Using Random Forests to Estimate Win Probability Before Each Play of an NFL Game

Dennis Lock[1]
Dan Nettleton[1]

[1]*Department of Statistics, Iowa State University*

January 23, 2014

Before any play of a National Football League (NFL) game, the probability that a given team will win depends on many situational variables (such as time remaining, yards to go for a first down, field position and current score) as well as the relative quality of the two teams as quantified by the Las Vegas point spread. We use a random forest method to combine pre-play variables to estimate Win Probability (WP) before any play of an NFL game. When a subset of NFL play-by-play data for the 12 seasons from 2001 to 2012 is used as a training dataset, our method provides WP estimates that resemble true win probability and accurately predict game outcomes, especially in the later stages of games. In addition to being intrinsically interesting in real time to observers of an NFL football game, our WP estimates can provide useful evaluations of plays and, in some cases, coaching decisions.

Author Notes: The authors wish to thank the reviewers, whose comments led to an improved manuscript, and Brian Burke, whose website *advancednflstats.com* served as partial inspiration for our work.

1 Introduction

The probability that a particular team will ultimately win an NFL game can be difficult to estimate at a specific moment. Undoubtedly fans, coaches, and players alike implicitly consider this probability as a game unfolds. We develop a statistical method for estimating this win probability (WP) prior to any play of an NFL game. As an example of what our methodology can produce, Figure 1 illustrates our WP estimates of a Baltimore Ravens victory in Super Bowl 47 prior to every play of the game.

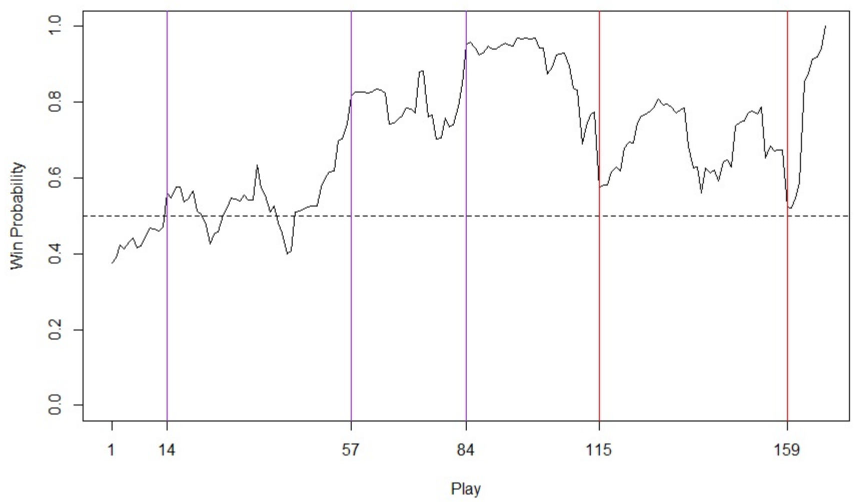

Figure 1: Estimated Baltimore Ravens win probability from every play of Superbowl 47. Highlighted plays with score and Ravens win probability: 1 = Start of game (0-0, $WP = 0.382$), 14 = Ravens score first (7-0, $WP = 0.552$), 57 = Ravens intercept the ball on the first play following their second touchdown (14-3, $WP = 0.743$), 84 = Ravens Open half with 109 Yard kickoff return (28-3, $WP = 0.935$), 115 = 49ers recover a fumble after back to back touchdowns (28-20, $WP = 0.572$), 159 = 49ers first and goal with 2 minutes remaining (28-23, $WP = 0.524$).

In addition to being of interest to fans as they watch a game in progress, our WP estimates could be used to evaluate specific plays and coaching decisions. For instance by comparing WP estimates in Section 7, we examine whether certain penalties should be accepted or declined and whether an offensive team should kick a field goal on fourth down or attempt to get the first down. While we discuss only a few specific examples, similar analyses can be used by coaches to strengthen decisions and enhance game strategy.

The idea of WP estimation for major sports is not new. Early uses of win probability were primarily in Major League Baseball but have existed since the beginning of the 1960's (Lindsey (1961)). Recent books on baseball analytics dedicate entire sections or chapters to the topic of win probability (Schwartz (2004), Tango, Lichtman, Dolphin, and Palmer (2006)). Analysts and fans of the other major sports have also begun to examine and use win probability more recently. For NBA and NHL examples, see Stern (1994) and Buttrey, Washburn, and Price (2011), respectively.

Motivation for this paper came partially from Brian Burke's NFL win probability metric found at *www.advancednflstats.com*. Burke constructs a play-by-play win probability using mostly empirical estimation. His win probability focuses on in-game variables score, time remaining, down, yards to go for a first down and field position. His general strategy is to partition the observations of a training dataset into bins based on values of his predictor variables score, time remaining, and field position. The proportion of training observations in a bin that correspond to a win for the team on offense provides an estimate of the win probability for the offensive team whenever the score, time remaining, and field position of a new situation are consistent with the bin. Adjustments to WP based on first down conversion probabilities are included to account for down and yards to go. Some extrapolation and smoothing are used to incorporate information from situations in other bins similar to the situation for which a prediction is desired.

We attempt to enhance Burke's approach in several ways. First, rather than subjectively binning the training observations, we let the data define a partitioning using a method that attempts to minimize prediction error. Second, we include the pre-game point spread to measure the quality of both teams competing. Thus, unlike Burke's approach, our method provides WP estimates that differ from 50% for each team at the beginning of a game. Third, our method permits the use of additional variables and provides a natural assessment of the importance of each variable. Finally, the approach we propose can be applied in a largely automatic and straightforward manner to other sports when sufficient training data are available.

The heart of our WP estimation methodology is the random forest (Breiman (2001a)). The random forest is a good candidate for our prediction function for many reasons. First, and perhaps most important is the well documented predictive ability of a random forest (see, for example, Breiman (2001b), Svetnik, Liaw, Tong, Culberson, Sheridan, and Feuston (2013) Diaz-Uriarte and de Andres (2006), Genuer, Poggi, and Tuleau (2008) Genuer, Poggi, and Tuleau-Malot (2010)). Second, a random forest can combine many predictor variables with unknown interactions in a non-linear data driven manner. Third, the random forest method provides natural and effective assessment of variable importance (Breiman (2001b)). Finally, the method runs on minimal assumptions, handles outliers well, and predicts based on empirical evidence (Breiman (2001a), Liaw and Wiener (2002), Cutler, Edwards, Beard, Cutler, Hess, Gibson, and Lawler (2007)).

Successful use of the random forest has begun to appear in sports analytics recently. Some examples include predicting major league success in minor league baseball players (Chandler and Stevens (2012)), predicting hall of fame voting in baseball (Freiman (2010), Mills and Salaga (2011)), and predicting game outcome in non-American football matches (Hucaljuk and Rakipovic (2011)). It should be noted each of these examples used a random forest of classification trees. Our approach differs somewhat because our WP estimates are generated by a random forest of regression trees.

In Section 2, we discuss the training data used to construct our WP estimates. In Section 3, we describe the random forest estimation method. In Section 4, we examine the performance of our estimator. Sections 5 through 7 evaluate the importance of variables, and their effects on win probability, changes in win probaiblity during the course of games, and using win probability estimates to analyze coaching decisions. The paper concludes with a discussion including alternative approaches, future considerations, limitations, and conclusions in Section 8.

2 Training Data

The analyses in this paper are based on all play-by-play data from NFL seasons 2001 through 2012 obtained from *ArmChairAnalysis.com*. Except where noted otherwise, data from the 2012 season were set aside as a test set, and only data from 2001 to 2011 were used as a training set. This training set consists of $n = 430,168$ plays from $2,928$ games with $p = 10$ predictors for each play extracted or constructed from the play-by-play data. A list and description of each predictor variable is included in Table 1. We use Y to denote the $n \times 1$ response vector and X to denote the $n \times p$ matrix of predictor values. Each row of the data matrix $[Y,X]$ corresponds to one pre-play situation observed with respect to the offensive team. The response is an indicator of victory so that the i^{th} element of Y is 1 if the team on offense before play i won the game and 0 otherwise.

Variable Name	Variable Description
Down	The current down (1st, 2nd, 3rd, or 4th)
Score	Difference in score between the two teams
Seconds	Number of seconds remaining in the game
AdjustedScore	$Score/\sqrt{Seconds+1}$
Spread	Las Vegas pre-game point spread
TIMO	Time outs remaining offense
TIMD	Time outs remaining defense
TOTp	Total points scored
Yardline	Yards from own goal line
YTG	Yards to go for a first down

Table 1: Description of predictor variables

3 Random Forest Method

Random forests generate predictions by combining predicted values from a set of trees. Each individual tree provides a prediction of the response as a function of predictor variable values. We use a forest of regression trees, where each individual regression tree is generated as follows.

1. Draw a bootstrap sample of observations from the training dataset and group all sampled observations in a single node N_0.
2. Randomly select m predictor variables from all p predictors.
3. For each x among the m selected predictors and for all cut points c, compute the sum of squared errors

$$\sum_{k=1}^{2} \sum_{i \in N_k} (y_i - \bar{y}_k)^2,$$

   where N_1 is the set of training observations with $x \leq c$, N_2 is the set of training observations with $x > c$, and $\bar{y}_k$ is the response mean for training observations in N_k $(k = 1, 2)$.
4. Choose x and c to minimize the sum of squared errors in step 3, and split the training observations into two subnodes accordingly.
5. Repeat steps 2 through 4 recursively at each resulting node until one of two conditions are met:
   (1) the number of observations in node k is less than a chosen tuning parameter *nodesize*, or
   (2) all the response values corresponding to observations in the node are identical.

The final nodes that result from this recursive partitioning process are referred to as terminals nodes. This series of splits can be presented graphically as a binary tree, where each split constructs the "branches" and the final "leaves" represent the terminal nodes. Once the tree is constructed from the training data, a predicted response for a future observation can be found by tracing the observation's path down the branches of the tree to a terminal node (based on the observation's predictor variable values) and computing the average of the training responses in that terminal node. The prediction of the forest is then obtained by averaging the predictions of all trees in the forest.

As discussed by Lin and Jeon (2006) and Xu, Nettleton, and Nordman (2014), random forests are similar to adaptive nearest-neighbors methods, which predict the response for a target observation by averaging the responses of the "nearest" training observations. Such methods are adaptive in the sense that the definition of "nearest" is based on a concept of distance in the predictor variable space that accounts for the relationship between the predictors and the response inferred from the training data. Predictor variables unrelated to the response are ignored while predictor variables strongly associated with the response play a major role when evaluating the distance between observations. In our application, the random forest win probability estimate for a given target play is a weighted average of game outcomes associated with past plays that are judged by the random forest algorithm to be similar to the target play. These similar training set plays are those that make up the terminal nodes of trees in the forest that contain the target play. The game outcomes for the training set plays most similar to the target play (i.e., those that occur most often in terminal nodes associated with the target play) get heavily weighted while outcomes for dissimilar plays (i.e., those seldom in a terminal node with the target play) receive little or no weight.

We construct our random forest using the function *randomForest* in the R package *randomForest*. The random forest has two tuning parameters, m the number of candidate predictors at each split and *nodesize* the maximum terminal node size. We chose both parameters using a cross-validation strategy described as follows. Play-by-play data from the 2011 season were set aside, and WP estimates for plays from the 2011 season were generated using random forests constructed from plays in 2001 through 2010 with various choices of m and *nodesize*. Based on the resulting misclassification rates, we chose $nodesize = 200$ (well above the R *randomForest* regression default of 5) and $m = 2$ (slightly below the default value of ($\lfloor p/3 \rfloor = 3$). The *randomForest* default of 500 regression trees were constructed with these two tuning parameter choices.

The decisions to use regression trees and to use the constructed variable *AdjustedScore* were also based on our cross-validation performance. For our data, the main difference between regression and classification trees is the predicted response in the terminal node of a regression tree is the proportion of response values equal to 1, while a classification tree reports a 1 if the proportion is greater than 0.50 and a 0 otherwise. The variable *AdjustedScore* was included to improve the performance of the method primarily in the later stages of games. Because we know *a priori* that a nonzero lead increases in value as the seconds remaining in a game decreases, we considered using

$$AdjustedScore(\gamma) = \frac{Score}{(Seconds+1)^{\gamma}}$$

for $\gamma = 0, 0.1, 0.5, 1, 1.5$ or 2. We ultimately selected $\gamma = 1/2$ because this value of γ minimized our cross-validation misclassification. Note that this cross-validation analysis favors using *AdjustedScore* and *Score* over *Score* alone because

$$AdjustedScore(\gamma) = Score \text{ for } \gamma = 0,$$

which was one value in our candidate set from which $\gamma = 1/2$ was selected.

4 Win Probability Prediction Accuracy

Measuring the accuracy of estimated win probabilities is a difficult task. For instance, when it appeared the 49ers were about to score late in Super Bowl 47 we estimated a 48% chance of victory. This may have been an accurate estimate of their win probability despite the fact that they lost the game 9 plays later. One basic way to measure accuracy is to calculate mean squared error (0.156) from all plays in our test set, where the example above contributes $(0-0.48)^2$ to the numerator of that mean squared error. Another option is to look at the mean squared error as the game progresses (Table 2). Mean squared error should decrease as the game progresses because we gain more information and move closer to the final response.

Quarter	1st	2nd	3rd	4th
MSE	0.201	0.177	0.143	0.107

Table 2: Test set MSE by quarter

Possibly a better way to measure accuracy is to bin the plays in the test set by estimated win probability and then calculate the proportion of wins in each bin. This proportion of wins is a representation of the unknown true win probability for the plays in a given bin. If a method performs well, we would expect estimated win probabilities that define bins to be similar to the actual proportion of wins within bins. For example, among plays with an estimated WP ≈ 0.75, aproximately 75%

should be associated with an offensive win. Figure 2 shows a plot of estimated win probability (binned in 5% increments) for plays in the test set and the proportion of offensive wins among plays in each bin. Correlation between proportion of wins and the WP at the center of each bin is extremely high ($r = 0.998$), and the random forest WP estimates are clearly well calibrated.

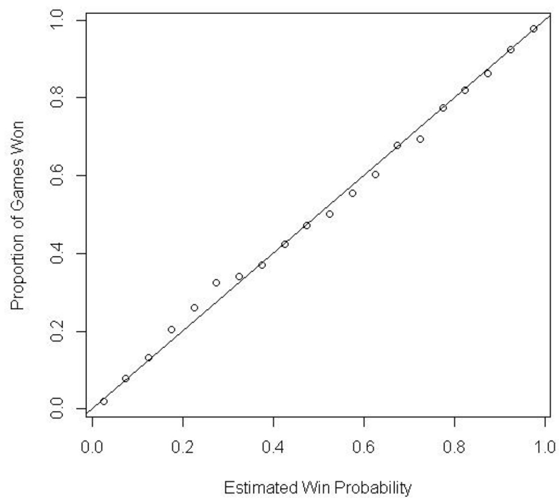

Figure 2: Binned estimated win probability and proportion of games won in each bin (line = a perfect fit).

5 Assessing Variable Importance and Relationships with Win Probability

In addition to the performance measures discussed in Section 4, it is interesting to examine how WP estimates change when one variable expected to have an effect on win probability is changed while holding the others constant. For example, Figure 3(a) shows how WP changes as the difference in score changes while holding all other variables constant. The other plots in Figure 3 show the effect of varying seconds (b), spread (c), down (d), yards to go (e), or yards from own goal line (f). Primarily of note is that each variable changes win probability in the direction we would expect, with *Score* having the greatest effect. That being said their are many other interesting features to note. For example, in Figure 3(b), we see that WP changes little over time until around the 4th quarter for each of the score differences examined. The black line in Figure 3(b) shows that having the ball in a tied game at your own 20 is advantageous (WP > 0.5) until just before halftime when it provides no advantage to either team (WP ≈ 0.5). When varying point spread in Figure 3(c) we see many plateaus, especially in the more extreme point spreads where the random forest is primarily grouping an interval of point spreads together as equivalent. Also no team is given a pre-game win probability greater than 80%, regardless of the point spread. Figure 3(f) shows that improving field position is noticeably more important with less time remaining, with a considerable increase around the opposing 40 yard line (entering field goal range).

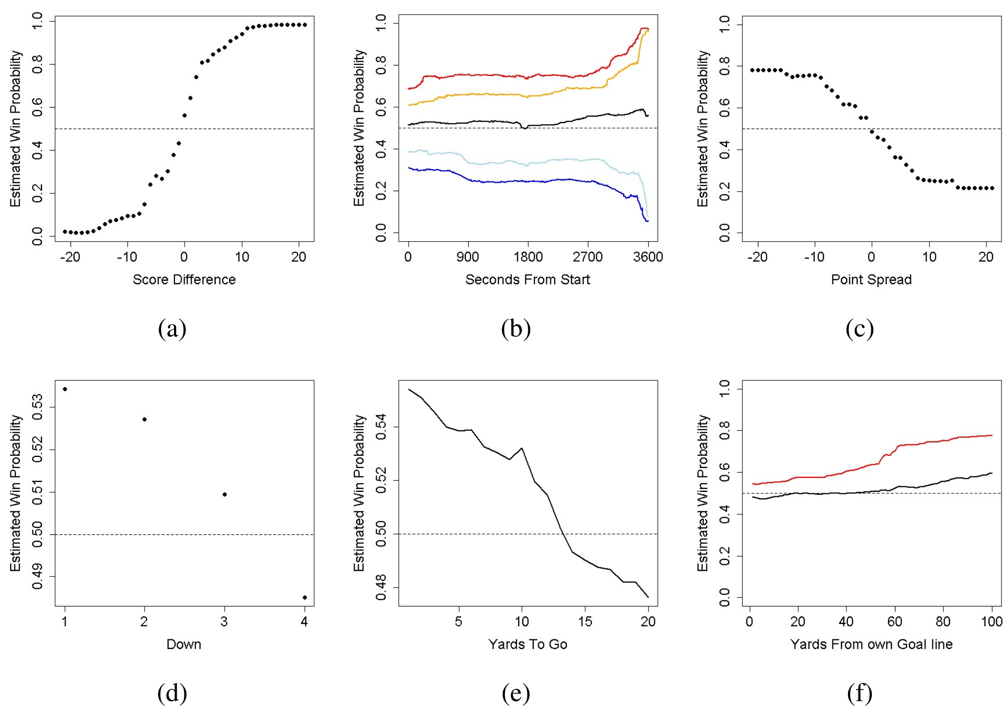

Figure 3: Changing one variable at a time with others held constant $Down = 1$, $YTG = 10$, $Yardline = 20$, $TOTp = 28$, $Seconds = 300$, $Score = 0$, $Spread = 0$, $TIMO = TIMD = 3$ unless otherwise specified. Variable changed in plot (a) *Score*, (b) *Seconds* (blue: $Score = -7$, Light Blue: $Score = -3$, Black: $Score = 0$, Orange: $Score = 3$, red: $Score = 7$), (c) *Spread* ($Seconds = 3600$), (d) *Down*, (e) *YTG* ($Down = 3$), and (f) *Yardline* (black: $Seconds = 1700$, red: $Seconds = 120$). Note the y-axis is focused on a narrower range in plots (d) and (e).

The importance of score difference is apparent graphically from the plots in Figure 3, but we can also numerically estimate variable importance. We chose to calculate importance for the kth variable as follows.

1. Randomly permute the values of predictor variable k within the test set and re-predict WP.
2. For each play i calculate the squared error after permuting the values of variable k minus the original squared error.
3. Repeat steps 1 and 2 many times (we chose 100 repetitions), and find the average increase in squared error for each play i.

This provides a play-wise variable importance for all plays in the test set. Overall variable importance can be found by averaging across all plays, and variable importance for specific types of plays can be found by averaging over plays of given type. Table 2 shows overall and quarter-specific measures of variable importance for each of our 9 variables (note that the variable *AdjustedScore* is just a function of *Seconds* and *Score* so was not permuted separately but recalculated for permutations of either *Seconds* or *Score*). Overall and in three of the four quarters, *Score* is the most important variable, however in the first quarter *Spread* is actually more important than *Score*.

Variable	Overall	Qtr 1	Qtr 2	Qtr 3	Qtr 4
Score	0.13653	0.04697	0.09773	0.15348	0.23539
Spread	0.02462	0.05361	0.02919	0.01436	0.00459
Seconds	0.00657	0.01105	0.00570	0.00341	0.00643
Yardline	0.00265	0.00276	0.00139	0.00208	0.00428
TOTp	0.00160	0.00195	0.00000	0.00152	0.00334
Down	0.00031	0.00038	0.00017	0.00018	0.00045
TIMO	0.00019	0.00040	0.00000	0.00005	0.00062
TIMD	0.00013	0.00023	0.00000	0.00017	0.00062
YTG	0.00009	0.00010	0.00002	0.00006	0.00018

Table 3: Variable importance (overall and by quarter)

One major advantage of calculating variable importance in this way is that we can examine how two variables interact by observing the importance of one variable at specific values of the other variable. For example, in Table 3 we can see that the difference in score becomes more important as the game progresses while the point spread becomes less important. Figure 4 shows a plot of the interaction between *Spread* and *Seconds*, looking at the importance of *Spread* at each second. Not surprisingly the importance of *Spread* is relatively high at the beginning of games (when not much other information is available) but diminishes to near irrelevance in the closing seconds.

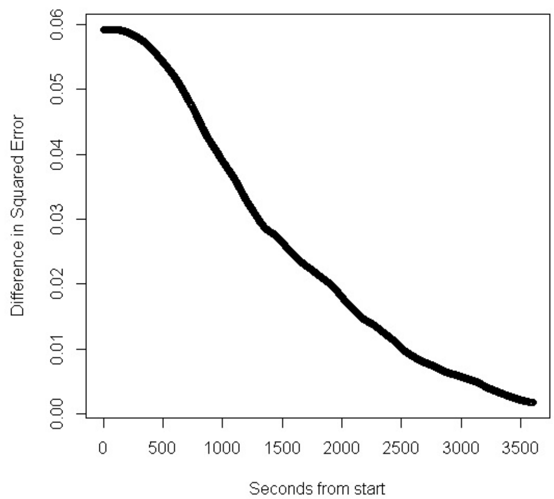

Figure 4: Variable importance of *Spread* by seconds from start (smoothed using a loess smoother).

6 Examining Changes in Win Probability

We can use change in win probability (ΔWP) to judge the most influential plays within a specific set of plays. For example, David Tyree's catch in Super Bowl 42 (ΔWP = 0.113), rated the greatest Super Bowl play in NFL history by Fox Sports, was actually not even the most influential play of that game. The touchdown pass to Plaxico Burress 4 plays later had a much greater increase in win probability (ΔWP = 0.389). If we were to choose the greatest Super Bowl play based on ΔWP, it would be James Harrison's 100 yard interception return for a touchdown just before halftime in Super Bowl 43 (ΔWP = 0.511). Similarly, we can judge the best play of the entire 2012 season to be Cecil Shorts' 39 yard touchdown reception to take a 1 point lead over the Vikings with 27 seconds remaining in the 4th quarter (ΔWP = 0.710).

Using the predicted win probability values from an entire game, we can plot how win probability changed as the game progressed. The plots of two of the more exciting Super Bowls from the last 12 seasons are presented below. Figure 5 shows Super Bowl 44 between the Indianapolis Colts and New Orleans Saints, and Figure 6 shows Super Bowl 42 between the undefeated New England Patriots and New York Giants.

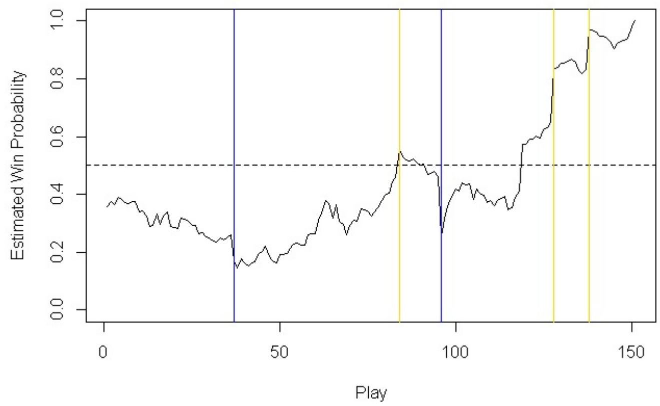

Figure 5: Estimated win probability by play for Superbowl 44, blue vertical lines represent Indianapolis touchdowns and gold vertical lines represent New Orleans touchdowns.

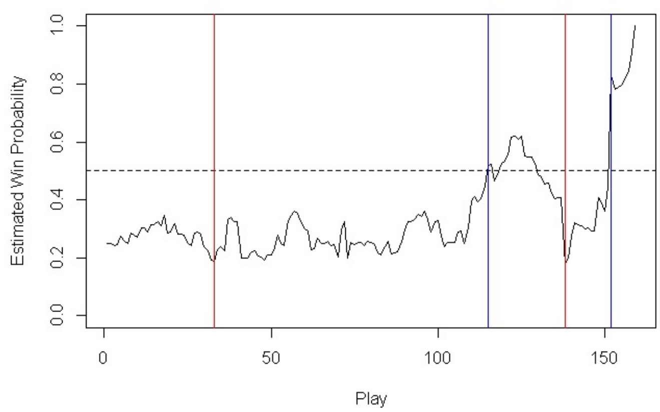

Figure 6: Estimated win probability by play for Superbowl 42, blue vertical lines represent New York touchdowns and red vertical lines represent New England touchdowns.

7 Win Probability Analysis of Coaching Decisions

WP and ΔWP can be used to evaluate some coaching decisions. Suppose, for example, that the offense is flagged for a holding penalty while throwing an incomplete pass on third and 10 from their own 20 yard line. In a game between evenly matched teams (*Spread*= 0) with the score tied 7 to 7 at the beginning of the second quarter, should the defense decline the penalty and force a fourth and 10 situation, or accept the penalty to put the offense at third and 20 from their own 10 yard line? Random forest WP calculations can provide guidance and favor accepting the penalty even though declining would almost surely guarantee a punt and a change of possession. The WP for the offense at fourth and 10 from the 20 is estimated to be 46% compared to 43% at third and 20 from the 10.

It is possible, of course, that the coach on the field will base his decision on additional information not available to the random forest. If the offense is facing a strong headwind, for example, accepting the penalty may be even more favorable than indicated by the random forest WP estimates. On the other hand, early-game injuries in the defensive backfield might make declining the penalty better than asking an inexperienced defense to make another stop. However, regardless of additional information, the random forest does provide useful baseline information that indicates that, in the past, teams facing third and 20 from their own 10 have lost more often than those in situations like fourth and 10 from the 20.

A very similar accept vs. decline WP analysis can be used in other situations, such as when a punting team commits a penalty that, if accepted, would almost surely result in a re-punt. The receiving team's WP at first and 10 from their current field position can be compared to their WP if the penalty were assessed and the punting team were to face fourth down again closer to their own goal line. Because the best choice may depend on the capabilities of the special teams involved, the WP analysis may not be able to give a definitive answer. However, the WP calculations can serve as a useful starting point for making an informed decision.

As another example, suppose the offense trails 14 to 10 and faces a fourth and 3 from the opponent's 10 yard line. Should they take the almost certain 27 yard field goal to cut their deficit to 14–13 or try for a first down? When the offense's WP at fourth and 3 is greater than their WP would be following a successful field goal, going for the first down is likely the better choice. Figure 7 shows that kicking the field goal is a good decision in the first half while going for the first down is better with about 10 or fewer minutes to go in the game. The two options are approximately equivalent throughout the third quarter.

As noted for the first examples of this section, the WP analysis provides baseline guidance constructed from past performance that could be overridden when special circumstances or specific strengths, weaknesses, and tendencies of the competing teams are taken into account. It is important to remember that our training data are observational rather than experimental. Teams facing similar situations in the past have not been randomly assigned to courses of action by an experimenter. Thus, we cannot be certain that decisions to go for the first down rather than kicking the field goal with, say, 8 minutes to go in a game *caused* the higher success rate experienced by teams that chose to go for the first down. However, in general, we

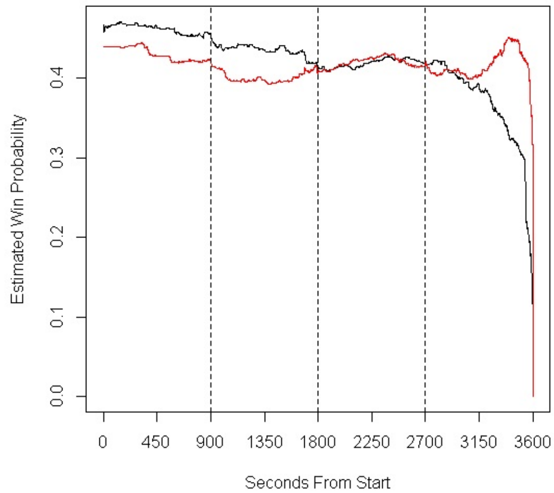

Figure 7: Estimated win probability before (red) and after (black) a successful 27 yard field goal on fourth down and 3 when the offense is trailing by a score of 14 to 10 prior to the kick (Other variables: $TIMO = TIMD = 3$, $Spread = 0$). Dotted lines indicate the changing quarters.

believe coaches should avoid attempting plays that, even if successful, will result in a decrease in their team's WP.

8 Discussion

Win probabilities estimated through our random forest method are similar to those calculated by Brian Burke. Figure 8 shows estimated WP for Superbowl 45 ($Spread = 2.5$) using both a random forest and Brian Burke's estimation. In general, because we include the point spread variable, our methodology provides better WP estimates near the beginning of games, especially in games with a clear favorite.

One major advantage of our approach is that it is fairly simple and could easily be replicated in other sports provided sufficient data is available. Due to its nature and performance, random forest methodology offers a unified approach to predict in-game win probability across many sports. In other work in progress, we have used random forests to estimate WP in the NHL and NBA, with success similar to that reported here for the NFL.

In addition to the WP calculator that served as motivation for our NFL work, Burke has also developed WP calculation methods for the NBA and NHL. While

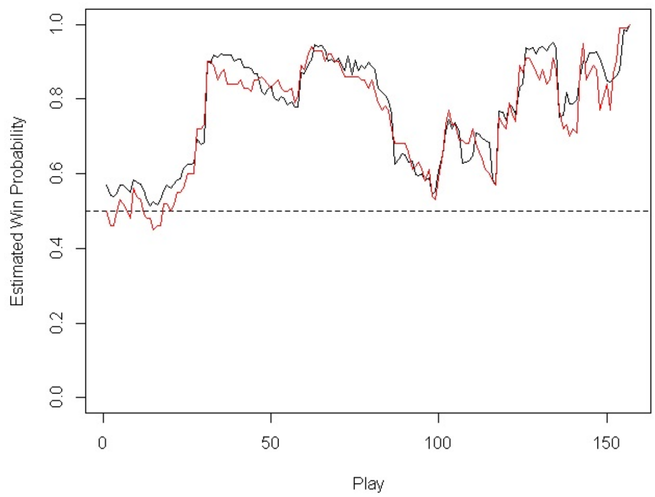

Figure 8: Estimated win probability by play for Superbowl 45 with both random forest estimation (Black) and Brian Burke's estimation (Red).

both a random forest and Burke's method predict win probability through historical values, there also exist methods which predict through simulation (some good examples include *accuscore.com* and *predictionmachine.com*). These methods differ in that WP is estimated utilizing win proportions from simulated game outcomes, rather than win proportions from historical situations. Because the details of the methods underlying the simulations are proprietary, it is not possible for us to evaluate the performance of these methods relative to our random forest approach.

Using either a random forest method or Brian Burke's method, specific WP values are estimated without regard to previous plays; only the current situation rather than the events leading up to that situation is considered. This may not be detrimental as there are many papers that discount the effect of momentum in the NFL (Johnson, Stimpson, and Clark (2012), Fry and Shukairy (2012)). A related issue is that each game has approximately 150 sequential observations all associated with the same response, and the random forest treats these observations as independent.

We attempted multiple adjustments to our random forest approach, some of which were meant to account for potentially detrimental effects of dependence in the data. To account for momentum, we attempted including the win probability estimates at the end of each quarter as predictors for later stages in the game. To account for multiple observations within each game, we constructed 3,600 separate random forests for each second, where the set of plays in each forest contains one observation per game, chosen as the closest observation to that second. This guaranteed that each forest consisted of 2,928 independent observations (under the assumption that plays from different games are independent). Future predictions were found by generating a prediction from the forest that corresponds with the current second. Methods with separate random forests by down or quarter were also attempted, such that, for instance, 4th down plays were estimated only from other 4th down plays. With each of these adjustments, we either saw no improvement in performance with added complexity (Momentum adjustment, separate forests by quarter) or a decrease in performance with added complexity (separate forests by second, separate forests by down), so no adjustment was included. To be thorough we constructed a few simple models unrelated to random forests, such as logistic regression on win/loss and linear regression on final score difference, each with large decreases in performance.

We also considered information other than *Spread* to account for team quality. The most basic was just adding variables such as current record, points per game, yards allowed per game, etc. to the variables in the forest. The most advanced was combining team quality variables using either a logistic regression or pre-game random forest to come up with a pre-game win probability variable to include as another predictor in a subsequent within-game forest. None of these alternative approaches showed an improvement over a model with only the point spread.

Throughout this paper, WP estimates for a play in a given game i were generated from training data that did not include plays from game i. For example, as previously discussed, data from 2001 through 2011 were used as the training set for generating predictions in 2012. Similarly, WP estimates for Super Bowls prior to 2012 were generated by using plays from 2001 through 2011 excluding the 11 Super Bowls. In future application, our random forest could be retrained after each week of NFL games to include play-by-play data from 2001 up through the most recently completed NFL games.

In conclusion, we have developed a method of estimating win probability that performs well and is simple to replicate. Regardless of how pre-play win probabilities are estimated, the uses of these values are numerous and could improve the way we look at the game.

References

Breiman, L. (2001a): "Random forests," *Machine Learning*, 45, 5–32.

Breiman, L. (2001b): "Statistical modeling: the two cultures," *Statistical Science*, 16:3, 199–231.

Buttrey, S. E., A. R. Washburn, and W. L. Price (2011): "Estimating NHL scoring rates," *Journal of Quantitative Analysis in Sports*, 7:3.
Chandler, G. and G. Stevens (2012): "An exploratory study of minor league baseball statistics," *Journal of Quantitative Analysis in Sports*, 8:4.
Cutler, D. R., T. C. Edwards, Jr., K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler (2007): "Random forests for classification in ecology," *Ecology*, 88:11, 2783–2792.
Diaz-Uriarte, R. and S. A. de Andres (2006): "Gene selection and classification of microarray data using random forest," *Bioinformatics*, 7:3.
Freiman, M. H. (2010): "Using random forests and simulated annealing to predict probabilities of election to the baseball hall of fame," *Journal of Quantitative Analysis in Sports*, 6:2.
Fry, M. J. and F. A. Shukairy (2012): "Searching for momentum in the NFL," *Journal of Quantitative Analysis in Sports*, 8:1.
Genuer, R., J. Poggi, and C. Tuleau (2008): "Random forests: some methodological insights," *arXiv*.
Genuer, R., J. Poggi, and C. Tuleau-Malot (2010): "Variable selection using random forests," *Pattern Recognition Letters*, 31:14, 2225–2236.
Hucaljuk, J. and A. Rakipovic (2011): "Predicting football scores using machine learning techniques," *MIPRO, 2011 Proceedings of the 34th International Convention*, 1623–1627.
Johnson, A. W., A. J. Stimpson, and T. K. Clark (2012): "Turning the tide: big plays and psychological momentum in the NFL," *MIT Sloan Sports Analytics Conference 2012*.
Liaw, A. and M. Wiener (2002): "Classification and regression by randomForest," *R News*, 2:3, 2225–2236.
Lin, Y. and Y. Jeon (2006): "Random forests and adaptive nearest neighbors," *Journal of the American Statistical Association*, 101, 578–590.
Lindsey, G. R. (1961): "The progress of the score during a baseball game," *Journal of the American Statistical Association*, 56, 703–728.
Mills, B. M. and S. Salaga (2011): "Using tree ensembles to analyze National Baseball Hall of Fame voting patterns: an application to discrimination in BBWAA voting," *Journal of Quantitative Analysis in Sports*, 7:4.
Schwartz, A. (2004): *The numbers game*, New York: Thomas Dunne Books.
Stern, H. (1994): "A brownian motion model for the progress of sports scores," *Journal of the American Statistical Association*, 89, 1128–1134.
Svetnik, V., A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston (2013): "Random forest: a classification and regression tool for compound classification and QSAR modeling," *Journal of Chemical Information and Modeling*, 53:8.

Tango, T., M. Lichtman, A. Dolphin, and P. Palmer (2006): *The Book: Playing the Percentages in Baseball*, New York: TMA Press.

Xu, R., D. Nettleton, and D. J. Nordman (2014): "Predictor augmentation in random forests," *Statistics and Its Interface*, Accepted.