time, and thereby estimate win probabilities, as effectively as possible. First, rather than setting a floor of 450 seconds remaining, which abruptly changes the time discounting curve into a straight line, we stretched the time discounting function so that it doesn't approach 0 seconds at the end of regulation and then hit a floor, but instead it approaches 450 seconds all along. See the lower-right graph in Figure 4. Second, because probabilities were actually a bit too modest (i.e., not sufficiently extreme) very late in games, we steepened the time discounting function in the final two minutes of regulation. See the lower-right graph in Figure 4. Third and finally, we set a fixed value for time discounting in overtime (.25). When evaluated using all plays, regulation and overtime, this final modified PFR model yielded a MSE of .1511.

Discussion

Statistical theory suggests that, all else being equal, using more training data will lead to more accurate predictions, but due to rule changes and differences in play style over the years, we know that all else isn't equal. Training models using older data may not increase their accuracy in predicting recent outcomes, and there may be a limit to how much data is optimal to provide. When we put these possibilities to the test, which seasons' data were used to build and test a random forest model didn't matter much and only explained a minute amount of variance in accuracy. Outcomes were more predictable in some years than others, but there was no systematic trend across seasons. As expected, increasing the number of training years helped, but with rapidly diminishing returns.

All of this suggests that NFL rule changes since 2001, which tend to aid offenses and increase scoring, as well as the changes in offensive play calling and pace of play, have had little effect on the successful modeling of win probabilities. It was expected, and is often assumed, that "no lead is safe" in today's NFL. Pundits like to discuss how late-game comebacks are always possible no matter the deficit due to the way the game is played today. Our findings suggest that a team's chances of winning at a given point in a game are about as (un)predictable today as they have been since 2001.

Surprisingly, there was no evidence of an accuracy-complexity tradeoff between the PFR and random forest WP models. The random forest is a machine learning algorithm trained by randomly sampling subsets of the training data, fitting a model to these sampled subsets, and aggregating the predictions (Breiman, 2001). The PFR model, however, simply constructs a normal curve whose center represents the game situation (score differential, expected points for the current drive, and time-discounted Vegas line) and whose spread represents the degree of uncertainty in predicting game outcomes (also discounted for time). The expected points are derived from a series of linear equations. Statistical theory, and naïve intuition, might suggest that using the more complex model, particularly with such large training and test datasets, would attain greater predictive accuracy than using the simple model. What we found, however, was nearly identical accuracy levels for the random forest and PFR models. We conclude that, at the very least, the original PFR model is comparable in predictive accuracy to a random forest model, that there is no real accuracy-complexity tradeoff between them. Considering how much simpler and easier the PFR model is to use, the PFR model is more than "good enough" for fans and writers to use. For these fans, such as ourselves, there is nearly no benefit to using the more complicated random forest while watching a game or discussing, debating, or writing about it afterward. Instead, with a user-friendly interface, fans can accurately estimate win probabilities in real time or at their leisure.

Close inspection of results suggested that the way the PFR model discounted time limited its predictive validity late in games. After modifying the PFR model's handling of time, our modified PFR model performed slightly better than the random forest model for plays throughout regulation and in overtime. Again, with a user-friendly interface, this should be very helpful for interested fans.

In fact, NFL teams themselves might benefit by examining win probability through something like our modified PFR model. Organizations have begun to expand their analytics departments, with many teams embracing data analysis to influence many types of decision making. Of course, NFL teams have the ability to use any model they'd like, including the statistical/technological power to generate and implement random forests. However, the modified PFR model is both as accurate and easier to run. This model could aid coaches without backgrounds in data and analytics to both better understand and become comfortable with the use of win probability models. It may be more difficult to convince a football coach without a statistical background to use a complex random forest model than a comparatively simple one based on the PFR model. And considering the negative stigma many "purists" have against analytics infiltrating the game, the simpler model might be more readily accepted.

Strengths

The first strength of this study is the sample. Even after trimming the dataset, the number of plays and seasons available for testing provides an extensive sample of actual NFL game data. Our dataset expands on that of Lock and Nettleton (2014) and offers up enough information to draw conclusions on win probability estimations over more than one and a half decades. Building the models with actual NFL game data ensures that estimations of win probabilities are generalizable to NFL games.

Another strength to this study is the complete access we were granted to work with both models. PFR provided their expected points per drive equations, and the rest of their model can be found on their website in its entirety. The random forest package on R is simple to access, though more complicated to understand and run than the PFR model. Complete access to each model is a clear strength of this study, as it allows for transparency and accuracy in testing their predictive accuracies.

This access to the models leads into our next strength, the ability to compare the accuracy of distinct models freely and without bias. Due to the proprietary nature of many win probability models, it is hard to find enough information on building models to replicate them for your own research. Ganguly and Frank (2018) critiqued popular win probability models for the lack of available information for analysts and researchers to compare models to. Often, the public is just presented with a percentage chance a team will win with no explanation of how it is derived or what the number means. The public also rarely has access to how accurate that model tends to be over time. Our access to the models themselves, along with a complete and comprehensive dataset, allowed us to compare win probabilities in an unprecedented way.

Weaknesses

While we had more access than prior studies, the major weakness of this research is that we only had access to two different win probability models. Both the PFR and random forest models are available for public use, and by contacting the PFR researchers we were given access to the expected points equations in order to run the model ourselves. The original NFL win probability model constructed by Brian Burke and examined by Lock and Nettleton (2014) is now owned by ESPN, and they do not provide details on the model. Therefore, we were unable to include it in our research. We don’t know whether it was more or less accurate then the PFR or random forest models, nor how it might have changed over time. In fact, we don’t even know how the win probability calculator available at the PFR website works. It includes overtime as an option, for example, so it’s unclear how the formulation of the model detailed on their own website has been adapted to accomplish this and whether it might have been modified in other ways. We offer full transparency in a domain that, for propietary reasons, tends to maintain tight control over the key details. We are not trying to develop and share as accurate a model as possible, and we are not withholding details to seek a competitive advantage against rivals in the business of sports or sports journalism.

While our comprehensive sample is certainly a strength of the study, the fact that it only dates back to 2001 limited our ability to study the effects of rule changes and differences in play style. Some major rule changes, such as the emphasis on defensive pass interference and quarterback safety, have been implemented during the time span in our sample. The way the game is played, however, has changed over longer time periods. There have been larger changes decade-to-decade than year-to-year. It would be interesting to work with a sample that goes back at least to the 1970s to better test win probability models as the offensive emphasis has changed from rushing to passing.

A third limitation of the present study is human error. The dataset we purchased had errors in the recording of time during the 2000 season. Because of these errors, we dropped the 2000 season altogether. There is the possibility for other errors in recording, though. There were possibilities for error when translating the expected points equations PFR sent to us into our usable R code. Despite extensive checking and debugging, some minor concerns could not be resolved (e.g., expected points for an offense become slightly better with worse field position on first down inside their own 10 yard line). Personal communication with the helpful individuals at PFR suggested that this issue involves either a lack of data due to the rare circumstance, or a mistake in the equations themselves. On the whole, we believe these equations are logically compelling and work quite well, but that doesn’t mean they are without flaw.

Future Directions

Though we hope to have improved the transparency and accuracy of win probability estimation, there are many areas for future research to expand on this foundation. First, it may worthwhile to further modify the PFR model. Perhaps there are better ways to discount time or to deal with the challenges of overtime. Our modifications seem to work well, but further research may yield something even better.

Second, in addition to the random forest and PFR models, there are other ways to estimate win probability. Future research that examines the validity of other models, such as Burke's model, now owned by ESPN, could be very informative. Burke's model is considered to be the original approach and it directly influenced the work of PFR and Lock and Nettleton (2014). Examining how well Burke's popular ESPN-endorsed model fares would be worthwhile, whether or not its details are made publicly available.

Third, whereas our efforts to improve existing models focused on the PFR model, it may be possible to re-specify the random forest model itself to achieve greater predictive accuracy. Lock and Nettleton (2014) tried using different parameterizations and including different variables, and we did not come up with new ideas to go beyond all that they had done. It remains possible that someone else might find a way to boost accuracy to the point where the increased complexity of a random forest model would be warranted.

Fourth, it would be interesting to explore the use of more special-purpose win probability models. For example, perhaps win probability could help to test the popular belief in momentum. Our research (Ruscio & Brady, 2020) found no evidence of between-game momentum in the NFL, but win probability models could help to test for within-game momentum. By including vs. excluding factors dealing with momentum shifts, one could test whether these help to predict outcomes. Another example would be a win-probability calculator designed for use exclusively on 4th down plays, to help decide whether to go for it or kick (and, in the latter case, whether to attempt a field goal or punt).

Finally, we intend to create a user-friendly tool for easy calculation of win probabilities using our modified PFR model. As noted earlier, this could be helpful, or at least entertaining, for NFL fans to play around with either during a game or afterward. Many game scenarios, real or hypothetical, can be constructed and we'd like to make it as easy as possible to estimate their corresponding win probabilities.

References

Alamar, B. (2010). Measuring risk in NFL playcalling. *Journal of Quantitative Analysis in Sports, 6*(2), 1-9. doi:10.2202/1559-0410.1235

Bennett, J. M., & Flueck, J. A. (2005). Player game percentage. In *Anthology of Statistics in Sports* (pp. 87-89). Society for Industrial and Applied Mathematics.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Burke, B. (2014, December 7). Win probability and win probability added explained. Retrieved from http://www.advancedfootballanalytics.com/index.php/home/stats/stats-explained/win-probability-and-wpa

Ganguly, S., & Frank, N. (2018, February). The problem with win probability. In *2018 MIT Sloan Sports Analytics Conference*.

Lindsey, G. R. (1961). The progress of the score during a baseball game. *Journal of the American Statistical Association*, *56*(295), 703-728.

Lock, D., & Nettleton, D. (2014). Using random forests to estimate win probability before each play of an NFL game. *Journal of Quantitative Analysis in Sports*, *10*(2), 197-205.

Moscowitz, T. J., & Wertheim, L. J. (2011). *Scorecasting: The hidden influences behind how sports are played and games are won*. New York: Crown.

The P-F-R Win Probability Model. (n.d.). Retrieved from https://www.pro-football-reference.com/about/win_prob.htm

Winston, W. L. (2012). *Mathletics: How gamblers, managers, and sports enthusiasts use mathematics in baseball, basketball, and football*. Princeton University Press.

Figure 1. Increase in accuracy with more training years provided to random forest models.

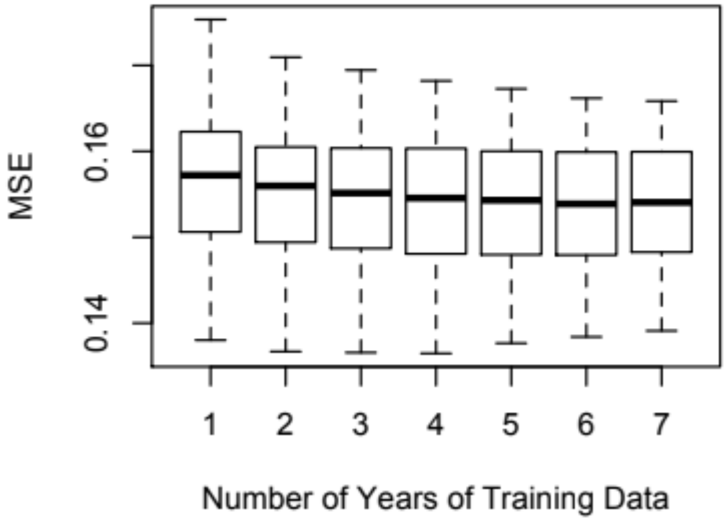


Figure 2. Calibration curves for random forest model (top) and PFR model (bottom). Model data are plotted as circles, with a theoretically perfect calibration curve shown as a dashed line.

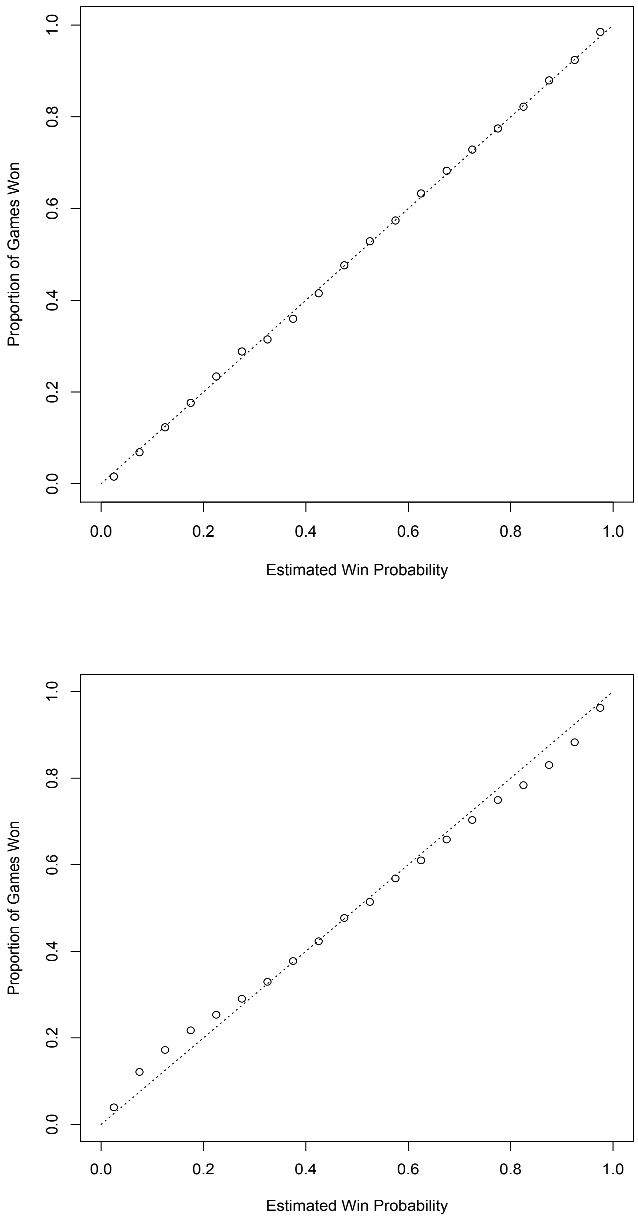


Figure 3. Difference in accuracy over the course of a game for the random forest and PFR models.

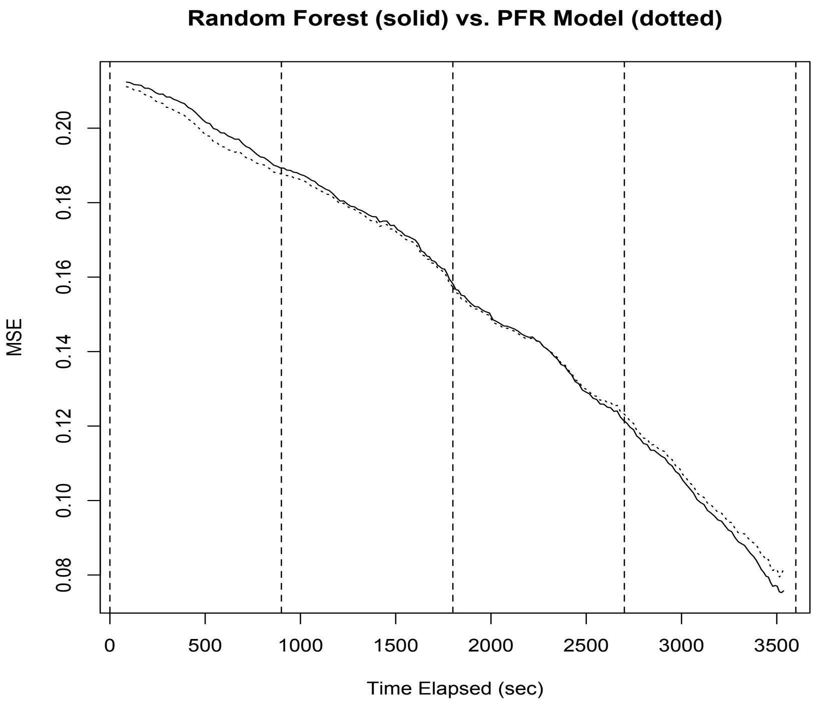


Figure 4. Time discounting function for the original PFR model (upper left) and modified PFR models using a time floor at 450 seconds (upper right), a curve stretched to reach the time floor value for 450 seconds (lower left), and the stretched curve with steeper discounting in the final two minutes of regulation (lower right).

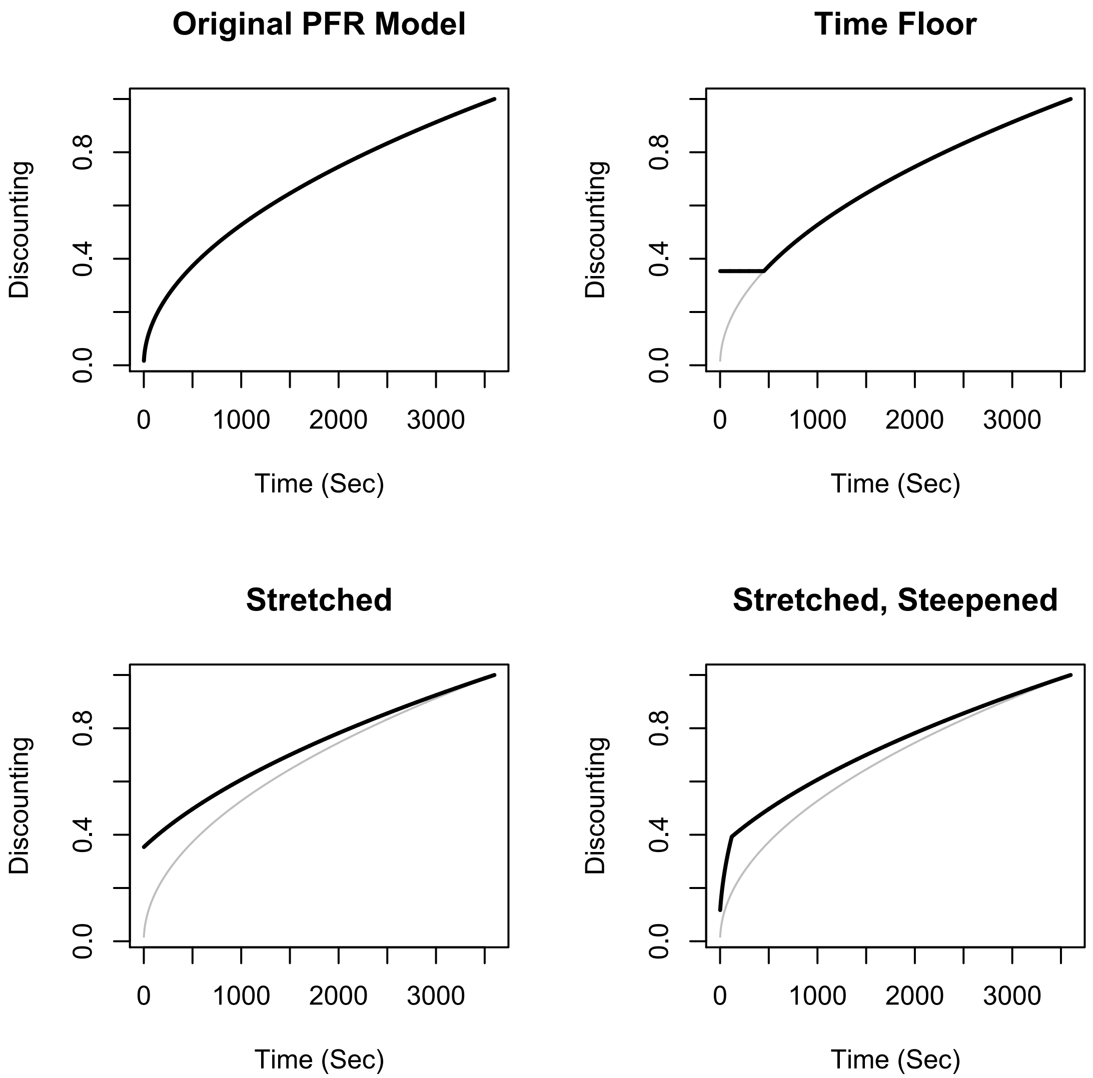
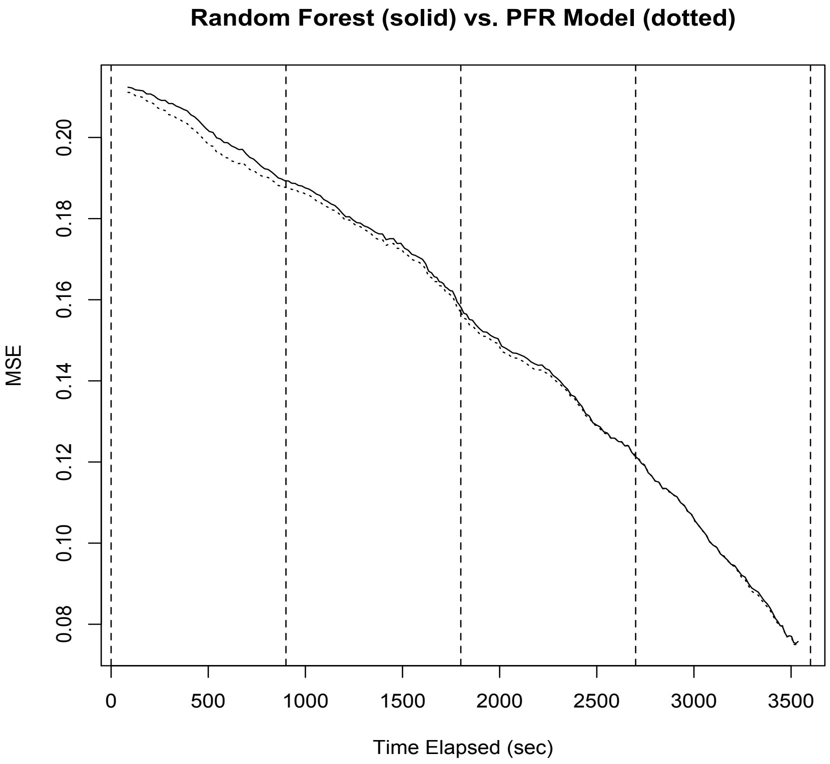


Figure 5. Difference in accuracy over the course of a game for the random forest and fully modified PFR models.


Random Forest (solid) vs. PFR Model (dotted)
MSE
Time Elapsed (sec)