Optimizing 4th Down Decision Making in the NFL: A Machine Learning Approach

Ilan Barr, Eric Fang

Abstract

In the high stakes world of American Football that is the NFL, the decisions coaches make, no matter how small, can be the difference between winning and losing. Specifically, the decision to go for it, or not, on 4th down represents a critical juncture that can significantly influence the game's outcome. In this study we use game theoretical concepts to create a novel machine learning framework designed to optimize 4th down decision making. Leveraging the NFL fastR dataset which includes play-by-play data from 2013-2023 NFL seasons, we employ regression models to simulate the decision-making processes coaches go through on game day. Our finished product takes in situational information such as the game-state, historical team data, weather, and stadium conditions to output a decision recommendation (punt, go for it, or attempting a field goal) that maximizes the likelihood the team will win. In the case where going for it is recommended, we then make another recommendation to pass or run the ball depending on the formation of the defense. In a league where the difference between winning and losing has enormous financial implications, we have found, NFL coaches are not making the correct decisions nearly enough. These coaches are the top 32 in their respective profession in the world, and are under performing. Our hope is to inform them to encourage more competitive decision making going forward.

Introduction

In the ever-changing world of American Football, the decision making behind the fourth down play is crucial in securing victory for a team. Whether the offense chooses to go for it, kick a field goal, or punt, can alter game outcomes because of the high leverage nature of the 4th down play. The offense must decide whether they want to willingly give the ball to the other team, in a more advantageous position (punt), attempt to keep the ball but risk giving the other team the ball in a disadvantageous position (go for it), or "settle" for 3 points in a field goal attempt instead of the potential 6,7, or 8 points awarded for a touchdown. Additionally, the defense needs to respond, and also choose a course of actions for their strategy on 4th downs. NFL coaches for a long time have adopted a very conservative strategy. They choose to punt the football away to the other team in most situations, relying on their defense to limit the amount of their opponent scores. Willingly giving the ball to the other team is not always the best decision. In some scenarios a calculated risk/reward payoff can indicate that taking a gamble on 4th down and going for it can actually increase a team's expected win probability.

Historically, traditional coaching approaches during fourth down usually leaned towards being more conservative. Being risk averse leads to less opportunities to look "stupid", a fear for NFL coaches, because there is seldom stability in their jobs. An average of 6.8 coaches are fired every season, or roughly 21% of team's fire their coach in a given season. While taking more aggressive actions will not work out every time, it is important to not be results oriented. For example thinking going for it and not convert the 1st down is was the wrong decision is not correct. Our model aims to empirically determine whether or not it is the correct decision, however it is important to note the validity of the decision is completely independent of the outcome.

With this study, we intend to revolutionize the way 4th down decisions are made using a game theoretical approach. Coaches are well versed in how football games work and use good rules of thumb such as looking at how much time is remaining or how many yards are left to make their decisions. Thus, the choice that the offensive coach makes, the fourth down decision, can be mathematically modeled using game theory in conjunction with calculating the mixed strategy of the offensive and defensive teams. We can estimate the action which would yield the best win probability for a coach's team perform based on historical data and football game theory.

Our study is based on the assumption that coaches under utilize the multitude of data that is available when making the fourth down play decision. Football is an extremely nuanced sport where there are thousands of different possibilities for each occurrence of a 4th down. Some 4th down decisions are fairly obvious and a coach wouldn't need our model to help them with their decision, however, in some situations the decision is not so clear. Making the correct decision when in critical moments, increases a team's probability of winning, and no matter how marginal, will yield to more success in the long run for the football program.

Our model relates a multitude of factors such as game state information, team's strengths and weaknesses, game context, and opponent tendencies. The goal of the research is to provide a better way to model the future of the sport's

decision making. The research also attempts to analyze the decision making in a more theoretical approach and providing a strong framework for other similar sporting decisions to be played out, for example, modelling whether a football team should go for 2 rather than kick an extra point.

The 4th down decision making situation can be modeled as an example of a Stackelberg game. The offense (leader) faces a decision on 4th down to either go for it, kick a field goal, or punt. There is an element of incomplete information because the defense does not know the exact play the offense will run. They can infer based off the formation and the personnel group the offense lines up in (it is fairly obvious to the defense when the offense lines up in punt or field goal formation, however, there is always the possibility of a fake). Specifically, if the offense sends their punter or kicker onto the field, the defense can deduce a punt or kick respectively will occur, and defend the play accordingly. Additionally, there are certain offensive formations that can tip the defense off about what type of play the offense might run. An example of this is when the offense lines up in empty formation, a setup involving a quarterback, five total eligible pass catchers lined up widely on either side and five pass protectors. This formation does not include a running back, thus it may indicate to the defense that barring a quarterback run, the offense is planning a passing play. Each teams holds a set of beliefs about themselves and their opponent, including what their opponent is likely to do given historical data. The offense must choose a decision that will maximize their payoff, taking into account the opposition's reactions in the event they succeed, and fail. Once the probabilities of success and failure are estimated, an informed and game theoretically optimal solution/decision can be made.

To achieve this, we collected play-by-play data from NFL fastR which provides a comprehensive dataset on American football games. After cleaning the data and going through the process of modelling, we would predict the expected value of each fourth down action.

During the data exploration phase of our project, we created some graphics to understand the relationship between running and passing the football at different yardages in the field. We also looked at the success rate of both yardages and down / distance.

Model

For the sake of simplicity, we will first look at the pure strategy that each team should take. That is we want to pick the strategy/option that maximize win probability. In Figure 5, we look at how this could happen with a sequence of states and decisions that each team performs.

Describing Figure 5, we can see that the offense during the fourth down, first decides to either punt, go for it, or attempt a field goal. This decision is solely based on the offense. In the deciding to punt or attempt a field goal, the defense is unable to respond and we assume that the fourth down decision ends in these cases. In the case where the offense goes for it, we want to analyze how the defense's strategy might affect the payoff of the offense going for it.

For the defense's strategy, we look at the defenders in the box. Once the offense has decided they will go for it,

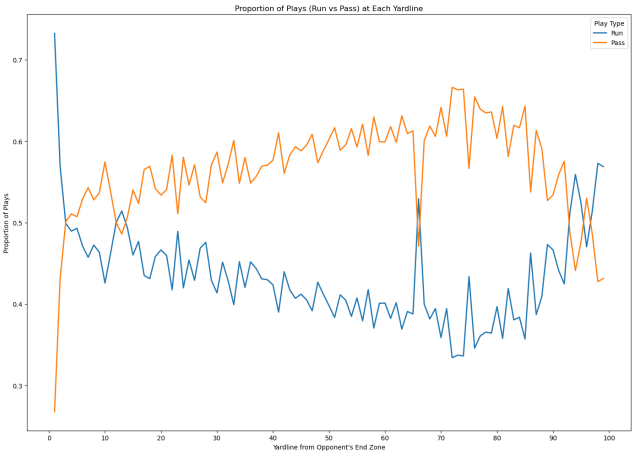

Figure 1: Proportion of Plays (Run/Pass at each yardline

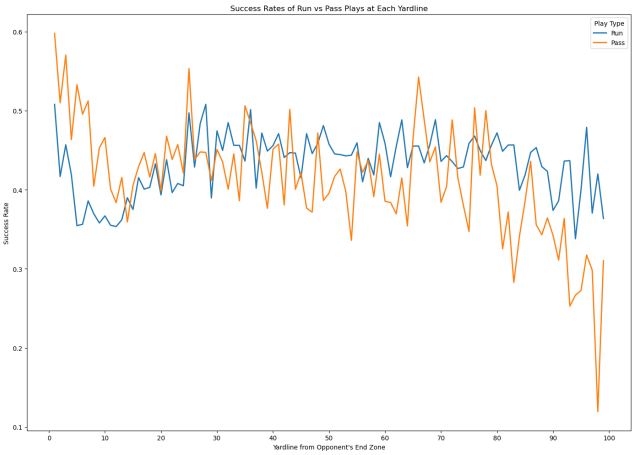

Figure 2: Play success rate at each yardline (run/pass)

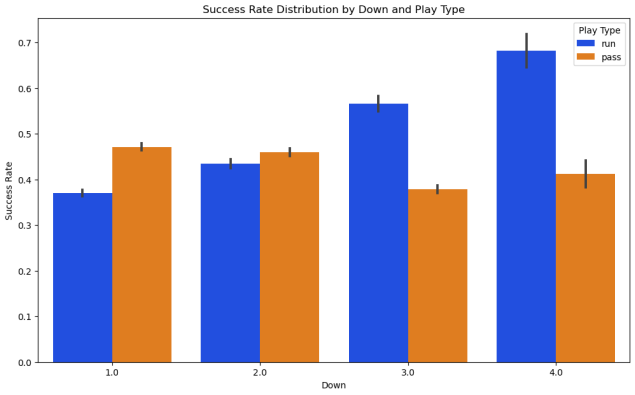

Figure 3: Success rate pass/run on each down

and sends their normal offensive players onto the field, the defense has a chance to respond with their own personnel group. Lastly, as before the offense starts the play, they are able to observe the number of defenders in the box and if they want to can change their play call to either a pass or a run. We have created a logistic regression model which takes in the following features:

- yardline_100: Distance to the opponent's end zone
- ydstogo: Yards needed for a new set of downs
- goal_to_go: 1 if it's a goal-to-go situation, otherwise 0
- n_defense_box: Number of defenders in the box

This model predicts the probability of success for both passing and running and outputs the option to maximize success.

Offensive Strategy

We first needed to model the pure strategy the offense would perform. That is we want to model the recommended action a coach would pick given the state of a game.

In order to complete our goal, we created two layers of models. The first layer estimates the probability a certain action will be successful. E.g. going for it on 4th down, or kicking a field goal (we did not model punt success % and are disregarding blocked punts for this project). The second layer model predicts the next game state given an event occurred. For example, if a field goal is made, the next game state is the team who kicked the field goal's score increases by 3, 5 seconds run off the clock, and the opposing team will get the ball at their own 25 yard line with a 1st and 10 (for simplicity we also are assuming touch backs on all kickoffs). For FG_make, FG_miss, and GFI_success when in a goal to go situation, we did not need to create a model, as the next game state is elementary to figure out (FG_miss results in opposing team gaining possession 7 yards behind previous line of scrimmage and GFI_success leads to possession team scoring a touchdown, gaining 7 points, and kicking off). For GFI_success in a non goal-to-go situation, GFI_fail, and punt we created regression models to predict the next game state after each event.

Next, for a given game state, we output 5 possible "next game states" which would be the game state assuming FG_miss, FG_make, GFI_fail, GFI_success or Punt. We ran these game states through nflfastR's calculate_win_probability model to output a predicted win probability for each possible next game state.

Lastly, we used our field_goal_success and GFI_success models to predict the probability of success/failure for kicking a field goal or going for it. Then, in conjunction with the outputted win probabilities for each next game state we were able to calculate the final recommendation:

Win Probabilities:
$GFI_{wp} = GFI_{fail} \times (1 - GFI_{percentage}) + GFI_{SUCCESS} \times GFI_{percentage}$
$FG_{wp} = FG_{SUCCESS} \times FG_{percentage} + FG_{FAIL} \times (1 - FG_{percentage})$
$PUNT_{wp} = Punt$

$$\text{Reccomendation} = \text{argmax}(PUNT\_wp, FG\_wp, GFI\_wp)$$

The mixed strategy comes into play when attempting fake punts or field goals. Since fakes only occur about 1% of the time, the defense likely does not expect a fake on any given attempt, however, they must be ready for it on the chance that it occurs. The threat of a fake influences how the defense sets up on any given punt or field goal. Other than faking a punt or field goal attempt, the model focuses on pure strategy. For example, given an input into the model of a specific 4th down instance, the model will usually give a pure strategy recommendation of going for it, kicking a field goal, or punting. In other words, every scenario where the parameters are the same as this one, the model will recommend the offense does the same thing. In some situations, it is advantageous to adopt a mixed strategy approach where occasionally taking increased risk and faking a field goal attempt or punt while attempting to catch the defense off guard can yield a higher win probability in expectation.

In building the first model, we used features such as historical success in completing that play (going for it on 4th down or kicking a field goal), the weather conditions, yards to go, time remaining, stadium roof type, and historical success on defense defending against similar play types. We also looked at league trends regarding success rates in these categories.

The first step taken to build the model was pull the data set from NFL fastR. We chose to use play by play data from every game from the 2013 season through the most recent 2024 season.

We chose to use the Logistic Regression and Random Forest machine learning models as our solution for this prediction problem. Our main reason for choosing these model is its robustness to sparse data. There are only 285 NFL games every season, thus, the likelihood two teams find themselves in the exact same scenario is extremely slim. Our model will look at similar situations and deduce what the success percentage may be, along with the suggested best action given these success predictions.

Table 1 shows an example of what the decision model does. Firstly given the game state, it predicts the success rate of each action (Field Goal, Go for it) and predicts the resulting win % of the resulting games state. In this example, if the teams punts, the model predicts a win % of 78%. If the team goes for it, they will have an average win % of 77%. This is calculated by taking a weighted average of the probability they are successful/not and the resulting win %.

$$\begin{aligned}\text{Go for it WP} =&(\text{Success \%} \cdot \text{Succeed WP}) \\ &+ ((1 - \text{Success \%}) \cdot \text{Fail WP})\end{aligned}$$

$$\text{Go for it WP} = 0.27 * .87 + (1 - 0.27) * .74 = 0.77$$

Background on Win Probability Model

We have chosen to use NFL fastR's win probability model, instead of creating our own. Their model is based on Yurko, Ventura, and Horowitz's nflscrapR model from the 2018 paper, "nflWAR: A Reproducible Method for Offensive Player

	Win %	Success %	Win % if	
			Fail	Succeed
Punt	78	NA	NA	NA
Go for it	77	27	74	87
Field goal attempt	72	0	72	83

Table 1: Example Game State: Visiting team, up 7 points, has possession of ball 45 yards from opponent's end zone. 4th & 11, qtr 3 12:39 remaining both teams have 3 timeouts. Recommendation: Punt (1% WP)

Evaluation in Football (Extended Edition)". These authors made breakthroughs into statistical analysis in football in their creation of the Expected Points Added and Win Probability Added metrics. These stats essentially predict a team's expected points to be scored next at any given point in the game. Negative expected points imply the opposing team is expected to score points next. The features that went into this model were field position and down and distance. Ben Baldwin and the team at Open Source Football expanded on the nflscrapR by creating an xgboost model using the following features for their win probability model:

- Seconds Remaining in Half
- Seconds Remaining in Game
- Yard Line
- Score Differential
- Ratio of point differentials:
  -
    $$\text{diff\_time\_ratio} = \text{point\_differential} \times e^{4 \times \frac{3600 - \text{game\_seconds\_remaining}}{3600}}$$
- Down
- Yards to Go
- Timeouts remaining for each team
- Which team will receive 2nd Half Kickoff
- Which team is at home
- Spread time:

$$\text{posteam\_spread} \times e^{-4 \times \frac{3600 - \text{game\_seconds\_remaining}}{3600}}$$

Note: the spread time is only included in a model which takes into account the pre-game Vegas betting line on the game. This allows the win probability model to adjust the weights properly to account for the possibility that one team is better/worse than the other.

We make the assumption that our coach does not have a bias towards any of the particular strategies. Making this assumption means that we believe the coach is risk-neutral which obviously may not always be the case. As acknowledged in the introduction, coaches may lean towards a particular strategy. In an extension of our model, we can add a risk profile on various coaches and model a more accurate bias towards certain strategies. But for simplicity, we will keep it risk-neutral.

Thus, with these probabilities, we can now model a pure strategy, maximizing the winning probability of the offensive team. If we wanted to be more nuanced in our approach,

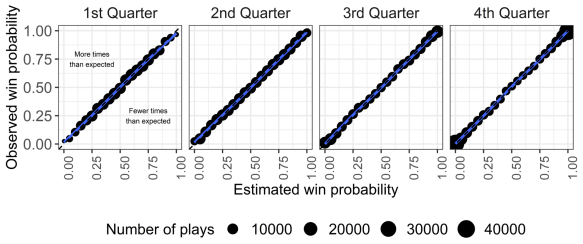

Figure 4: NFL FastR Win Probability Model Results

we can normalize the probabilities, that is if two choices are close in probability of winning, the offensive team would perform a close to a coin flip to determine the strategy they would choose.

But since we assume that the coach does have knowledge to this model and the xgBoost model is accurate, we can assume a pure strategy of simply picking the largest probability. That is:

$$\arg\max_{strategy} \{PuntW\%, GoForItW\%, FieldGoalW\%\}$$

In addition to choosing the strategy, there is also a chance when the offense can fake or not fake. We will at first model this as 1% faking and 99% not faking due to historical data. And in this case, we would later perhaps expand our model to try to capture a better mixed strategy in order to determine more optimal times to fake vs not fake.

Defensive Mixed Strategy

We next need to estimate the mixed strategy of the defense in our model. Compared to the offensive strategies, the defense is much more reactionary in the formation and the particular planning which they can perform.

This estimate is harder to determine because currently in only 1% of all punts or field goals the offense attempted fakes. Thus, the defense must have some mixed strategy when defending field goals or punts to account for the possibility of the fake.

Understanding that, we can use this same historical fact so that 99% would expect that the fake occurs 1% of the time. Similar to the offensive strategy, it is important to note that based on the environment, the mixed strategy of the defending team can be different and this model currently does not take account for that.

We found since 2013, 21 fake field goals or punts have occurred. Surprisingly, of the 21, 13 have been successful. In

this context, success is defined as "Binary indicator whether $epa > 0$ in the given play". If EPA is positive this means, the 1st down or touchdown would have been converted, else the EPA would have been negative.

Payoffs

Next, we need to estimate the payoffs or the utilities that are observed by either team during each sequence given the state of the game. The most basic way to think about payoffs is to have most of the utilities being offensive favored. This is because when we determine the offense's strategy, we are maximizing the win probability and thus pure strategy to maximize the offense's utility. Thus, when actions go as expected (without fakes), $U_O = 1, U_D = 0.5$. And then when there are fakes, $U_O = 0.75, U_D = 0.75$. These numbers representative of our assumptions that have been made when choosing the strategies.

We plan to refine the payoff numbers by thinking more about the game state and the most beneficial components and perhaps some way to model these payoffs to be more accurate towards the game state similar to the fake strategies of the offense and defense.

With all the components, we can then model the game similar to what figure 5 looks like.

Implementation

Tuning

One important part of machine learning Models is hyperparameter tuning. We plan to use regularization and choosing our hyper-parameters by using 5-fold cross validation. We are still working on creating the two models and are using the nflfastR data in order to accomplish this.

Stackelberg Game Equilibrium

To determine the optimal choices for offense and defense in the fourth down decision, we will calculate the Stackelberg Game Equilibrium. We plan to perform backwards induction on the chances of faking based on the defender's mixed strategy. In any given strategy, we know that the defense has the choice of expecting the actual play or expecting a fake.

Results/Evaluation

We are still working on building out the model that calculates the win percentages. For now we are relying on the already built one by Open Source Football.

We plan on evaluating our model based on accuracy, F1 score, and AUC-ROC curve on the xgBoost with a testing and training split on the data. These evaluation metrics would be useful in testing the predictive ability of the model. The model accuracy would be interesting to analyze as there are many factors in choosing the training and testing sets as mentioned when determining our model, the rarity of the presence of the event is important.

Additionally, the model needs to take account unacknowledged factors. This includes the mentioned risk profiles of a particular coach and team. Additionally, the strengths and weaknesses of particular teams and their profiles also should be taken account when making a decision. Our model attempts to uniformly view NFL teams in order to produce higher amounts of data, but if our model sees weak results, it might be necessary to take into account particular players, kickers, and the team which is playing. This may require more work in trying to model the exact strength of a team in order to input as a parameter in the model.

The results of the modeled game equilibrium would provide insight on whether or not teams should fake or not given a situation. The goal is to provide better insight in the theory and strategy in American Football.

Our final model, which given a specific game state, gives a recommendation of kicking, punting, or attempting a field goal, is shown in the results in Figure 6. Holding the game state constant we modelled what our model's recommendation would be for each combination of Yardline100, the number of yards to go to get to the endzone, and Yards to go, the number of yards to achieve a first down. The example given is in a tie game, start of the 2nd quarter, the defending team was favored by 3 points in the pregame betting odds, both teams have 3 timeouts remaining. It is important to note we used a version of the win probability calculator that takes into account the pregame betting odds, as a heuristic to which team is inherently more likely to win. The results from our model are very aggressive. The recommended choice is to never punt when there are less than 5 yards to go.

Our model may benefit from further refinement as we think this has gone a little too far in the opposite direction. While the typical NFL coach is in our opinion (and the model's opinion) too conservative, this model is probably too aggressive. If an NFL team decided to adopt the strategy derived from our model, it would make for an interesting experiment in the revolutionizing of the game of football strategy, but we are likely many years from strategy anything close to this.

Related Works

The study of using predictive models to evaluate the game theory behind football is a relatively recent discipline. Jordan, Melouk, and Perry wrote about "Optimizing Football Game Play Calling" in 2009. In this paper, they analyze how the imperfect information present in the game leads to imperfect decision making. They go on to explore the process of modeling football games as a 2-player zero sum game using a game theoretic system.

As previously mentioned, the theory of EPA or expected points added and WPA or win probability added are the basis of our study. Expected points added are generated by making plays that exceed the expectation for the average football team, and WPA is similar. Coaches are not the ones out on the field actually making the plays, but they are an integral part to a football team. Designing and calling plays that gives their team the maximum chance to succeed is an enormous part of their job. Exceeding expectations literally generates EPA and WPA for a team. The idea Yurko, Ventura, and Horowitz came up with allow for the proper analyses of football games. The statistics can make it easier to discern whether the correct decisions were made or not in a game.

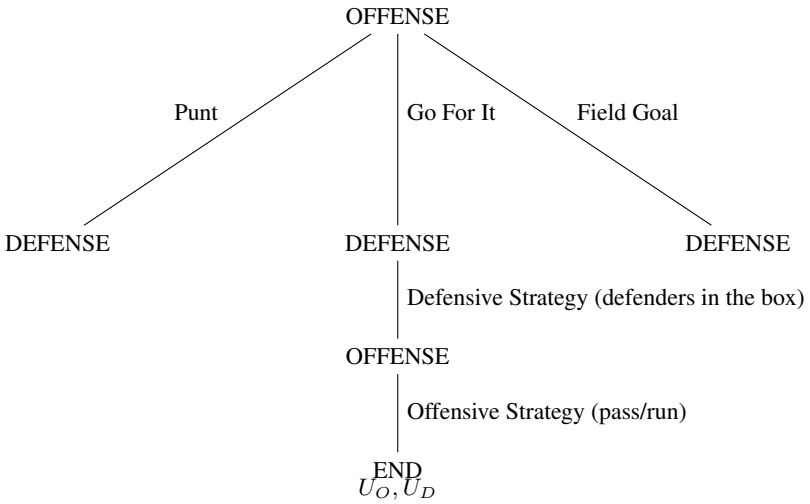

Figure 5: Simplified Stackelberg Game Model for Fourth Down Decisions

Another study looked at the fourth down problem through pure simulations using nflsimulator which is a publicly available simulator that maps out NFL games based on given parameters. In "Fourth Down Decision Making: Challenging the Conservative Nature of NFL Coaches" by Palmquist at the University of Denver, ran simulations and evaluated its use of the results during the fourth down. Through simple evaluation, he found that going for it was much better than the defensive strategies. And that coaches tended to make sub optimal decisions by not going for it most of the time. One key insight that was made is that overall data has more punts/field goals are much more prominent than going for it in historical data which may cause some uncertainty in the data.

In a 2019 paper named "What was lost? A causal estimate of fourth down behavior in the National Football League" by Yam and Lopez", found that teams that chose to go for it on fourth down generally saw a mean increase in win probability by about 1.9% compared to teams that did not. They also found that the decision to go for it introduces greater variability in win probability outcomes, which is consistent with the perception that coaches who make this choice accept higher risks. In the long term, a bootstrap analysis estimated the long-term effect of always going for it within recommended scenarios, suggesting most teams would have increased their total wins from 2004 to 2016. This shows the additional evidence that being more aggressive tends to favor the offensive team.

Looking more into the risk aversion of coaches, a 2011 study called "Are NFL Coaches Risk and Loss Averse? Evidence from Their Use of Kickoff Strategies" explored further in this passage is whether NFL coaches' conservative play-calling is due to them not optimizing decisions perfectly or because they are inherently risk-averse. The findings suggest that coaching decisions might be influenced by a combination of non-neutral risk preferences and imperfect optimization. Thus that it may not just be the risk profile that is affecting the coach's decision but also perhaps actual suboptimality that could be improved.

Another paper authored in 2024, called "Analytics, have some humility: a statistical view of fourth-down decision making" by Brill et. al. showcased that analysts are advised to exhibit humility and recognize the limitations of the data when making conclusive statements about the efficacy of fourth-down decisions. Small gains in win probability, even those under 1%, can accumulate significant advantages over a season, suggesting that these opportunities should not be overlooked. The paper focuses on the limitations that the dataset may have. Win probability estimates are subject to significant uncertainty due to the limited data available from football history. This uncertainty can make some decisions appear more beneficial than they might actually be, suggesting caution in using these estimates to make definitive recommendations. Thus, the study recommends fourth-down decisions only when there is high confidence that a particular choice has a better win probability than alternatives. If there is significant uncertainty in the recommendation, the decision should be left to the coach's discretion, recognizing the limitations of data-driven models and the on-field expertise of the coach. The paper mainly calls for a more balanced approach when evaluating and making recommendations for games. The discussion introduces probabilistic state-space models as a potential advancement over traditional statistical models for estimating win probability. These models calculate transition probabilities between game states from play-level data and simulate games to estimate outcomes. This is similar to what we try to model in our transition and state space inside of our model.

Looking more into more data implications, "Bigger data, better questions, and a return to fourth down behavior: an introduction to a special issue on tracking data in the National football League" by Lopez talks about how certain data points are not captured which may deter the results of the bot. The data does not capture certain aspects like the z-coordinate (height) or the exact positioning of helmets, arms, and legs, which limits the depth of analysis possible

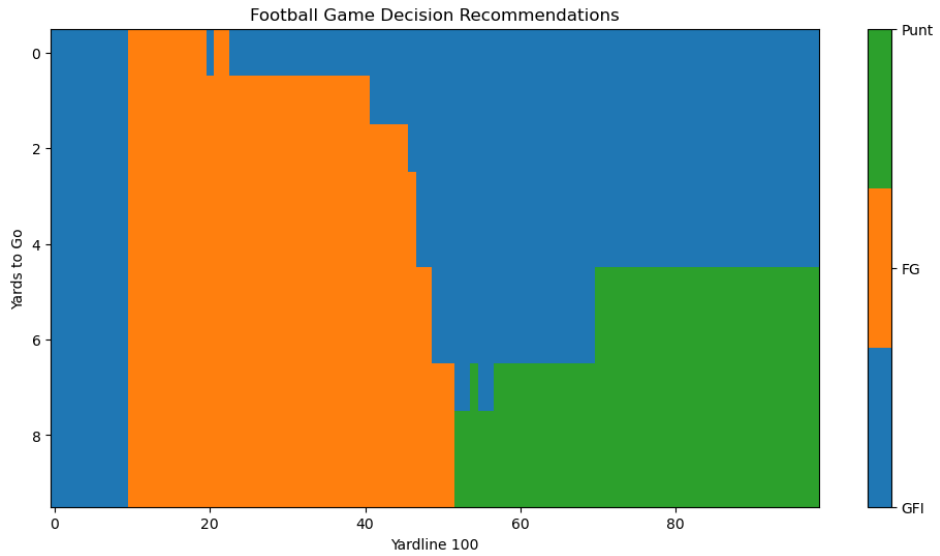

Figure 6: 4th down decision model given field position and yards to go

from the tracking data alone. The study found that sometimes the bot overestimated certain aspects of the play by this missing distance variable which is crucial in better understanding.

In a more game theoretical sense, we also looked at "Do Firms Maximize? Evidence from Professional Football" by David Romer which analyzed using dynamic programming to model the utilities of actions during the fourth down. The author similarly models the payoffs of the game as the value of situation g,t as of that situation must equal the expectation of the situation's realized value one situation later. The model framework which the paper is based on is reflective of our goals also by projecting and finding the future probability and the expected values of certain actions. Again also though, this paper finds that teams are more often more conservative than usual in making the fourth down play. The paper finds generalizations such as: On their own side of the field, teams should consider going for it if they have less than 4 yards to go, which gives practical rules of thumb for readers/coaches.

Conclusion, Limitations, and Future Work

As studied in related works, we recognize that there may exist confounding variables or bias inside of the results that we observe. For example, historically, coaches would tend to make punt/field goal calls and those results are much more prominent than going for it, which may mean that coaches only go for it when it is highly guaranteed, causing some skewness towards aggressive plays potentially.

Another limitation with our model is that we do not relay a good confidence interval on how strong our model prediction is. As laid out in other articles, solely relying on analytics without good interpretability is not an effective way of performing coaching.

Future work would be doing more work in understanding the underlying data more and seeing how that may be impacting our model. For instance, modeling risk aversion from coaches that may impact the data.

Appendix

The github repository with all of our code is here: `http://www.github.com/ilanbarr/555T-NFL-project`