Ethan Lee and Janet Li
STAT 143 Final Project

When Should NFL Teams Run and Pass on Fourth Down?

Introduction

Fourth down plays have been an eternal point of controversy in the NFL. From Bill Belichick's infamously risky decision to go for it on 4th and 2 from his team's own 28-yard line to the Packers' field goal attempt from the 8-yard line while down by 8 with little time remaining, fourth downs often provide a platform for coaches to take risks by choosing a single action for their teams to take that could determine the overall outcome of the game.

Historically, coaches very rarely go for it on 4th down, not even on 4th and 1 (Figure 1). Instead, coaches prefer to choose the safer choices of a field goal or punt, which have higher probabilities of success despite lower potential rewards.

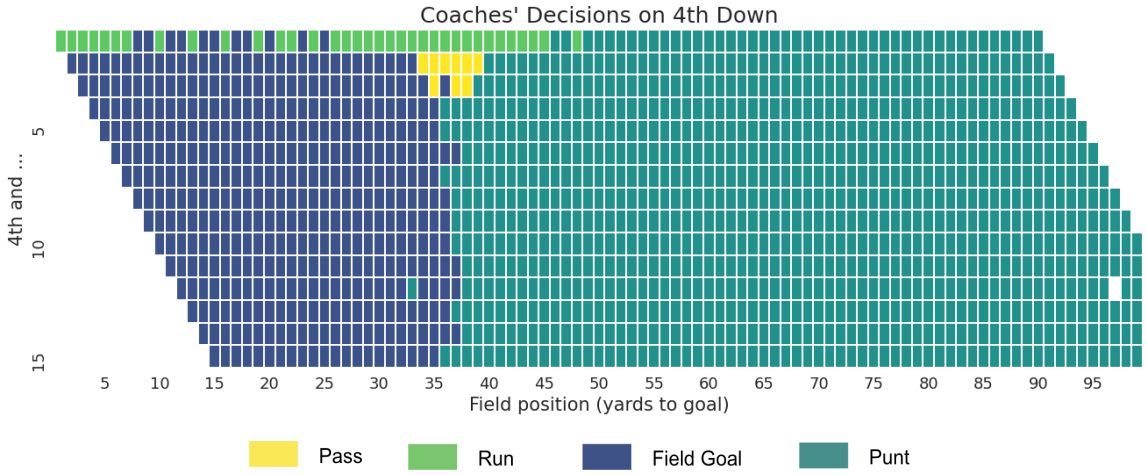


Figure 1. Most popular 4th down decisions by coaches at each yardline and yards to go.

To better understand how fourth downs should be treated in different circumstances, we used this project as an opportunity to explore two different methodologies – one using expected points and the other using the win probability resulting from a fourth down play – for determining the best possible actions to take on fourth down. The primary dataset used was a CSV file of all plays from NFL games played during the 2009-2018 seasons.

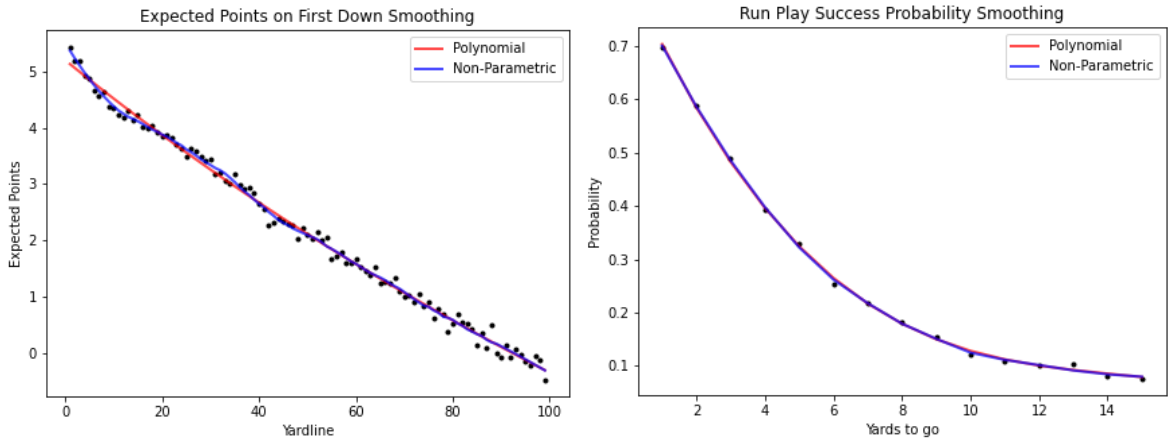
Methodology 1: Expected Points Model

To determine the best fourth down decision, we first calculated the expected value of each play type, and chose the play that maximized the expected value, in terms of points. We used the law of total probability to obtain the following general formula, given the situation S:

$$E(\text{play}) = P_{\text{play}}(S) \times E(\text{Value} \mid S) + (1 - P_{\text{play}}(S)) \times E(\text{Cost}_{\text{turnover}} \mid S),$$

where $P_{play}(S)$ is the probability of success (i.e. making the first down), E(Value | S) is the expected value of that success, 1 - $P_{play}(S)$ is the probability of failure, and $E(Cost_{turnover} | S)$ is the cost of giving the opponent possession. All four terms are conditioned on the situation S.

First, we calculated the expected value of a first down, in terms of points, given the current yardline. To do so, we took the average of the next score relative to the team in possession for all plays at each yardline, and used a non-parametric approach to smooth the values (Figure 2(a)).



(a) Expected Points on First Down **(b)** Run Play Success Probability

Figure 2. Cubic polynomial (parametric) and Savitzky-Golay filter (non-parametric) smoothing.

The expected values of a run or pass on 4th down, given the yardline and yards to go, are:

$$E(run) = P_{run}(S) \times E(Value_{1D} | S) + (1 - P_{run}(S)) \times E(Cost_{turnover} | S)$$
$$E(pass) = P_{pass}(S) \times E(Value_{1D} | S) + (1 - P_{pass}(S)) \times E(Cost_{turnover} | S)$$

We calculated the probability of successfully making the first down by finding the proportion of successful run or pass plays for each value of yards to go, then smoothing the probabilities using a cubic polynomial for each yardline (Figure 2(b)). Then, we multiplied the probability of success by the expected value of possession at the first down yardline.

The cost of failing to make the first down is the value of giving the opponent possession at that yardline. We took the average yards gained on a failed run or pass play into account, for each yardline and yards to go, smoothed the values using a cubic polynomial, and adjusted the opponent's starting yardline accordingly.

In addition, for passing plays, we also included the expected cost of an interception. To do so, we found the probability of an interception and used the average return yards after an interception to calculate the value of the opponent's starting field position. We also included the probability of an interception being returned for a touchdown, and the value of a touchdown. We used the nonparametric approach to smooth both interception probabilities, and the average return yards.

Assuming that punts are always successful, the expected value of punting the ball on 4th down, given the yardline, is equivalent to the value to the opponent of possession after the punt:

$$E(\text{punt}) = E(\text{Cost}_{\text{Punt}} \mid \text{yardline}) = -1 \times E(\text{Value}_{1D} \mid \min(100 - \text{yardline} + \text{avg punt distance}, 80)).$$

We calculated the average punt distance from each yardline, then smoothed the values using the nonparametric Savitzky-Golay filter. Using the average punt distance, we calculated the opponent's expected starting yardline after the punt, which we then used to calculate the expected value of giving the opponent possession.

Next, we calculated the expected value of a field goal, given the yardline:

$$E(\text{FG}) = P_{\text{FG}}(\text{yardline}) \times E(\text{Value}_{\text{FG}} \mid \text{yardline}) + (1 - P_{\text{FG}}(S)) \times E(\text{Cost}_{\text{turnover}} \mid S)$$

We calculated the probability of a field goal at each field position up to 50 yards, then smoothed the values using the nonparametric Savitzky-Golay filter. We set the probability of a field goal from beyond 50 yards to 0. The expected value of scoring a field goal is 3 points, but we need to subtract the value to the opponent of a first down after kickoff, which is typically 75 yards from the end zone. Finally, the cost of a missed field goal is the value to the opponent of a first down at the kick position, which is approximately 8 yards beyond the line of scrimmage or at the opponent's 20-yard line, whichever is closer (from the opponent's perspective).

Expected Points Model Results and Discussion

Our overall model recommendations are depicted in Figure 3. According to the model, coaches should attempt to go for it on 4th down much more frequently—in particular, coaches should always run the ball on 4th and 1 and 4th and 2. When teams are close to the end zone, the model favors passing plays over field goals, since the expected value of a touchdown is higher than the expected value of a field goal. Around the 30-40 yard line, the model again favors passing plays over field goals, since the probability of making the field goal is low. In addition, teams can still gain yardage on passing plays even if they fail to make it to the first down, causing the opposing team to gain possession further from the end zone. As we continue to move further from the end zone, the recommended decision shifts from field goal or pass to punt, since the cost of failure is much higher if a field goal or pass play fails far from the end zone. This is because the opposing team will gain possession close to the end zone, and will be very likely to score. However, this model ignores two very important factors in coaches' decision-making: the current scoreline, and the time remaining on the clock.

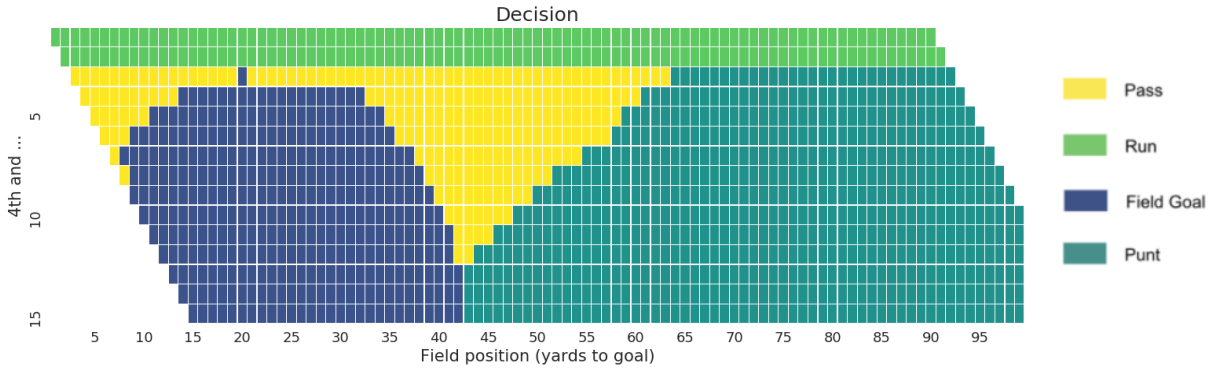


Figure 3. Model recommendations at each yardline and yards to go.

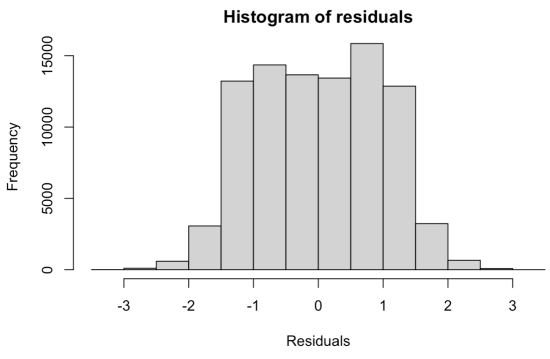
Methodology 2: Win Probability Model

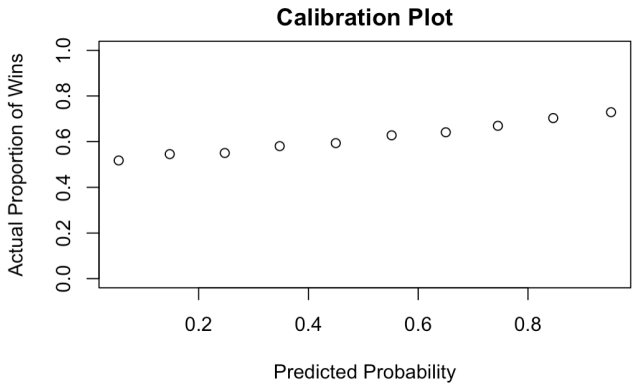
To account for the time remaining in the game and the score margin (relative to the team in possession), a win probability model taking these into consideration was designed. The desired model would accurately calculate the winning probability of a team with a first down at some point in the game, based on the score margin, time remaining, and yardline of possession. This model could then be used to calculate the winning probability resulting from some action taken by a team in possession on fourth down, since each fourth down results in some first down.

Two main approaches were considered for fitting the first down winning probability model. Because a winning probability is a value between 0 and 1, a simple logistic GLM predicting whether a team in possession wins a game based on the time remaining, score margin, the ratio of the score margin and the time remaining (as an interaction term) and yard line, as well as a logistic generalized additive model using the same predictors were considered. The models were fit on a randomly generated training dataset from a dataset of first down plays, with the train-test split being 70-30. Several model performance metrics are shown in the table below:

Model	**Log Loss**	**Train RMSE**	**Test RMSE**
GLM	1.185036	0.3948303	0.3938371
GAM	0.533655	0.3939117	0.3932839

The generalized additive model performed better than the generalized linear model based on log loss and test RMSE, which was expected since the relationship between whether a team won a game and factors such as time remaining and score margin would not be linear. Thus, the generalized additive model (GAM) was chosen to be used as the model for predicting winning probability based on the previously mentioned predictor variables. With regards to model assumptions, the residual histogram below shows the residuals are normally distributed, although there is concern for the independence assumption since certain score margins may appear more at certain times (which is why an interaction term was included). Lastly, the response variable is indeed binary, the sample size for training is sufficiently large, and there is little sign of overfitting due to the closeness of the train and test RMSEs. We can further observe the level of accuracy of the generalized additive model through the calibration plot below, which plots the average proportion of wins for teams in possession of the ball on plays with a predicted probability between 0 and 0.1, between 0.1 and 0.2, and so on, by the GAM.





Based on the calibration plot above, it is clear that the win probability model does not completely accurately predict whether a team in possession of the ball will win a game based on situational factors on first down. However, this can be explained by the fact that a single first down play typically will not be an indicator of whether a team in possession will ultimately win the game; furthermore, the calibration plot shows that teams that were in possession of the ball on first down plays that were predicted to have lower probabilities of winning the game did have actual proportions of wins that were lower than those of teams that had possession on first down plays that were predicted to have higher win probabilities (since the slope of the calibration plot is positive). Thus, the win probability model does seem to generally serve its purpose.

Using the generated win probability model, we then created a model for suggesting a specific action for NFL teams to take on fourth down. Specifically, because the win probability model predicts the win probability of a team in possession on first down, we could use similar equations to the previous expected points model by simply replacing the expected points of an action on fourth down with the resulting win probability caused by the action on fourth down. Then, for each possible yard line, yards to go, time remaining in the game, and score margin, the model predicts the win probability resulting from a punt, field goal attempt, run attempt, or pass attempt on fourth down using the following general formula, where WP is the win probability, P_{play} is the probability of success of a play, S is the indicator of whether the play succeeded, F is the indicator of whether the play failed, and "E(Factors)" is the expected values of the time remaining, score margin, yardline, and yards to go after the play succeeds or fails:

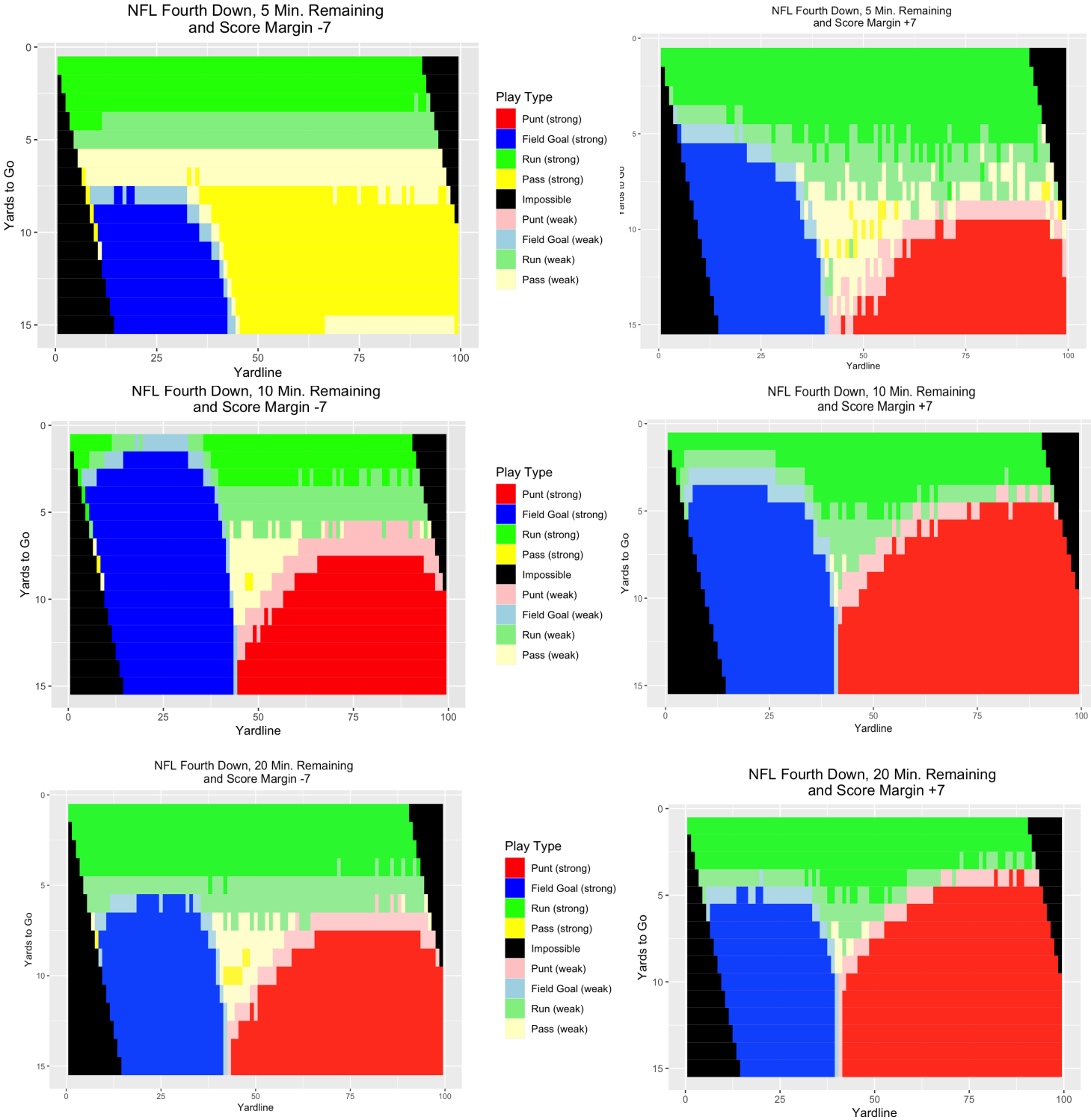
$$WP(play) = P_{play}(S) \times WP(E(Factors) \mid S) + (1 - P_{play}(S)) \times WP(E(Factors) \mid F)$$

Specifically, the win probabilities for each play were calculated by either calculating the new win probability on first down for the team originally in possession on fourth down if the fourth down play resulted in a new first down, or by calculating the complement of the win probability for the team defending the fourth down if the fourth down play resulted in a first down for the defending team. The smoothed probabilities of success for different plays, probabilities of interceptions, and probabilities of pick-sixes were calculated as in the expected points model, but on third and fourth down data only, in order to accurately reflect fourth down situations. Additionally, the average number of yards gained on a failed run/pass play given yardline and yards to go were considered by adjusting the yardline accordingly when calculating win probability for the defending team gaining possession after a failed run/pass play.

Win Probability Model Results, Discussion, and Conclusions

As a first result, we built an R Shiny web app that gives the new win probabilities when any of a punt, field goal, run play, or pass play are attempted using the user's inputs of yardline, yards to go, time remaining, and score margin. (The app takes a few minutes to load due to time needed to access datasets). Additionally, to visualize the results of the win probability-based fourth down model, we generated "heatmaps" of the suggested action to take for each yardline and number of yards to go currently in effect for a team in possession on fourth down, with the

time remaining and the score margin fixed for each plot. In the following plots, the row of a cell represents the yards to go, and the column represents the yardline that the team in possession is on. A bright-colored cell indicates an action that results in the highest win probability, where this win probability is at least 0.005 greater than the largest win probability resulting from taking a different action (a faded-colored cell indicates that this difference is less than 0.005).



We can compare the six plots above, in which we consider the cases when a team is up 7 or down 7 in a game that has 5, 10, and 20 minutes remaining. The plots show that when a team is down 7, they should take riskier plays than when they are up 7; furthermore, as the time remaining for a team that is down 7 goes down, they are encouraged to be riskier by running or passing the ball on larger yards to go and yard lines. When a team is up 7, they are generally encouraged to be riskier when there is more time remaining, as can be seen by comparing the plots for a team up by 7 with 10 minutes remaining and a team up by 7 with 20 minutes remaining. However, interestingly, the model actually suggests riskier plays to some extent when there are 5 minutes remaining and a team is up by 7; we can see that this might be explained by the fact that the plot for a team up by 7 with 5 minutes remaining has many faded colored cells, meaning that the possible actions on fourth down result in win probabilities that are extremely close to each other. Additionally, the model takes into account the average time taken by a run or pass play attempt, and so the decrease in time remaining caused by such a play may be evaluated as advantageous enough by the model to encourage a team with a lead with little time left to go for riskier plays to run out the clock. Thus, because of these factors, it would be advisable for any users of this model to less strictly follow it when there is little time remaining or when their team is in a situation corresponding to a faded cell on a heatmap, since in that case the differences in win probabilities caused by possible fourth down actions are extremely small, and so the user should rely on their own judgement.

The win probability fourth down model itself thus largely worked as expected—that is, when there is more time remaining and a team has a lead, the team is encouraged to run/pass more, although teams with leads and very little time remaining were actually encouraged to make riskier plays. As explained before, this may have been caused by very close win probabilities resulting from fourth down actions and an unreasonable consideration of time by the model with little time remaining, and so users should refer to this model in mid-game situations as mentioned before. When a team has a deficit, they are encouraged to play much more riskily, especially when there is little time remaining.

Conclusion

In regards to the original point of investigation, both models make clear that run plays should be attempted when there are less yards to go, while pass attempts should be made when there are more yards to go on fourth down. In addition, the various results previously discussed illuminate the differences between suggestions made by the expected points and win probability fourth down models.

Overall, the expected points and win probability models had similar field goal and punt boundaries. However, teams were encouraged to pass the ball more often in the expected points model, while teams were encouraged to run more often in the win probability model. The win probability model encourages running the ball on many 4th down situations with 1 to 5 yards to go, whereas the expected points model only encourages running the ball with 1 or 2 yards to go. One explanation for this disparity is that passing plays generally result in more yards gained, and

thus higher expected points on success. The win probability model may also favor run plays because of the lower risk—passing plays are riskier due to the chance of interception. This also may have been a result of training the run and pass success probabilities on only third and fourth down data, in which coaches may have more planned run schemes for successfully reaching a first down.

An important point is that the models generated as a result of this project should not be considered as definitive guides to what actions teams should be taking on fourth downs. Rather, they can serve as useful references or suggestions for teams at fourth down, but users should take into account various other factors not accounted for by the models, which are noted below.

Model Flaws and Future Analyses

The primary flaw of the expected points model is the failure to account for time remaining and score margin, two key factors in any NFL game, both of which were addressed by the win probability model. Some flaws of the win probability model include that it behaves unexpectedly in situations where a team has a large lead but has very little time remaining; the risky plays suggested by the model in these circumstances may be due to the extremely high likelihood that the team with the lead will ultimately win, causing the team to be encouraged to try riskier plays that take a long time in order to take time off the clock. Another flaw in the win probability model may be the choice of a generalized additive model for win probability on first down. The smoothed functions fitted by the model for each predictor variable may not necessarily be completely correct; additionally, although the log loss of the generalized additive model was lower than that of the GLM, it was still rather high, indicating that although the generalized additive model is incorporated in the popular nflscrapR package's win probability model, future extensions could incorporate other model types for the win probability.

There are various other ways to improve upon the models used in this project and expand into future research. First, the models do not take into account the abilities of players and teams involved, but these abilities should greatly affect a coach's decision on fourth down; for example, a team with a kicker who is known to miss long-range field goals might decide to punt or go for it more often instead of attempting a field goal. Team strengths could be accounted for by using pre-game betting odds as factors to identify how stronger and weaker teams should act on fourth down; additionally, player abilities could be evaluated by incorporating statistics such as field goal percentage and average rushing or passing yards per play for each team.

Some further extensions would be to take into account the probability of success for a punt (which could go out of bounds or be blocked) and the number of timeouts remaining for each team as a predictor variable (which should affect how important the time remaining in the game is for the team on fourth down). Given the necessary data, these smaller extensions could most likely be implemented by simply adding them into the model as additional predictors.