Using a Special-Purpose Win Probability Model to Improve Fourth-Down Decisions in the National Football League

John Ruscio and Matthew J. Guiliano

The College of New Jersey

Draft Date: 1/22/2021

This working paper is provided to help understand our shiny apps titled "NFL Fourth Down Calculator" and "NFL Fourth Down Graph". Comments, questions, or suggestions are warmly welcomed. Please contact ruscio@tcnj.edu.

Abstract

The perception that NFL coaches are too conservative on fourth downs, choosing to kick rather than go for it, has been gaining momentum. The increasing attention being paid to to sports analytics, combined with salient examples of success following bold play-calling, appears to be shifting the conventional wisdom toward going for it more often. The present study examines the question of when to do so by evaluating the relative merits of three different models that offer advice. Two widely known models are based only on field position and yards to go for a first down, and following their advice was only weakly associated with increased win probability estimates or season-long win totals in analyses of 72,938 fourth-down plays over the past 19 NFL seasons. A third model uses machine learning to take into account a wide array of pertinent information, and following its advice was much more strongly associated with wins. This machine learning approach could be used by broadcasters or viewers in real time, by obsevers critiquing decisions made in pivotal moments, or by NFL teams to test intuitions about fourth-down strategy or mine data in myriad ways to develop more effective strategies as they prepare for upcoming games.

1. Introduction

Throughout a drive in the National Football League (NFL), the stakes increase each time the team on offense fails to either score or earn a first down. To assess the average impact of plays on each down (first, second, third, and fourth), data from ArmChairAnalysis.com for all plays in regular season and postseason games spanning the 2001 to 2019 seasons were examined using the win probability method provided by Pro Football Reference (n.d.). Impact was measured using the mean absolute change in win probability following a given play. For example, advancing the ball downfield, defensive penalties, earning a first down, or scoring will typically increase a team's probability of winning, whereas plays that lose yardage, offensive penalties, punts, or turnovers will typically decrease a team's probability of winning. Because win probability can increase or decrease, the absolute value of the change from one play to the next was calculated so that the mean would reflect the typical impact across all possible outcomes after a given down. Sequential plays were only included in our calculations if they took place within the same half of the same game. The mean absolute change in win probability was only .020 for 296,295 first-down plays and .023 for 223,952 second-down plays, but it increased to .047 for 144,783 third-down plays and peaked at .052 for 77,954 fourth-down plays.

This preliminary analysis shows that just over 10% of all plays take place on fourth down and these are among the most consequential moments in a game. Depending on the circumstances, it may be difficult for a coach to decide what to do in these situations. There are two basic options. They can go for it by calling a rush or pass play to try to earn enough yards for a first down or, if in a "goal-to-go" situation, a touchdown. Alternatively, they can kick it by calling for either a punt or a field goal try. Fourth-down decisions are higher-impact than those

on earlier downs is that there is so much at stake. Consider what can happen for each course of action.

If the coach goes for it, the team might score a touchdown, bringing the drive to a successful conclusion, or at least earn a first down that keeps the drive alive. On the other hand, the team might come up short, or turn the ball over through a fumble or interception, in which case the opposing team takes possession where the play ends. This is not only a failure to score any points on that drive, but also a failure to push the opponent back so that their drive will start with worse field position.

If the coach chooses to punt, this foregoes an opportunity to score any points on that drive but usually succeeds in pushing the opponent back into worse field position. Less common outcomes include having the punt returned for significant yardage, perhaps even a touchdown, which is worse than having gone for it and come up short, or recovering a fumble after the punt is fielded, thereby gaining significant yardage and a first down.

If the coach chooses to attempt a field goal, there is a chance to score 3 points, which is better than nothing but less than the value of a drive that ends with a touchdown. If the field goal misses, the opponent takes possession at the spot where the kick was attempted, usually 7 or 8 yards behind the line of scrimmage, which means no points are scored and the opponent gets even better field position than if the team tried going for it but came up short. A less common outcome is to have a field goal attempt blocked and the ball returned for significant yardage, perhaps even a touchdown.

The present study is designed to explore what coaches tend to do in fourth-down situations, evaluate advice to be more aggressive by going for it more often in accordance with fairly simple guidelines, and propose a more nuanced approach to making these decisions that

takes into account a wider range of pertinent factors. Section 2 reviews theory and research on fourth-down decision making. Section 3 describes the data source for the present study and Section 4 describes the models to be evaluated. Section 5 presents the findings and Section 6 discusses their implications.

2. Background

Many observers have noted that NFL coaches have, historically, tended to make conservative decisions on fourth downs, meaning that they are much more likely to kick than go for it (Moscowitz & Wertheim, 2011). It may be that this was a smart approach in earlier eras, when NFL games featured much less scoring. In a tight, low-scoring contest, the battle for field position increases the value of punting, and a field goal try that yields 3 points could be decisive. Going for it may be a needlessly risky choice. Another potential reason for conservative decision making is that coaches are risk averse and therefore prefer to kick because it is considered the safe approach (Easterbrook, 2015; Urschel & Zhuang, 2011). If the coach kicks in a key situation and the team loses, fans and sports media will not blame the coach for being reckless. Instead, the players will be blamed for failing to execute throughout the game.

In a study of the fourth-down decisions of college football coaches, Owens and Roach (2018) found that coaches with greater job security (e.g., those who have won a Lombardi Trophy) were more aggressive than those with weaker job security. The former may be in a better position to weather the storm if they go for it in a key situation and the team loses anyway, whereas coaches on the "hot seat" may be more likely to stick to the conventional wisdom, the conservative choice to kick, because they fear the repercussions of even being perceived as taking unnecessary chances. For many decades, the conventional wisdom has been that kicking

is safer, going for it is risky. Behaving accordingly appears to be especially important for those coaches who have less job security.

The conventional wisdom itself, however, may be changing. As early as the 1970s, data analysts began to suggest that NFL coaches should be more aggressive on fourth downs. Carter and Machol (1978) were the first to examine this systematically. They adopted an expected points framework and concluded that coaches should go for it more often. Romer (2006) examined more—and more recent—data using an expected points framework and reached essentially the same conclusion. His results even provided simple guidelines for when coaches should go for it, depending only on field position and yards to go for a first down. Burke and Quealy (2013) extended Romer's analysis to a wider range of game situations (e.g., examining actual fourth-down plays throughout a game rather than only third-down plays from the first quarter) and once more reached similar conclusions. They, too, suggested guidelines for when coaches should go for it that depend only on field position and yards to go, and their user-friendly interface is widely known as the *New York Times* Fourth-Down Bot (NYT Bot). Causey et al. (2015) updated this approach with more recent data. Finally, Yam and Lopez (2019) examined matched pairs of plays on which coaches chose to go for it or kick. They used an extensive list of pertinent variables to match plays into pairs (e.g., field position, yards to go, score, time remaining in the game) and then examined the change in win probability resulting from each play. For kicks, the distribution was unimodal, centered around a mean change in win probability of -.01. Going for it, in contrast, yielded a bimodal distribution with a mean of +.01 and greater variance. This suggests that coaches may be well advised to be more aggressive, and it also demonstrates the sense in which this is risky: The variance in outcomes is greater. Relatively speaking, kicking affords greater predictability.

Yam and Lopez (2019) report that there is no evidence that NFL coaches are following the advice provided by Romer (2006) or the NYT Bot. It is true, as they found, that the percentage of the time that coaches go for it on fourth downs fluctuated apparently at random through 2017. However, there was a noticeable uptick in the two seasons that followed. As shown in Figure 1, both 2018 and 2019 saw levels higher than in any season from 2001 to 2017. Though this may be an anomaly, there are reasons to believe that coaches are in fact becoming more aggressive in their fourth-down decision making. For example, as head coach of the New England Patriots, Bill Belichick has been both aggressive on fourth downs and enormously successful, leading the team to nine Super Bowl appearances and six wins. This success, and its link to bold play-calling, has not been lost on anyone. Perhaps even more compelling is what Doug Pederson, head coach of the Philadelphia Eagles, accomplished. In the 2016 and 2017 seasons, the Eagles chose to go for it more than any other team in the NFL, and Pederson led the Eagles to their first-ever Super Bowl win. The Eagles, who entered that Super Bowl as 4.5 point underdogs, executed a bold game plan to upset the Patriots, who made surprisingly conservative play calls. Pederson chose to go for it on two key fourth-down plays and the Eagles converted both times. One of those plays, the "Philly Special" pass to quarterback Nick Foles that yielded a touchdown, is now commemorated with a statue outside the Eagles' home stadium. It would be difficult to imagine a more dramatic illustration of why coaches should go for it on fourth down. Pederson's unwavering boldness may have begun to change the conventional wisdom, affording other NFL coaches the safety they need to begin to heed the advice of data analysts. Though it may be too soon to know this for certain, it would be surprising if coaches returned to more conservative fourth-down play calls in upcoming seasons.

It is against this backdrop that the present study seeks to evaluate the evolving conventional wisdom. Would coaches be well-advised to follow the advice of Romer (2006) or the NYT Bot? These guidelines have been proposed to encourage coaches to be more aggressive, yet they only take into account field position and yards to go. To what extent would the purported advantages of being more aggressive be offset by a narrow focus on two variables, to the relative neglect of many other situational factors widely understood to be relevant to making smart fourth-down decisions, such as the score and time remaining in the game? The links between following these models' advice and expected changes in win probability, as well as in observed win totals, are examined. In addition, a more complex model that uses machine learning to take into account a richer array of situation-specific factors is developed and tested. The present study differs from those using expected points in that the random forest model is designed to predict win probability. By comparing estimated win probabilities for going for it versus kicking, this model allows one to test how important this decision can be in influencing outcomes as well as to investigate the conditions that favor each course of action. Analyses demonstrate how the model can be used to test intuitions and uncover new insights.

3. Data Source

This study was performed using data tables from ArmChairAnalysis.com that included every play from all 5,324 regular season and postseason NFL games spanning the 2000 through 2019 seasons. There were 83,371 fourth-down plays before we cleaned the data for analysis. A total of 4,181 plays from the 2000 season were dropped because of mistakes in coding the time variable (e.g., time remaining on the game clock frequently remained identical for each play in an entire offensive possession), as were 200 plays from games that ended in a tie (as they could not be used in the development of a win probability model) and 2,252 plays in games for which

weather data were missing (and therefore could not be included in our modeling). This yielded a sample including 9,468 rush or pass plays (labeled "go for it") and 63,500 punts or field goal attempts (labeled "kick"). Finally, 3,770 entries in the database on which the offense neither went for it nor kicked (e.g., a penalty or time out was called) were removed; these were generally followed by an entry for an actual fourth-down play. This left a final sample of n = 72,968 fourth-down plays that took place in 4,911 games from 2001 through 2019. Coaches chose to go for it on 13.0% of these plays and kick on the other 87.0%.

The variables used in this study were obtained by merging information from two data tables. The play-level data table provided points scored by the team on offense, points scored by the team on defense, yards to go for a first down, field position, quarter, minutes on the game clock, and seconds on the game clock. The game-level data table provided wind speed, temperature, over/under, and point spread. For games played in a domed stadium, wind speed was set to 0 and temperature was set to 72. Following Lock and Nettleton (2014), several new variables were calculated: time (the number of seconds remaining in the game; 0 for plays in overtime), score (points for offense minus points for defense), adjusted score (score divided by the square root of time, with 1 second added to time to avoid division by 0 for overtime plays), total points (points for offense plus points for defense), and win (whether the team on offense ultimately won the game). See Table 1 for a summary of all variables.

4. Models

Three models were included in this study. The first two models use only two variables—field position and yards to go for a first down—to recommend whether to go for it or kick. The third model takes into account a much wider range of factors thought to be relevant to this decision.

4.1. Romer's (2006) Model

The first model was developed by Romer (2006). This model was operationalized by converting the smooth curve in Romer's Figure 4 into a set of discrete values for all possible combinations of field position and yards to go. Everything on or beneath the curve was coded as a recommendation to go for it, and everything above the curve was coded a recommendation to kick. The results are shown in our Figure 2, top graph. For the present sample of fourth-down plays, Romer's model recommended going for it on 40.1% of all plays.

4.2. Burke and Quealy's (2013) Model

The second model was developed by Burke and Quealy (2013). This is known more commonly as the NYT Bot, the interface the authors created for use in real time. The advice offered by this model (as of August 6, 2020) was retrieved online; see Figure 2, second graph. For six particular combinations of field positions and yards to go, the model is indifferent between going for it and kicking. To be as generous as possible, these instances were coded as kicks because it yielded the most model-friendly results in subsequent analyses. For the present sample of fourth-down plays, the NYT Bot recommended going for it 41.1% of the time. There were only 285 plays (0.4% of all plays) where the model was indifferent, and had these been coded as go for it rather than as kick, the NYT Bot would have recommended going for it 41.5% of the time.

4.3. Random Forest Model

The third model was a random forest win probability model similar to the one developed by Lock and Nettleton (2014). Random forest models use machine learning to identify relationships between predictor and criterion variables in large samples with many variables (Breiman, 2001). The algorithm can detect complex patterns, including nonlinear predictor-

criterion relationships and interactions among predictors. The random forest algorithm begins by taking a bootstrap sample that matches the original sample size; cases are sampled at random, with replacement. A tree is constructed by splitting this sample into two subsamples at each of a series of nodes. At each new node, a random selection of the available variables is examined to identify which one can be used to split the cases in a way that maximizes the difference in scores on the criterion variable for cases above vs. below a threshold value; all possible threshold values are examined for each variable being considered. Each split results in two new nodes nested within their parent node. For example, at the top of the tree, two variables might be selected at random, such as time and score. All possible thresholds for these variables are examined to determine which one would split the sample into subsamples that differ as much as possible in the criterion variable, in this case winning vs. losing the game. Perhaps the first node is split into subsamples consisting of plays when the offense is winning vs. tied or losing. This would create two new nodes, and the splitting process would be repeated, independently, for each of them by randomly selecting a new set of variables to examine for the next split. This splitting process proceeds until each node can be split no further without reducing its size below a specified minimum, yielding a series of terminal nodes. The larger the sample, the more terminal nodes will be created. The goal is to obtain terminal nodes that are as homogeneous as possible (i.e., some terminal nodes will contain plays that nearly always led to wins, other terminal nodes will contain plays that nearly always led to losses).

Once a tree is complete, the entire process restarts with a new bootstrap sample to generate the next tree. This is repeated a specified number of times to yield the desired number of trees in the random forest model. To make predictions using this model, the algorithm proceeds case by case, tree by tree. For a particular case of data, the algorithm begins by working

its way from the top of one tree to a terminal node at the bottom by following the splitting rules at each node along the way. The prediction for this case, for this tree, is the mean value on the criterion variable for all cases in the same terminal node. In the present context, that would mean averaging the win variable (coded as win = 1, loss = 0) to obtain an estimated win probability. This is repeated for each tree in the forest, each of which provides a new estimate of win probability. These values are averaged to yield a single estimate for the entire random forest. This is done for each case, in turn.

Whereas Lock and Nettleton (2014) developed a general-purpose random forest win probability model, applicable to all plays, the goal of the present study was to develop a special-purpose win probability model for fourth-down plays. Seven variables used by Lock and Nettleton (down, score, time, adjusted score, point spread, total points, field position, and yards to go) were used along with two weather variables (wind speed and temperature) because they can affect kicks and therefore might be expected to influence fourth-down decision making.

The critical decision variable (go for it vs. kick) was handled in two ways that ultimately yielded highly similar results. The first method was to include decision as a variable in the random forest model. Two separate win probabilities could then be estimated for a given play by providing values for the other nine variables along with either a decision to go for it (which yields one estimated win probability) or a decision to kick (which yields a second estimated win probability). The second method was to split the sample into two subsamples, one for plays on which the decision was to go for it and another for plays on which the decision was to kick. Separate random forest models were developed for each subsample using the other nine variables. Two separate win probabilities could then be estimated for a given play by providing values for the other nine variables to both random forests, each of which yields an estimated win

probability for an alternative decision (go for it vs. kick). Because the end results for the two methods were extremely similar, only those for the first method are reported because it is simpler to generate and work with a single random forest.

The randomForest R package (Liaw & Wiener, 2018) was used to generate the random forest, following Lock and Nettleton (2014) by setting mtry = 2 (which allows the model to randomly select two variables at each node to evaluate for making the best split into two new nodes), nodesize = 200 (which ensures that terminal nodes contain at least n = 200 plays), and ntree = 500 (which creates a forest containing 500 trees). Because this approach is prone to overfitting, the full sample of fourth-down plays was randomly split into 10 subsamples. A random forest model was generated using data from nine subsamples and then used to make predictions for the one remaining subsample. This was repeated 10 times so that predictions were obtained for cases in each subsample, and all of these were "out of bag" predictions. For simplicity, this paper will refer to the random forest model as though it was a single model, with the understanding that its predictions were obtained using a series of 10 such models.

As noted earlier, the random forest model was used to estimate win probabilities separately for each decision type (go for it vs. kick). Using the win probability estimates corresponding to the coaches' actual decisions for each play, the random forest model fit the data well, yielding MSE = .145. This is comparable to the MSE of .156 obtained by Lock and Nettleton (2014) using a similar set of variables in a similar sample. Another way to contextualize this level of accuracy is to compare it to the variance of the criterion variable, which for the approximately evenly-split binary win variable is .250. Thus, the MSE can be expressed as a model-fit r^2 = (.250 - .145) / .250 = .420. Finally, the calibration curve shown in

Figure 3 shows the excellent fit between predicted win probability and actual game outcomes; for the data points in this graph, $r = .9995$.

To use the random forest model to offer advice, the estimated win probabilities for each play were compared and choice associated with the higher value was coded as recommended. For the present sample of fourth-down plays, the random forest model recommended going for it 30.9% of the time.

5. Results

5.1. Comparing Models' Recommendations with Coaches' Decisions

Coaches chose to go for it (GFI) only 13.0% of the time, and that occurred mostly when they were at the edge of field goal range, or closer, and there were few yards to go for a first down. However, as shown in Figure 2 (bottom graph), there was wide variability in other circumstances, with coaches sometimes choosing to go for it. In this and many subsequent graphs, data points code the percentage of GFI decisions using size (larger data points for higher GFI%), shape (different symbols for each quartile of GFI%), and color (darker data points for higher GFI%). Also note that data points have been smoothed by aggregating results for adjacent data points (e.g., the point plotted for *yfog* = 65 and *ytg* = 3 includes results for all nine combinations of $34 \leq yfog \leq 36$ and $2 \leq ytg \leq 4$; see Table 1 for a list of variable names and descriptions).

The first two models, which are based on field position and yards to go, were much more aggressive than coaches. The Romer model recommended GFI 40.1% of the time and the NYT Bot recommended GFI 41.1% of the time. This is more than three times the rate at which coaches chose to GFI. The recommendations of the random forest model were less aggressive than the other two models, GFI = 30.9%, but still more than twice as high as the rate for coaches.

Beyond overall aggressiveness, the pattern of model recommendations was not as simple as one might expect. Table 2 shows the crosstabulations of model recommendations and coaches' decisions. Even though the Romer model and the NYT Bot were much more aggressive than coaches, they actually recommended kicking it nearly 30% of the time when coaches chose to GFI. This is surprising because it is a less aggressive choice being recommended by models that are known for being more aggressive. What does make these models more aggressive on the whole, however, is that they recommended GFI more than 35% of the time on the far more numerous occasions when coaches chose to kick. The random forest model, on the other hand, disagreed even more often with coaches' decisions to GFI (recommending kicks in more than 60% of those instances) but agreed more often with coaches' decisions to kick (recommending GFI in less than 30% of those instances). In total, the random forest model was in slightly higher agreement with coaches' decisions (66.1%) than the Romer model (65.5%) or the NYT Bot (64.8%).

5.2. Maximizing Win Probability and Wins

For plays when the models' recommendations differed from coaches' decisions, the estimated win probabilities were compared to see what impact those choices might have had (see Table 2). This is hypothetical, of course, but affords some insight into whether following models' advice is warranted. For both the Romer model and the NYT Bot, decisions to kick when coaches chose to GFI would have increased win probability. On the other hand, for both models, decisions to GFI when coaches kicked would have reduced win probability. When aggregating across all fourth-down plays on which these two models' recommendations and coaches' decisions differed, the net effect was a very small reduction in win probability, or a net effect extremely close to 0.

It would be circular to use the win probabilities generated using the random forest model to evaluate that model's recommendations. Because the decisions recommended by the random forest model were selected as those that were associated with higher estimate win probability, it is naturally the case that following these recommendations when they disagreed with coaches' decisions would yield an increase in estimated win probability. However, one can ask a related question that yields greater insight: Do coaches who behave *as if* they're following each type of model tend to win or lose more games?

The percentage of coaches' decisions that agreed with each model's recommendations was calculated separately for each team, during each season of play, for all 607 team-seasons in the data. This level of agreement was plotted against the number of wins for each team-season, and the results are shown in Figure 4. On each scatterplot, a dark line shows the simple linear regression and light lines connect scores at the quartiles along the variable on the *x* axis to the expected number of wins using the regression line.

The regression line for each of the three models slopes upward, indicating the decisions that agreed more closely with each model's recommendations were associated with winning more games. There are dramatic differences in the strength of the association between agreement and wins, however. For the Romer model and the NYT Bot, $r = .17$, and there is only a modest difference in win totals between scores at the upper and lower quartiles of agreement ($IQR \approx 1$). For the random forest model, on the other hand, $r = .80$, and there is a more substantial difference in win totals between scores at the upper and lower quartiles of agreement ($IQR \approx 3$).

In contrast to a comparison of models based on estimated win probabilities, which is biased in favor of the random forest model that was used to generate those estimates, there is nothing biased about a comparison based on wins. This is simply a look backward, using actual

data rather than estimates, to see how strongly a tendency to behave in accordance with each model's advice is associated with winning games.

5.3. Contextual Factors

The fact that adherence to the advice of all three models was associated with winning more games suggest that there is support for being more aggressive on fourth down. To provide some context for the magnitude of gains one might expect with more aggressive fourth-down decision making, we also plotted the percentage of GFI decisions against wins. This relationship is shown in the final graph of Figure 4, and it reveals that teams that chose to GFI more often lost more games (r = -.46, $IQR \approx 2$.). This suggests that indiscriminately being more aggressive may be unwise, that each model is associated with more wins by virtue of knowing when to GFI and when to kick. The challenge, of course, is figuring out what game conditions favor each choice.

The Romer model and the NYT Bot are fairly crude tools, as they only pay attention to two variables (field position and yards to go) and neglect many factors that most observers would consider to be obviously relevant (e.g., score, time remaining). The random forest model, however, can accommodate far more information. This part of the paper closes by demonstrating how the model results might be mined to uncover strategically important patterns. The third graph in Figure 2 summarizes the random forest model's recommendations for all fourth-down plays. By comparing this type of graph across systematically chosen subsets of plays, however, one can determine which game situations influence the random forest model's recommendations. This could be done in a confirmatory mode, to test intuitive understandings of fourth-down decision making, or in an exploratory mode, to search for trends.

To begin this illustrative series of analyses, consider what happens when you break down the plays by quarter. Figure 5 shows that there is a pretty substantial difference in the model's

recommendations as the game progresses, and at least some of this may be counterintuitive. On the whole, the model recommends being more aggressive in the first half of the game than later, with the most conservative advice of all in the fourth quarter.

The model's advice in the fourth quarter, however, is far from uniform. Figure 6 shows that it depends very strongly on whether the team on offense is trailing or leading. Notice that the difference between those two situations is not so much in the frequency that the model recommends GFI (24.5% when trailing, 23.0% when leading), but rather in the field positions where it recommends GFI. When trailing, the model is much more likely to recommend going for it when more than 30 yards from the end zone, whereas the reverse is true when leading. For teams leading by a touchdown or more, the model very seldom suggests GFI (8.7%), and in those cases it is mostly when on the edge of field goal range.

As a final example, consider what happens when you break down plays not by time or score, but by the relative strength of the two teams. Even a crude estimate of this variable, namely the pre-game Vegas point spread, reveals extremely large differences in model recommendations. Figure 7 plots findings for teams expected to win big (favored by 7+ points; GFI = 13.7%), win by a little (favored by <7 points; GFI = 21.2%), lose by a little (GFI = 34.0%), or lose big (GFI = 67.4%). This enormous gradient exists before even refining the samples to take into account current score, time remaining, or other important factors.

6. Discussion

To those who have paid attention to decades of data analyses suggesting that coaches should be more aggressive in their fourth-down decision making, it appears that change may have begun. Time will tell whether the past two seasons are just the beginning of a bolder era, or whether coaches return to more conservative play-calling. Given their understandable concerns

about job security, one should expect NFL coaches to do what they believe their audience expects them to do. The fact that many observers are paying greater attention to data analytics, as well as the recent and highly visible success of bold decision-making in the NFL, the audience may expect to see teams go for it on fourth downs more often than in the past.

How coaches will adapt, specifically how they decide when to go for it, is not obvious. The advice to "be more aggressive" is too vague to be very helpful. The advice offered by Romer's (2006) model and the NYT Bot is based on two variables that are clearly important—field position and yards to go—and these models usually arrive at the same advice. In the present sample of fourth-down plays, they agreed with one another 91.5% of the time. Their ease of use is both a strength and perhaps their most serious limitation. It would not be difficult for a coach to consult either of these models, or perhaps a variation on this theme constructed via in-house analytic work, in real time during a game. Doing so, however, ignores all of the other factors that are widely believed to be relevant. The present findings suggest that it would be foolish for a coach to strictly follow the guidelines of a model constructed using only field position and yards to go. The evidence shows no reason to expect this to increase win probability relative to the decisions coaches would make otherwise. Similarly, there was only a weak link between the extent to which coaches' decisions agreed with these models and their teams' season-long win totals. In addition to being an arguably poor guide for coaching decisions, it would be unfortunate for observers to expect coaches to follow these models' advice or to judge them harshly for failing to do so.

Instead, a more successful approach might be grounded in greater respect for the knowledge that the highly experienced coaches in the NFL bring to the table. Coaches recognize that there are many relevant factors to consider (e.g., score, time remaining). A more nuanced

model that takes into account this wider range of pertinent information could be used to inform strategic decision making in preparation for games and, after the fact, to inform critiques of in-game decisions. The random forest model developed and tested in this study shows that there is considerable potential for discovering useful patterns in the data that can be exploited to make better decisions. Agreement between coaches' decisions and the recommendations of the random forest model was very strongly associated with win totals.

Unfortunately, NFL sideline rules prohibit the use of technology that would allow coaches to consult a model as complex as a random forest during a game. Such a model could be developed in advance and it would be very easy to estimate win probabilities for a specific game situation. It would take mere seconds to enter the relevant data (e.g., field position, yards to go, score, time remaining in the game) into a function that runs these data through an existing random forest to produce the estimated win probabilities associated with going for it and kicking. However, this process is computing intensive, hence infeasible for real-time use by NFL teams. Tools such as this could be used in real time by broadcasters to provide viewers an understanding of what sports analytics suggest a coach should do. Particularly when the analytics yield reasults contrary to the conventional wisdom, this could go a long way toward changing outdated or poorly informed expectations for what coaches should do. Similarly, after the fact, observers interested in critiquing coaching decisions made in key situations could consult a random forest model that takes into account relevant factors rather than judging coaches against comparatively simplistic models that only consider field position and yards to go for a first down.

For NFL teams, perhaps the best they can do under the present sideline technology rules is to use a random forest model in preparation for upcoming games, such as while reviewing film on one's own performance or that of scheduled opponents. By producing and studying graphs of

the kind presented here, analysts can evaluate the conventional wisdom and allow new ideas to emerge through the targeted exploration of a rich set of data. The particular graphs displayed in this paper are intended to be illustrative rather than exhaustive. What they demonstrate is that constructing a special-purpose random forest win probability model is a helpful first step toward mining the data for a wealth of information. The potential utility of a machine learning approach was demonstrated for the case of fourth-down decision making in the NFL, but it could be applied to other important decisions in NFL games as well as other professional sports.

References

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

Burke, B., & Quealy, K., 2013, How coaches and the NYT 4th down bot compare, http://www.nytimes.com/newsgraphics/2013/11/28/fourth-downs/post.html, accessed August 7, 2020.

Carter, V., & Machol, R.E., 1978, Optimal strategies on fourth down, *Management Science*, *24*, 1758-1762.

Causey, T., Katz, J., & Quealy, K., 2015, A better 4th down bot: Giving analysis before the play, https://www.nytimes.com/2015/10/02/upshot/a-better-4th-down-bot-giving-analysis-before-the-play.html, accessed August 7, 2020.

Liaw, A., & Wiener, M., 2018, Breiman and Cutler's random forests for classification and regression, R package version 4.6-14, https://CRAN.R-project.org/package=randomForest.

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*, 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.

Owens, M. F., & Roach, M. A., 2018, Decision-making on the hot seat and the short list: Evidence from college football fourth down decisions, *Journal of Economic Behavior and Organization*, 148, 301–314.

Pro Football Reference, n.d., The P-F-R win probability model, https://www.pro-football-reference.com/about/win_prob.htm, accessed August 7, 2020.

Romer, D., 2006, Do firms maximize? Evidence from professional football, *Journal of Political Economy*, *114*, 340-365.

Urschel, J.D., & Zhuang, J., 2011, Are NFL coaches risk and loss averse? Evidence from their use of kickoff strategies, *Journal of Quantitative Analysis in Sports*, *7*.

Yam, D. R., & Lopez, M. J., 2019, What was lost? A causal estimate of fourth down behavior in the National Football League, *Journal of Sports Analytics*, *5*, 153-167.

Table 1. Summary of variables in the random forest model.

Variable	Description
Play-Level Data	
ptso	Points scored by team on offense
ptsd	Points scored by team on defense
ytg	Yards to go for a first down
yfog	Field position, coded as yards from own goal
qtr	Quarter, with overtime coded as 5 (or even 6, when needed)
min	Minutes on the game clock
sec	Seconds on the game clock
Game-Level Data	
wspd	Wind speed (set to 0 for games play in domed stadium)
temp	Temperature (set to 72 for games play in domed stadium)
ou	Over/under
sprv	Point spread, coded for visiting team
Calculated	
time	Seconds remaining in the game: $((4 - qtr) \times 15 + min) \times 60 + sec$; 0 for plays in overtime
score	Points for offense minus points for defense: $ptso - ptsd$)
adjscore	Adjusted score: *score* / sqrt(*time* + 1)
totpts	Total points: $ptso + ptsd$
win	Whether team on offense ultimately won the game (win = 1, loss = 0)
type	Decision (go for it = 1, kick = 2)

Table 2. Comparisons of decisions made by coaches with those recommended by models.

	Total	Coaches' Decisions Go for it	Kick	Δ Win Probability (Model – Coach)
Romer Model				
Go for it	40.1%	9.3%	30.8%	-.005
Kick		3.7%	56.2%	.015
		Agreement = 65.5%		Mean = -.002
New York Times 4th Down Bot				
Go for it	41.1%	9.4%	31.7%	-.004
Kick		3.5%	55.3%	.016
		Agreement = 64.8%		Mean = -.002
Random Forest				
Go for it	30.9%	5.0%	25.9%	.016
Kick		8.0%	61.1%	.017
		Agreement = 66.1%		Mean = .016

Figure captions

Figure 1. Percentage of fourth-down plays on which NFL head coaches chose to go for it during the 2001 to 2019 seasons. Results for the final two seasons are plotted using open circles to accentuate their difference from the filled circles for prior seasons.

Figure 2. Graphs show recommendations for the Romer model, the NYT Bot, and the random forest model, followed by coaches' decisions. Because the Romer model and the NYT Bot only consider field position and yards to go, each data point corresponds to a decision. In contrast, the random forest model and coaches' decisions consider many additional factors, so each data point represents the percentage of decisions to go for it. Results have been smoothed by aggregating data for ±1 yard of field position and ± yard to go for a first down. Sparse data points ($n < 5$) are not plotted.

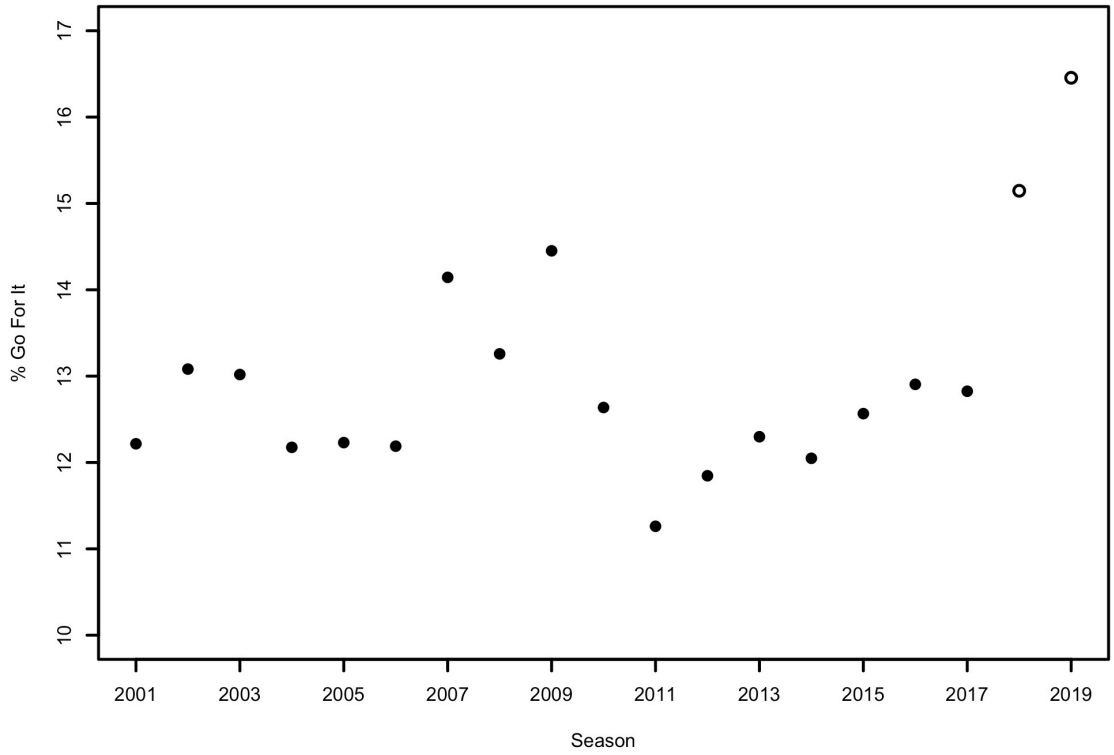
Figure 3. Calibration curve for the random forest model. The diagonal reference line represents perfect calibration.

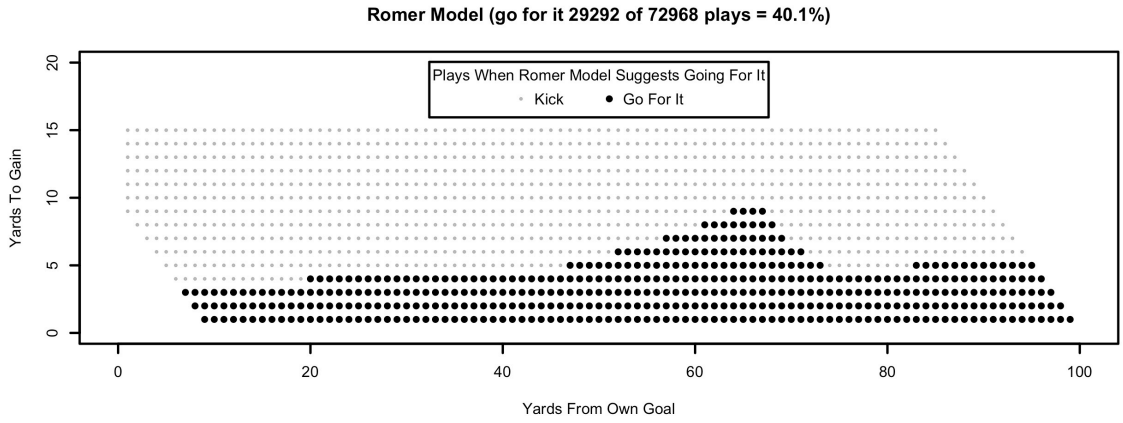
Figure 4. Graphs show associations between the number of wins for each of $n = 607$ team-seasons and the percent of all fourth-down plays on which coaches' decisions agreed with the advice of Romer's model, coaches' decisions agreed with the advice of the NYT Bot, coaches' decisions agreed with the advice of the random forest model, and coaches decided to go for it. Each graph includes a simple linear regression line (dark lines) and index lines showing how many wins correspond to each quartile along the *x* axis (light lines).

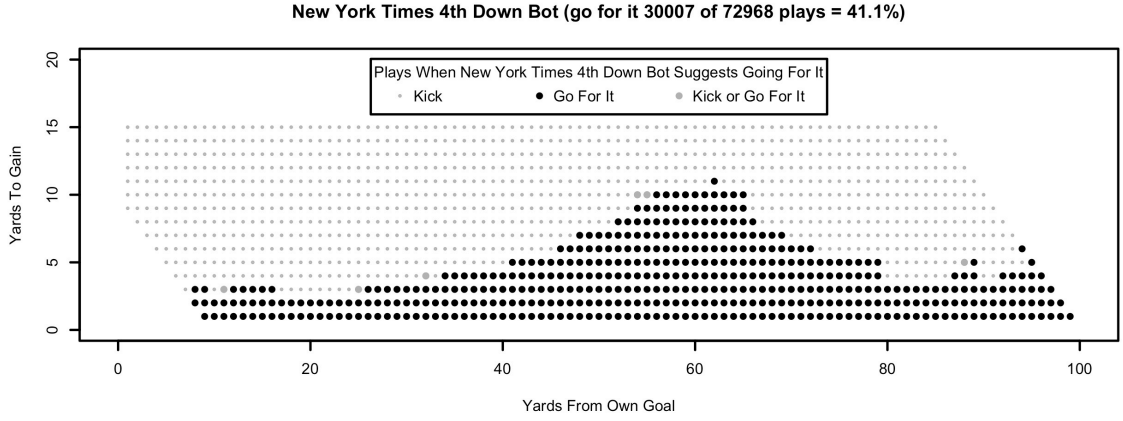
Figure 5. Graphs show findings for the random forest model broken down by plays in each quarter of the game. Results have been smoothed by aggregating data for ±1 yard of field position and ± yard to go for a first down. Sparse data points ($n < 5$) are not plotted.

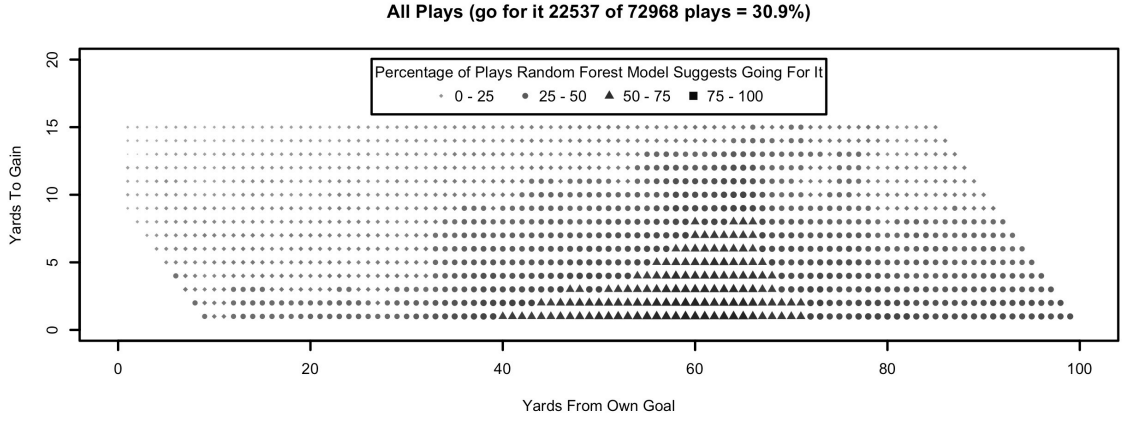
Figure 6. Graphs show findings for the random forest model for all plays in the fourth quarter followed by subsets of fourth-quarter plays in which the team on offense was trailing, leading, and leading by at least 7 points. Results have been smoothed by aggregating data for ±1 yard of field position and ± yard to go for a first down. Sparse data points ($n < 5$) are not plotted.

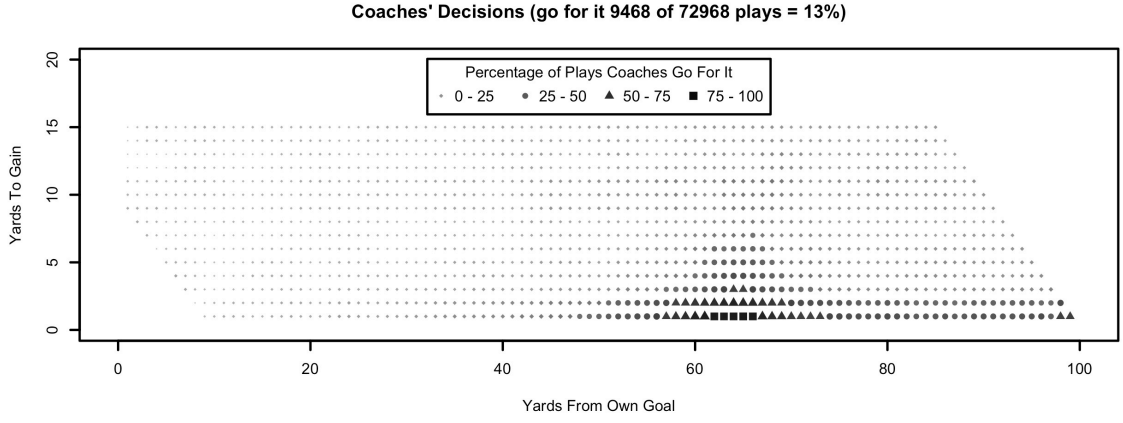
Figure 7. Graphs show findings for the random forest model broken down by the pre-game point spread (expected to win by at least 7 points, expected to win by less than 7 points, expected to lose by less than 7 points, or expected to lose by at least 7 points) to represent the strength of the team on offense. Results have been smoothed by aggregating data for ±1 yard of field position and ± yard to go for a first down. Sparse data points ($n < 5$) are not plotted.

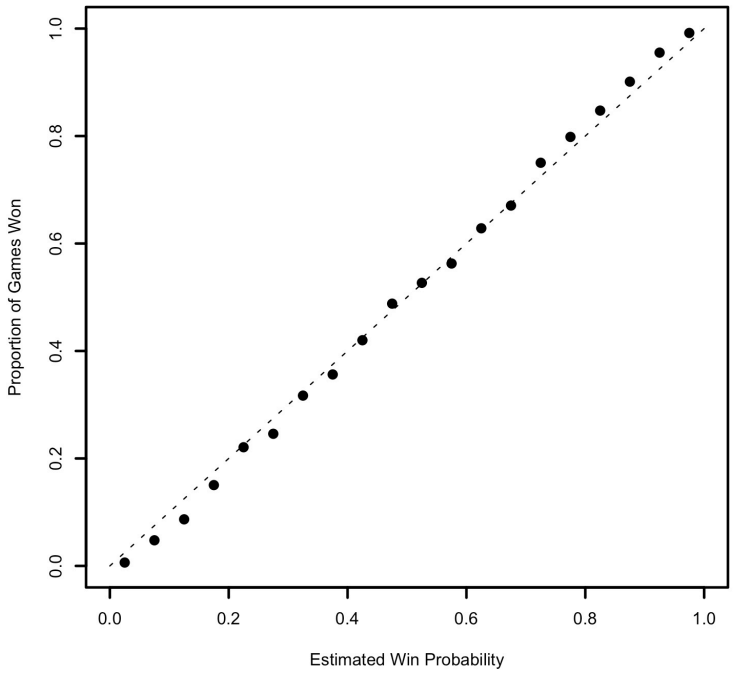

% Go For It
10
11
12
13
14
15
16
17
2001
2003
2005
2007
2009
2011
2013
2015
2017
2019
Season

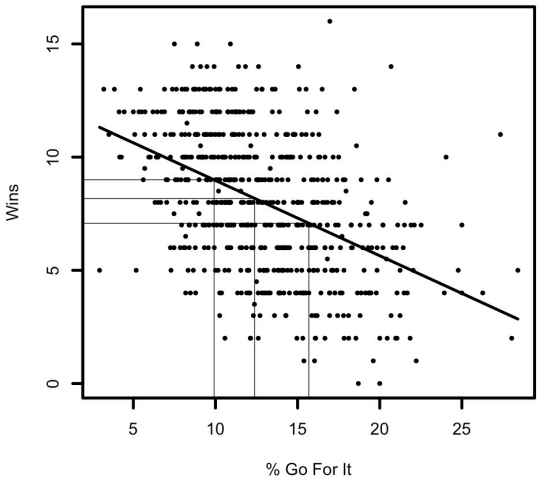

Romer Model (go for it 29292 of 72968 plays = 40.1%)
Plays When Romer Model Suggests Going For It
Kick
Go For It
Yards To Gain
Yards From Own Goal

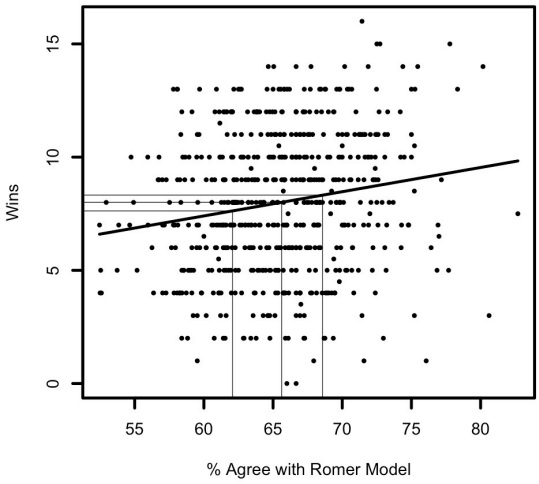

New York Times 4th Down Bot (go for it 30007 of 72968 plays = 41.1%)
Plays When New York Times 4th Down Bot Suggests Going For It
Kick
Go For It
Kick or Go For It
Yards To Gain
Yards From Own Goal

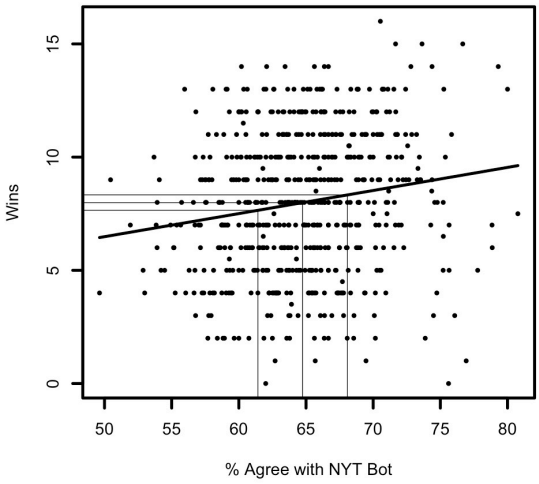

All Plays (go for it 22537 of 72968 plays = 30.9%)
Percentage of Plays Random Forest Model Suggests Going For It
0 - 25
25 - 50
50 - 75
75 - 100
Yards To Gain
Yards From Own Goal

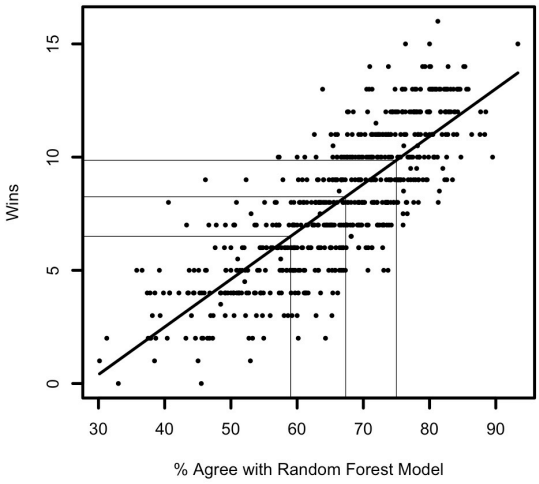

Coaches' Decisions (go for it 9468 of 72968 plays = 13%)
Percentage of Plays Coaches Go For It
0 - 25
25 - 50
50 - 75
75 - 100
Yards To Gain
Yards From Own Goal

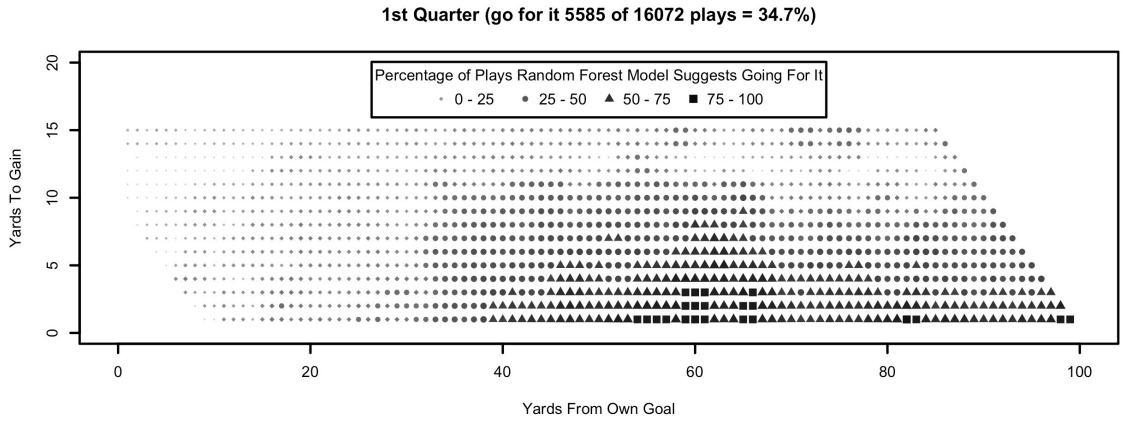

Proportion of Games Won
0.0
0.2
0.4
0.6
0.8
1.0
Estimated Win Probability
0.0
0.2
0.4
0.6
0.8
1.0

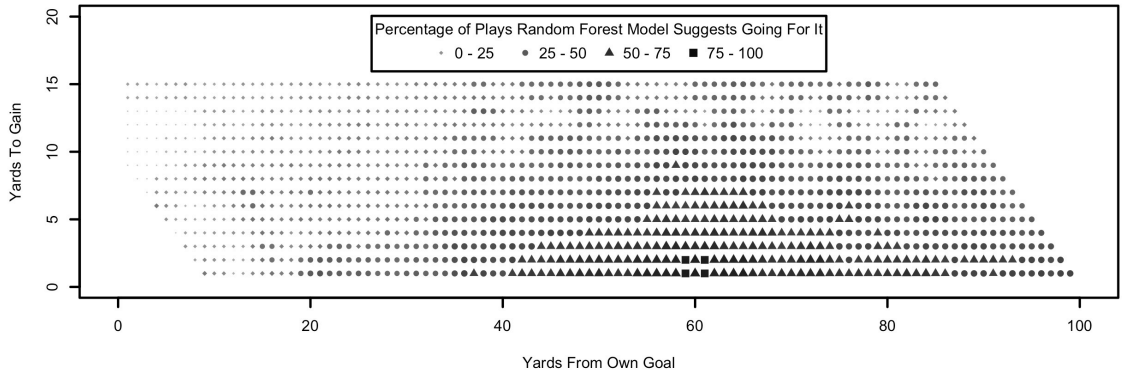

Wins
0
5
10
15
5
10
15
20
25
% Go For It

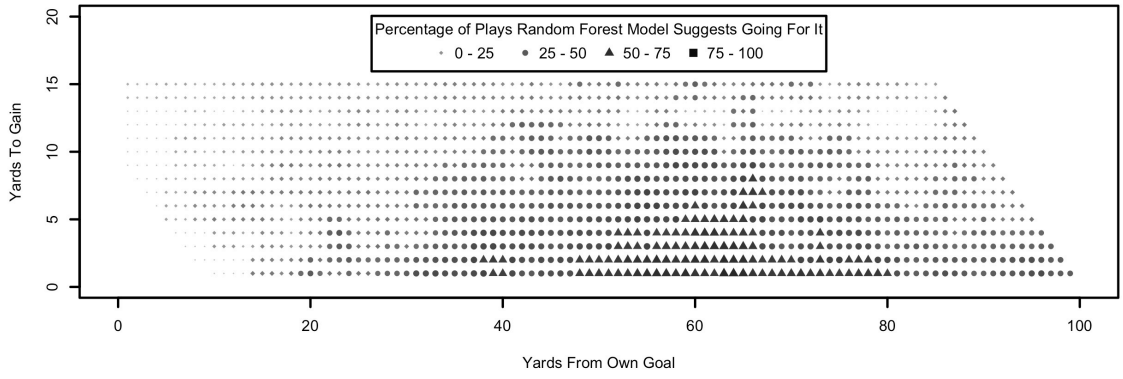

Wins
0
5
10
15
55
60
65
70
75
80
% Agree with Romer Model

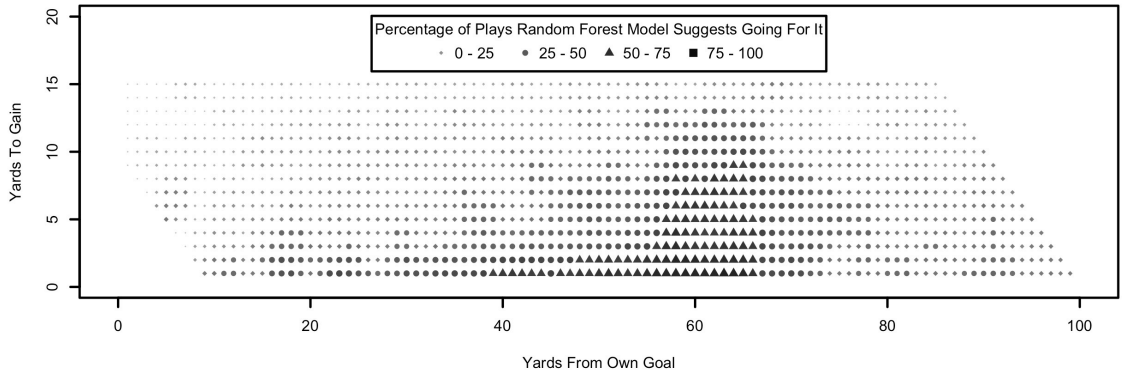

Wins
0
5
10
15
50
55
60
65
70
75
80
% Agree with NYT Bot

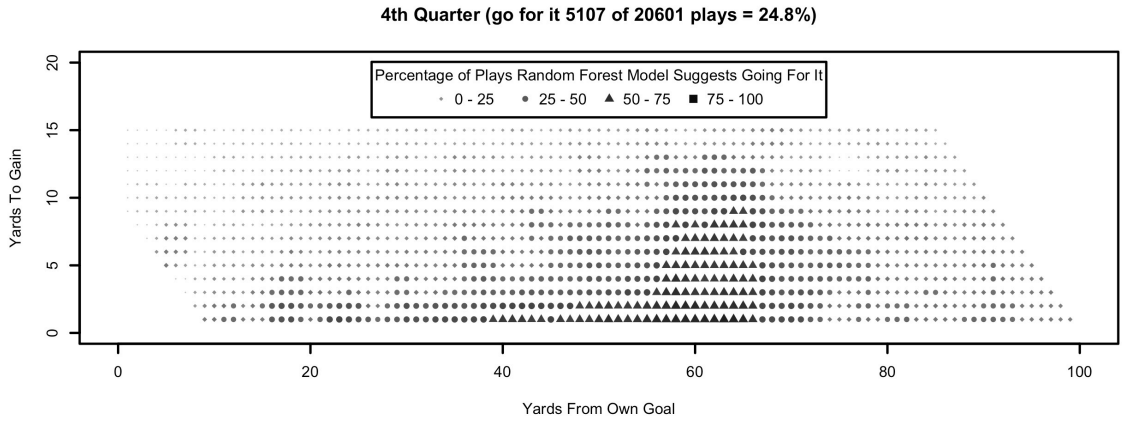

Wins
0
5
10
15
30
40
50
60
70
80
90
% Agree with Random Forest Model

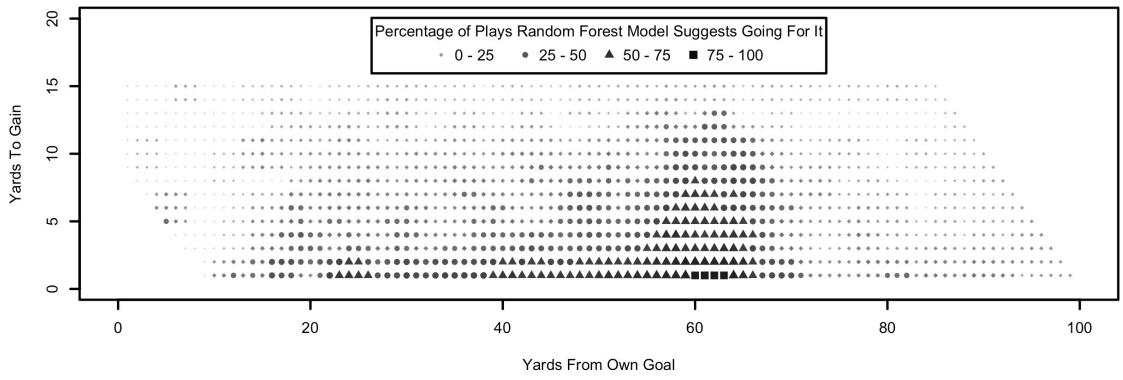

1st Quarter (go for it 5585 of 16072 plays = 34.7%)
Percentage of Plays Random Forest Model Suggests Going For It
0 - 25
25 - 50
50 - 75
75 - 100
Yards To Gain
Yards From Own Goal

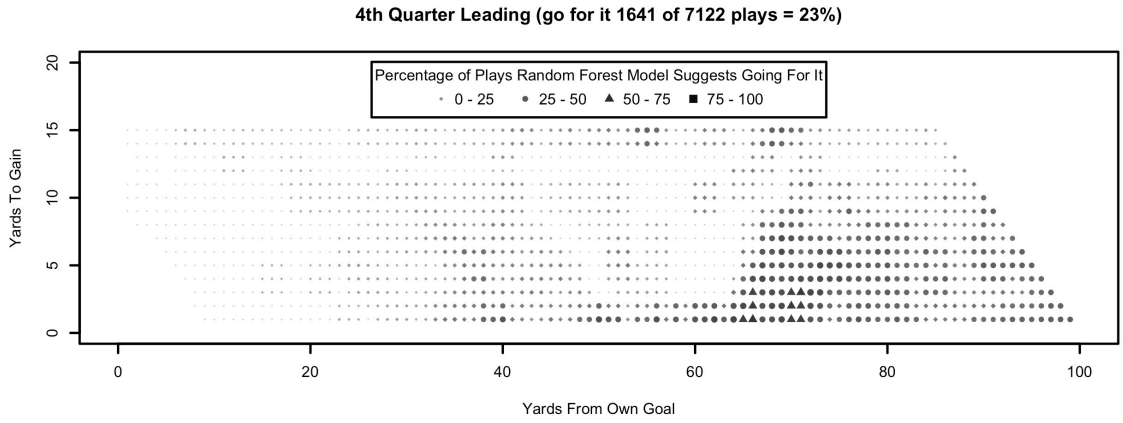

2nd Quarter (go for it 6988 of 19972 plays = 35%)
Percentage of Plays Random Forest Model Suggests Going For It
0 - 25
25 - 50
50 - 75
75 - 100
Yards To Gain
Yards From Own Goal

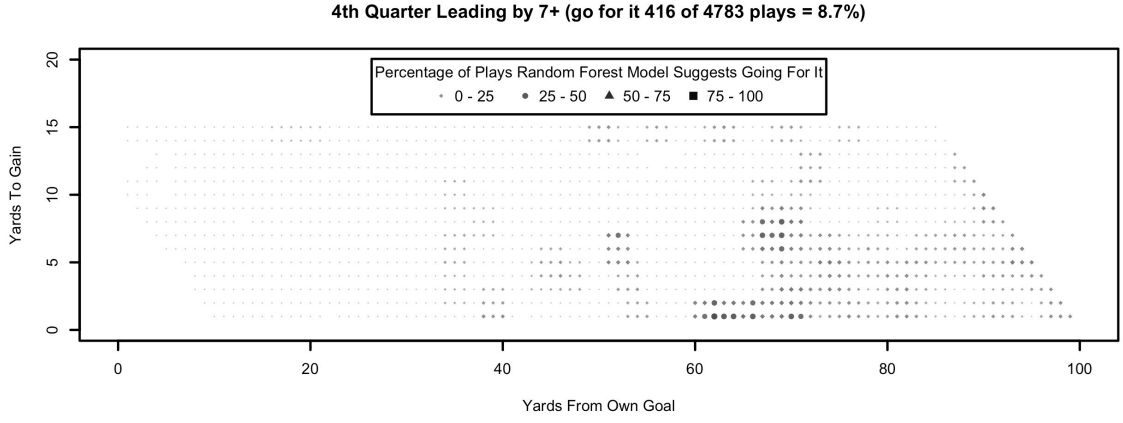

3rd Quarter (go for it 4785 of 15861 plays = 30.2%)
Percentage of Plays Random Forest Model Suggests Going For It
0 - 25
25 - 50
50 - 75
75 - 100
Yards To Gain
Yards From Own Goal

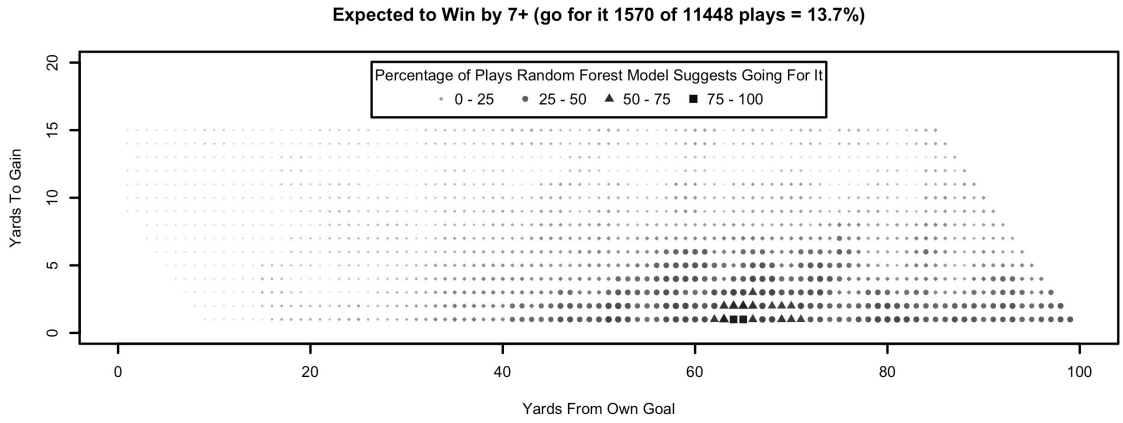

4th Quarter (go for it 5107 of 20601 plays = 24.8%)
Percentage of Plays Random Forest Model Suggests Going For It
0 - 25
25 - 50
50 - 75
75 - 100
Yards To Gain
Yards From Own Goal

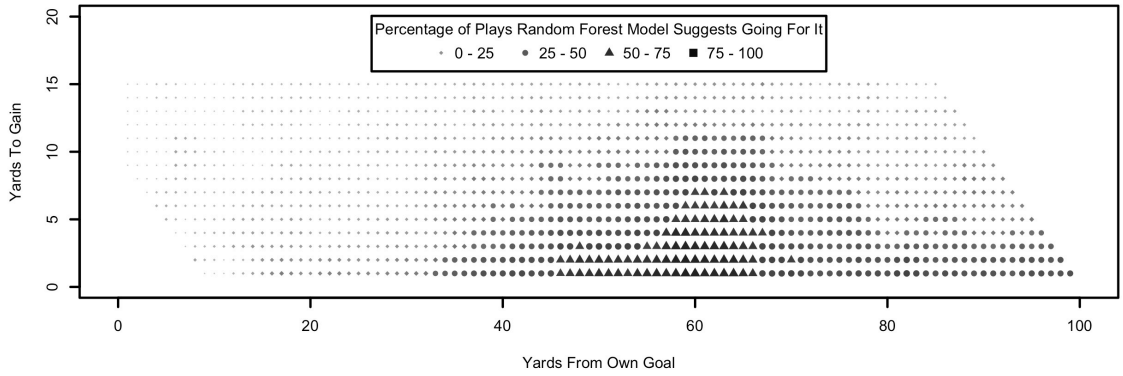

4th Quarter (go for it 5107 of 20601 plays = 24.8%)
Percentage of Plays Random Forest Model Suggests Going For It
0 - 25
25 - 50
50 - 75
75 - 100
Yards To Gain
0
5
10
15
20
0
20
40
60
80
100
Yards From Own Goal

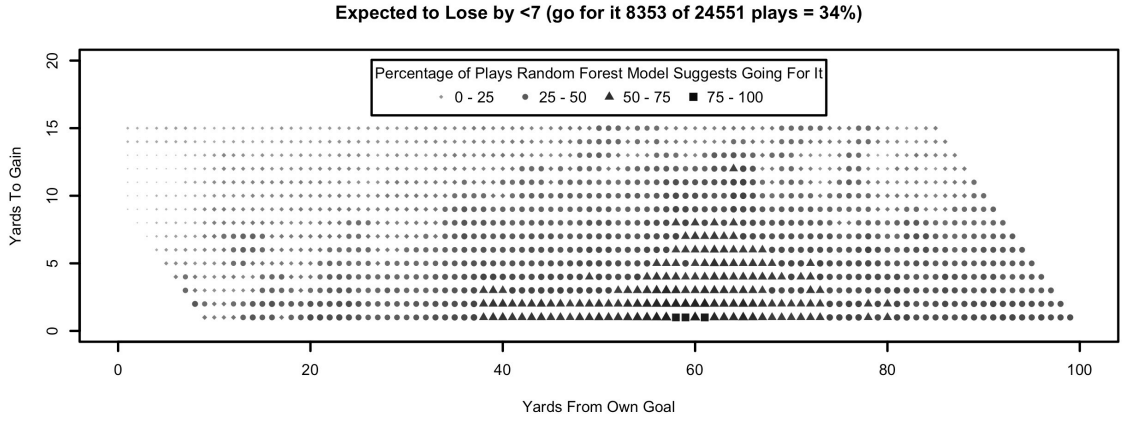

4th Quarter Trailing (go for it 3007 of 12275 plays = 24.5%)
Percentage of Plays Random Forest Model Suggests Going For It
0 - 25
25 - 50
50 - 75
75 - 100
Yards To Gain
0
5
10
15
20
0
20
40
60
80
100
Yards From Own Goal

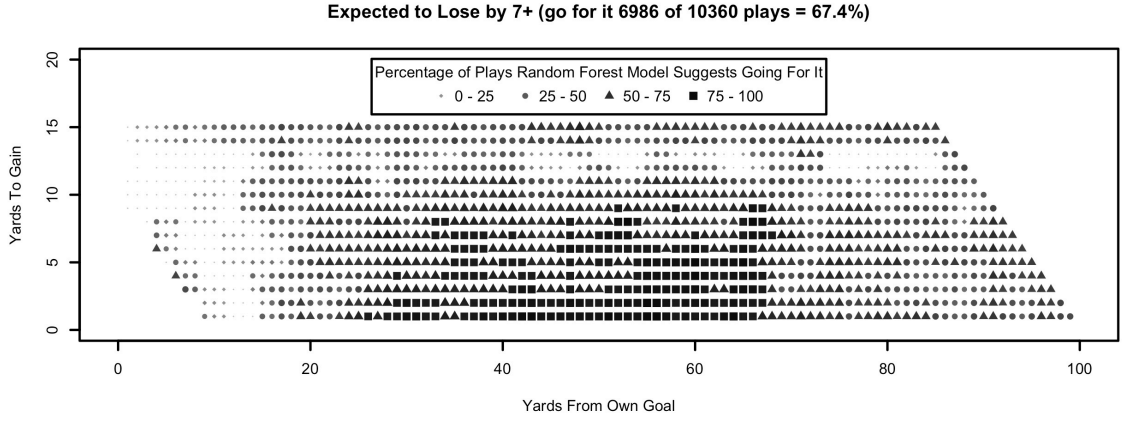

4th Quarter Leading (go for it 1641 of 7122 plays = 23%)
Percentage of Plays Random Forest Model Suggests Going For It
0 - 25
25 - 50
50 - 75
75 - 100
Yards To Gain
0
5
10
15
20
0
20
40
60
80
100
Yards From Own Goal


4th Quarter Leading by 7+ (go for it 416 of 4783 plays = 8.7%)
Percentage of Plays Random Forest Model Suggests Going For It
0 - 25
25 - 50
50 - 75
75 - 100
Yards To Gain
0
5
10
15
20
0
20
40
60
80
100
Yards From Own Goal


Expected to Win by 7+ (go for it 1570 of 11448 plays = 13.7%)
Percentage of Plays Random Forest Model Suggests Going For It
0 - 25
25 - 50
50 - 75
75 - 100
Yards To Gain
0
5
10
15
20
0
20
40
60
80
100
Yards From Own Goal


Expected to Win by <7 (go for it 5525 of 26003 plays = 21.2%)
Percentage of Plays Random Forest Model Suggests Going For It
0 - 25
25 - 50
50 - 75
75 - 100
Yards To Gain
0
5
10
15
20
0
20
40
60
80
100
Yards From Own Goal


Expected to Lose by <7 (go for it 8353 of 24551 plays = 34%)
Percentage of Plays Random Forest Model Suggests Going For It
0 - 25
25 - 50
50 - 75
75 - 100
Yards To Gain
0
5
10
15
20
0
20
40
60
80
100
Yards From Own Goal


Expected to Lose by 7+ (go for it 6986 of 10360 plays = 67.4%)
Percentage of Plays Random Forest Model Suggests Going For It
0 - 25
25 - 50
50 - 75
75 - 100
Yards To Gain
0
5
10
15
20
0
20
40
60
80
100
Yards From Own Goal