15.095: Machine Learning Final Project

Optimal Play Calling in the National Football League

Massachusetts Institute of Technology
MIT Sloan and Operations Research Center

Jonathan Chan & Joe Zaghrini

Table of Contents

List of Tables

List of Figures

Executive Summary:

Football is a sport that is a prime candidate to use optimization and machine learning to prescribe the optimal offensive play call. Using play by play data between 2009 and 2018, models were built to prescribe the optimal play call using Interpretable AI software. Optimal Classification Trees were used to determine the expect outcome of a drive based on down, distance and yard line. These play-by-play valuations were leveraged by a set of Optimal Prescriptive Trees to prescribe the optimal play for various scenarios related to realistic in-game shifts in strategy. These Optimal Prescriptive Trees were extended to determine the resulting play calling depending on the quality of a team's quarterback.

The results are play calling models that are interpretable and practable. They confirm a lot of the play calling intuition that exists in the game today such as field goals from the 37 yard line, running the ball for short yards to go and yard line, passing the ball for long yards to go, and yard line. In addition, with a good quarterback, teams should have a balanced offense of running and passing, while with a bad quarterback, teams should run or kick a field goal. Possible extensions for future model improvement are also identified in this report.

Problem Description:

Football is a complex sport with many facets to consider in decision making and is thus a prime candidate for using optimization and machine learning. In the National Football League (NFL), each team calls a play (i.e. strategy) for their players on each down. Fundamentally, coaches must decide whether to run, pass, or kick. Poor decisions can be extremely costly: two recent Super Bowls have been lost due to questionable late-game play calling. In addition, play calls are incredibly complex decisions from an "input" standpoint: coaches have many different sources of information to consider when selecting a play. The complexity and time-sensitivity (40 second play clock) decisions result in most decisions being made on "intuition," which is subject to human biases. Coaches also employ a greedy as opposed to global decision-making, choosing the best choice for a play, and not necessarily for the drive or the game. The purpose of this project is to explore the possibility of using machine learning to develop decision-making tools rooted in the vast data available to NFL teams.

Football Overview:

Football offers a certain number of points for each end-of-drive result: the offensive team can score a *touchdown*, which is worth 6 points, or a *field goal*, worth three. After a touchdown, a team may choose to kick an *Extra Point Attempt* (like a short field goal), worth a single point, or attempt a *Two-Point Conversion* (a short, one-down touchdown attempt), which yields two points. In the vast majority of cases, a team will successfully kick an Extra Point Attempt, so touchdowns most often result in 7 points scored by the offensive team.

In addition to these ways the offense can score, the defensive team can score in two ways. The first is a *safety*, where the ball is "down" in the offensive team's own endzone. This is worth two points for the defense, or -2 for the offense. Another way to score is to intercept a pass and run for a touchdown, colloquially referred to as a "*Pick Six.*" Since this is effectively another touchdown, this can be considered to be -7 points.

A *drive* begins when the offense starts with the ball, and ends with a Touchdown, Field Goal, Punt, or a Turnover. The start of each drive is a 1st Down. The offensive team then has four attempts (downs) to gain 10 yards. If they gain 10 yards within these attempts, they get another 1st Down. This continues until they fail to get to a 1st Down, kick the ball (on 4th Down), or score a touchdown. Note that the result of a drive can have no score, which is the vast majority of drive outcomes in a game.

Data:

To solve this problem, data was used from two sources. The first is a Kaggle dataset with clean play-by-play data in games from 2009 to 2018. The dataset has over 250 features for 450k NFL plays. This dataset was clean and there was considerable effort required to understand which features are important predictors. In addition, a second dataset from sportsreference.com was used to determine which quarterbacks were the best statistically in each year in order to filter all the plays of the good quarterbacks and bad quarterbacks to run the model.

Approach & Methodology:

Simplifications & Justifications:

To make the analysis simpler, more tractable, and feasible within the given time frame, we made several assumptions. These assumptions, along with some justification, are as follows:

- Extra point conversions – After scoring a touchdown (6 points), the scoring team can choose tween an Extra Point Attempt (one more point) or a Two-Point Conversion (two more points). For this model, all touchdowns were converted to be 7 points, as the majority of cases include a successful Extra Point Attempt.

- Plays within 4 minutes of the end of a half – These plays were filtered out as they require a different strategy than the rest of the game. When a team is winning, the traditional strategy is to run the ball in order to run the clock and when a team is down, the traditional strategy is throw the ball and go out of bounds or call a timeout to stop the clock. While these intricacies could be a future expansion of the model, this project will simplify the input data, instead capturing these strategic choices with a separate cost function.
- Plays with "No Play" & unique plays – "No Play" occurs frequently in the dataset for plays that resulted in a penalty or coach's challenge. These were removed since they cannot be predicted by a model and the strategy of a play should not be to gain yards via a penalty. In addition, unique plays such as QB Kneel, QB Spike, FG or Punt before fourth down were removed as they do not capture the regular situations in a game.
- Offensive plays – To simplify the problem, this project will only consider offensive play calling. Defensive play calling is more difficult with the dataset available, but is an opportunity for future enhancements.

Model Concept Development:

For this problem, prescription of the ideal play is the objective, so an Optimal Prescriptive Tree was used. Play calls must be made in the time allowed by a play clock (40 seconds), so the model must be quickly interpretable both for a coach to call the best play in the game, and to defend the call if the result is a poor outcome.

In addition, football is a multi-staged decision problem where the decisions and outcomes of each down affect the next downs. Thus, it is important to properly assign a value to a certain point in the game based on down, distance, yard line with the goal of maximizing the expected score with each play. For this, an Optimal Classification Tree was used to predict the expected value of a drive based on a play within the drive.

To generate the models, the following features from the original dataset were used: `game_id`, `yardline_100`, `down`, `ydstogo`, `goal_to_go`, `drive`, `passer_name`, `year`, `play_type`, `quarterback_rating` and `passer_player_name`.

Note: The objective of offensive play calling is to maximize the outcome score. The models are meant for minimizing objectives, so the scores were converted to negatives (ie. touchdown = -7 points, field goal = -3 points, etc.)

Optimal Classification Tree (OCT) Model:

An important observation of scoring in football is that the scores (presented earlier) are outcomes of *drives*. However, our decision is not on a drive-by-drive basis. Instead, each call (and decision) is for a single *play* within the drive. While the outcomes in the Football Overview describe the possible outcomes of a drive, the outcome of a play is most often another, slightly different play. For example, a "3rd and 8" play 75 yards from the Endzone may turn into a "1st and 10" play 65 yards away from the Endzone. To better model the decision, we needed to capture the value of each individual play.

Our approach was to treat the later stage decisions and outcomes, after the current play call, as an uncertainty. This is a simplifying assumption, and one that does not exactly fit the decision structure at hand, but this assumption preserved the feasibility of our problem without sacrificing on the structure. A relaxation of this assumption is explored later as an extension to the project.

The purpose of the OCT was to create a discrete set of possible outcomes that could result from a play and calculate an expected cost for each of these results. These expected values correspond to intermediate plays that do not terminate a drive. The inputs were the data of the current play, and the predicted outcomes are the results of the drive. By following the splits, a set of possible outcomes was developed, in which each play could be categorized. The resulting OCT can be found in the appendix. With the probabilities of each drive outcome on a given leaf, an expected value can be calculated based on any arbitrary cost function.

Optimal Prescriptive Tree (OPT) Models:

Cost Functions Motivation:

While the different drive outcomes all have numerical values attached to them, they do not capture the kinds of strategy changes that may come in a real game. More specifically, these scores merely capture what is on the scoreboard. In reality, teams want to maximize their probability of winning, not necessarily their point differential. Therefore, coaches may choose to take riskier plays late in the game when the trail by only a few points or may take safer options with lower "expected points" values when leading in a game. Making strategic mistakes risks costing a leading team the game or squander a trailing team's opportunity to make a comeback win.

Cost Functions Summary:

To see how an OPT can model these kinds of strategic adjustments, we developed several cost functions. In each cost function, the value of each outcome is slightly different. Table 1 summarizes these values:

Scoring Function	Touchdown	Field Goal	No Score	Safety	Pick Six
Base	7	3	0	-2	-7
Down by Four	7	0	0	0	0
Any Score	1	1	0	0	0
Down by Three	0.9	0.5	0	0	0

Table 1: Cost Function Importance

- In the *Base Case*, a team simply values each drive as the change in point differential, from the offensive team's perspective.
- In the *Down by Four* case, the offensive team is trailing the other team by four points late in the game. In these situations, a field goal is effectively worthless; the maximum they can score is 3 points. Likewise, a safety or a pick six are unimportant, as they don't change the outcome of the game. The only way to change the outcome of the game is to score a touchdown, resulting in a win.
- In the *Any Score* case, a team is trailing by one or two points. In this case, a touchdown or a field goal will win them the game. The team's goal, then is just to get *a* score (any score).
- Finally, in the *Down by Three* case, the offensive team trails the game by three points. A safety or pick six are similarly inconsequential to the team as in previous cases. However, a

touchdown and a field goal are scored differently. The field goal would tie the game, forcing overtime, while the touchdown would effectively give the team a lead in the final moments of the game. This scoring function was developed with the assumption that, in overtime, each team has an even chance of winning, while a touchdown raises the offensive team's chance to 90%.

Good vs. Bad Quarterbacks (QBs):

An additional analysis of the models is the effect of a quarterback's historical performance. The quarterback is the most important player in the game, and one would expect the play calling to shift based on the quality of the quarterback. To do build these models, the dataset of QB performance per year was gathered from sportsreference.com and the top 10 quarterbacks and bottom 10 quarterbacks were selected each year from 2014 to 2018 based on Quarterback Rating (QBR). While this may not be the most representative statistic of the best quarterback, it was clear from our understanding of football that the split was uncontroversial. In addition, we filtered out any quarterbacks with less than 100 pass attempts to filter for one off throws or plays, or where a quarterback did not have enough data points. This resulted in 27,000 and 12,000 rows of data for each group, respectively.

Results:

Each of the varying cost functions should see some different behavior. While there is no easy way to test the validity, they can be checked against conventional play-calling wisdom.

Full Dataset: "Base" Scoring Function

The "Base" scoring function, where a team seeks to maximize their expected points scored, sees some intuitive results. Choosing to run the ball tends to be a less varied play: they pretty reliably gain a few yards. This means a team should choose to run when they are close to their target (like a first down). A pass, on the other hand, has a lower chance of success. They often result in incompletions, where the pass is not caught. They also, however, have a longer reach. The expectation, then, is that a pass would be suggested in situations where there are relatively low stakes (like a first or second down) but a large gain is required. Lastly, a Field Goal is effectively only chosen on a fourth down.

The model confirms the conventional wisdom for passing and running. The main deviations are with regards to field goals. The model often suggests kicking a field goal before fourth down, which is not something teams regularly do. However, this is likely due to the `down` data being a continuous variable and not a categorical one. That the model *never* recommends kicking on first or second down is an encouraging result.

Altogether, the "Base" model does a very good job of approximating (and therefore confirming) much of the mainstream thinking in football.

Full Dataset: "Down Four Late" Scoring Function

The next tree, for the "Down Four Late" scenario, is *very* different. In this case, only a touchdown is worth any value to a team. The low-hanging fruit here is that the model should never say to kick a field goal; scoring 3 points when trailing by 4 at the end of the game just guarantees a loss by 1. But beyond that, the team's strategy completely changes, and the tree calls for a lot more runs. In fact, if the play is on third or fourth down, the model calls for a run regardless of the distance, even though one would expect to see more passing.

This particular strategy may be less realistic in practice. In real game situations, time pressure may necessitate a team throw the ball. By passing, they can advance more quickly. In addition, there are some tricky timing rules related to when the clock stops that incentivize teams to throw the ball.

Perhaps an interesting nuance of this model is the effect of an incomplete pass. Should a throw not be completed, the offense does not advance at all. In most cases, this isn't a big deal; a team can get close enough and kick a field goal. But in this case, where a field goal is worthless, the model opts for slow-and-steady strategy, seeking to gain a few yards each play until a touchdown can be scored.

Full Dataset: "Any Score" Scoring Function

In the "Any Score" case, a team is just looking to score (eg. already in the lead). The tree should prioritize a high *probability* of scoring, while ignoring the magnitude. Therefore, we would expect to see a predominance of field goals. This is exactly captured in the model. Perhaps, however, the model is too optimistic; it does not consider the distance away from the goal when choosing to kick a field goal. This is likely just an issue with the data; there is no data where a field goal was kicked from 75 yards, because it does not happen. In future versions, a way to fix this would perhaps to label "punts" and "field goals" as the same action (a "kick").

In the few cases where the model does prescribe a run or pass, it is always first or second down. The model is capturing the lack of risk associated with having multiple opportunities to get a first down, and that the chance of scoring a field goal increases as you get closer to the Endzone. This tree captures the main element of conventional wisdom, but not all of its nuances.

Full Dataset: "Down Three Late" Scoring Function

In the final case, "Down Three Late", the score is a weighted average of the probability of a field goal and the probability of a touchdown. There is no clear or well-defined global strategy. Coaches will, like in many difficult decision problems, take a greedy approach, seeking to maximize an immediate goal (like yards gained). This is a case where OPT does not have a good baseline comparison, in large part because our model seeks to couple those decisions and take a less greedy approach.

A simple case to test against the baseline is when the "greedy" goal is the same as the "global" one: at the end of the drive. In situations where a team is on third or fourth down and far from the endzone, conventional wisdom would suggest a touchdown is very unlikely. The model captures this, opting to kick a field goal. This does, however, depend on how far a first down is. In a similar case, where a first down is close by, the model prefers to pass and seek a first down (and chance for a touchdown). This model also seems to capture the risk-reward relationship of passing the ball, prescribing it more often on first down than on 3^{rd} or 4^{th}. Altogether, the results of this model aren't egregiously counterintuitive, and the model is a promising way to make global decisions instead of greedy ones.

Subset: Quarterback Analysis

In the quarterback analysis, the trees prescribe very different strategies, as one would expect given the QB's centrality to offensive strategy. For a bad quarterback, the model suggests throwing the ball only on first down and running the ball in every other situation. For a good quarterback, there is a more even prescribed distribution of passing and running. This matches intuition, as a pass on first down allows the offense to still get advance the drive with runs.

The result holds for other scoring functions. For example, in the “Down Four Late” case, the good quarterback subset resulted in a tree that prescribes a very balanced strategy with plenty of passing and running, even on later downs. With a bad quarterback, passes only occur on first down.

Figures for each of these cost functions, using the full dataset or the Good/Bad QB subsets, can be found in the appendix.

Extension Opportunities:

While we are overall pleased with the results of this project, it has raised several opportunities for improvement. A clear place to begin is to improve the accuracy of the OCT. The current OCT has a misclassification error of 0.65. Though this is good enough for a first iteration of a model, we believe that there is room for significant improvement. This can be accomplished with more feature engineering (some examples below) and varying the depth of the tree to value field position more precisely.

Player Performance

The current model does not consider in-game or in-season player performance. This is an area that can greatly improve the performance of the model. By accounting for a players year-to-date and in-game performance, the model could capture the effects of having a stronger or weaker player at a given position, which also matches how coaches think. This feature engineering would be performed for four positions: the Quarterback, Running Back, Wide Receivers, and Punters/Kickers. While the QB statistics were accounted for at a high level by performing subset selection on QBR, more can be done to understand the position better, especially given how vital the Quarterback is to a team’s offense. Since the running back is usually the person who runs in running plays, getting their statistics is a natural improvement to the model, especially since the decision can be made to run using a specific running back. Receivers are important because they are the second part of every pass play; a Quarterback throws the ball to a receiver. This would still be a difficult phenomenon to quantitatively capture. Good receivers tend to draw more defensive attention, which may result in the quarterback throwing the ball to that receiver less often. Developing a holistic statistic on receiver ability is therefore a very non-trivial problem. Finally, Punter/Kicker performance would capture their success rates, which are important for deciding when to kick field goals.

Modern Passing Game

Another limitation of the model is a stylistic shift in teams’ style of play. In the last five years, teams have moved from a “run-first” offense to a “pass-first” offense, mainly due to increased performance of quarterbacks and rules changes favoring passing. An opportunity exists to build models focused on modern NFL play calling rather than the data set from the last 10 years.

Time Control

As currently formulated, the model also does not consider time control. In late-game scenarios, the coach may choose to take certain actions to extend or shorten the end of the game (depending on whether they are leading). Examples of this are QB kneels, taking time-outs, and favoring incomplete passes and out-of-bounds receptions over getting down on the field, where the clock continues running. The current models, because they do not consider time control, suggest some counter-intuitive late-game strategies. This extension would require adjusting the cost functions for time and score in the game.

Defensive Strategy & Kicking

This model also ignores the performance of the opposing defense. Teams that are very good at throwing the ball often must run more when playing teams with strong secondary (passing) defense, and vice versa. Pass and run defense metrics ought to be used to prescribe an offense play call to exploit defensive weaknesses.

The current models never prescribe a punt because the expected value of a punt is almost always zero. The expected value of any other play will always be slightly better from an offensive perspective. In reality, punt can be treated as a defensive play, where the goal is to prevent the other team scoring by starting with advantageous field positioning for the defense. Since the expected value of field positioning from a defensive standpoint was never determined, the models never prescribe a punt. For a more robust and accurate model, accounting for defensive strategy could also improve the model's prescription of punts.

Maximizing Scoring Probability vs. Winning Probability

Altogether, an ambitious shift in the model's framing may be necessary. As currently formulated, the model seeks to maximize the score at the end of the drive, according to some engineered scoring functions. These are, at their core, attempting to approximate a decision where a team maximizes their chance of winning. A model that takes the probability of winning would directly capture this dynamic. This would make some of our other improvements, like including punting and time control, fit in more naturally into the structure of the model. It would also require that these other decisions be included in the data, in enough samples to be representative. This would create an additional burden on finding appropriate data.

Dynamic Programming Formulation

An algorithmic extension involves dynamic programming. As discussed in previous sections, a simplifying assumption was made to treat the outcome of a set of future decisions as an uncertainty to be predicted (using OCTs). In the real world, the results of future plays both are dependent on the current play and should influence the play. This is an opportunity for a global optimization problem, through which a series of sequential decisions are made. This would function as a dynamic programming problem.

The process, in an informal explanation, is as follows. A set of "terminal" outcomes are defined. This set would begin with outcomes that end a drive, like turnovers, Touchdowns, and Field goals. Then, a set of "penultimate" plays would be taken from the data. These are plays in which the entire set of possible outcomes (from data) is a subset of the termination outcomes. An example of a penultimate play would be Fourth and Goal, where a team either kicks a field goal, scores a touchdown, or turns over the ball. In each iteration, an OPT would be fit to the discrete penultimate plays, based on the value of their outcomes (which, by construction, are all terminal). The prescriptive cost (the first term in the OPT's objective) would be set as that play's value, and the iteration's terminal plays would be added to the set of termination outcomes. This process would repeat until every possible play in the game joins the termination set.

This has a couple of challenges, especially when compared to our analysis. In our analysis, we used the splits of an OCT to algorithmically determine what our intermediate outcomes would be. In the dynamic programming version, however, these intermediate outcomes would have to be predetermined. This could still be performed using another method, like an OCT, as a sort of initialization heuristic to define these discrete states. By using several of these heuristics, the modeler could formulate several DP problems and solve all of them, choosing the one with the best results (perhaps on which one best leverages the dataset). This would, of course, be quite intensive computationally. Finally, the choice of prescriptive cost is a non-trivial one. On one hand, considering only the first term in the OPT would focus on the decision made on the expected outcomes. On the other hand, including both the prescriptive and predictive terms of the objective in the play's cost would penalize plays that involve more uncertainty.

Conclusion

This project used a large dataset and Optimal Trees to inform decision-making in football. It incorporates all three components of analytics we have been taught thus far. It begins with data, uses OCT's to model the outcomes of plays in an NFL game, and uses OPT's to inform decision-making. Beyond this structure, it showcases and leverages several degrees of freedom in each model, demonstrating these methods' abilities to be engineered to solve complex decision problems. The final developments of this model effectively capture several aspects of play-calling principles, and extensions promise to better approximate (and even surpass) the decision-making skills of professional coaches. Altogether, the project demonstrates the power and flexibility of analytical methods to solve complicated decisions where the data is available.

Contributions:

Short Description	Long Description	Contributor
Paper	Writing all sections of the paper	Jonathan Chan / Joe Zaghrini
Poster	Assembling the poster materials	Jonathan Chan / Joe Zaghrini
Feature Engineering: Drive Outcomes	Determining the outcome of each drive as a summary for all the plays in that drive	Jonathan Chan
Feature Engineering: Quarterback Grouping	Determine the top and bottom quarterbacks for each year to be used in OPT	Jonathan Chan
OPT Modeling: Quarterback	Developed OPT models and insights for whether the quarterbacks were good or bad	Jonathan Chan
Feature Engineering: Cost Functions	Developed different cost functions to simulate scenarios experienced in play calling	Joe Zaghrini
OPT Modeling: Cost Functions	Developed OPT models and insights for the different cost functions	Joe Zaghrini
Modelling: OCT Prediction	Developed prediction of each play and their expected cost for a point in play	Jonathan Chan / Joe Zaghrini

References:

"Detailed NFL Play-by-Play Data 2009-2018", Max Horowitz, Kaggle.
https://www.kaggle.com/maxhorowitz/nflplaybyplay2009to2016
"NFL Passing", Pro Reference Football.
https://www.pro-football-reference.com/years/2014/passing.htm

Code:

Code for the project can be found on our GitHub:
https://github.com/jonathanchan1994/machinelearningproject

Appendix:

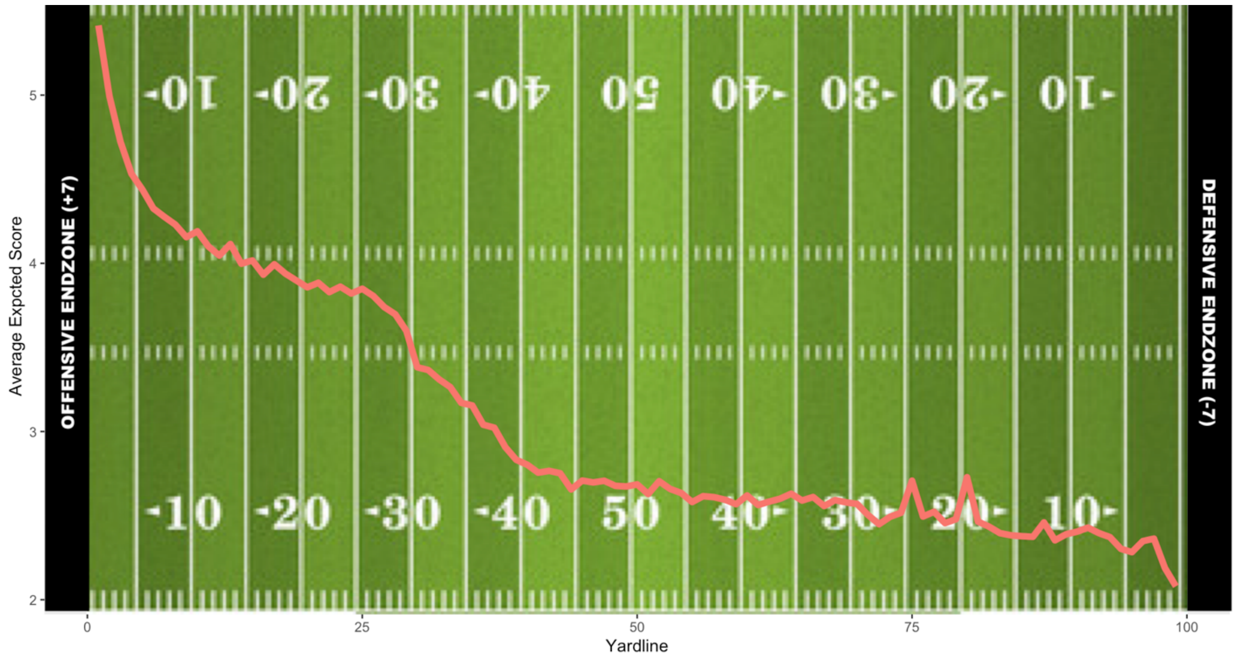

Figure 1: Expected Cost vs. Yard Line from OCT

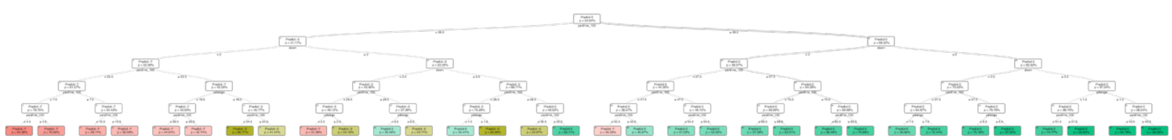

Figure 2: OCT for play outcome valuation

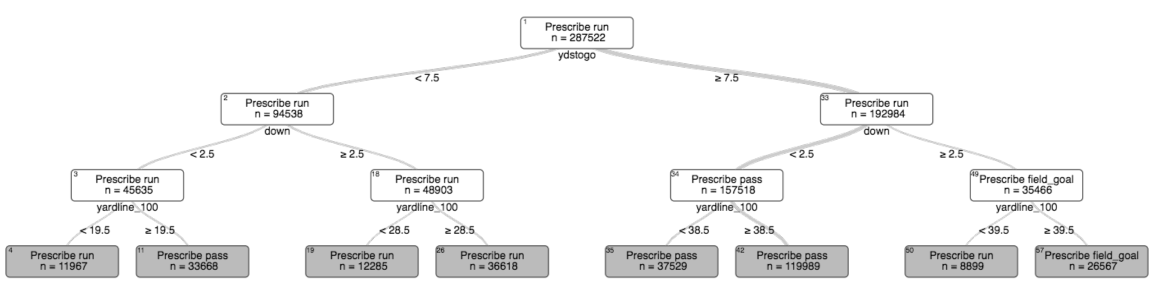

Figure 3: OPT for Base Case

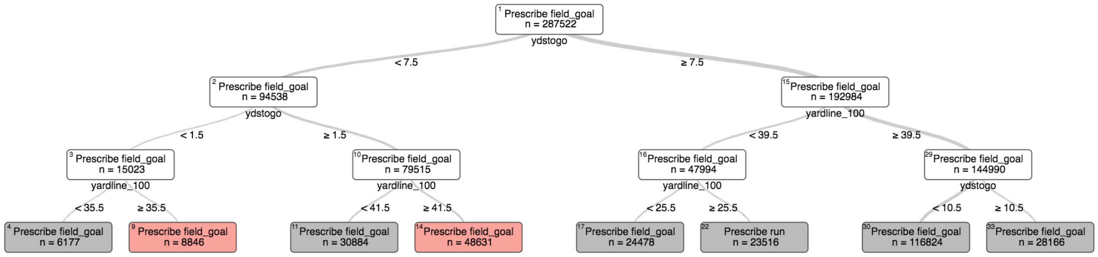

Figure 4: OPT for Any Score

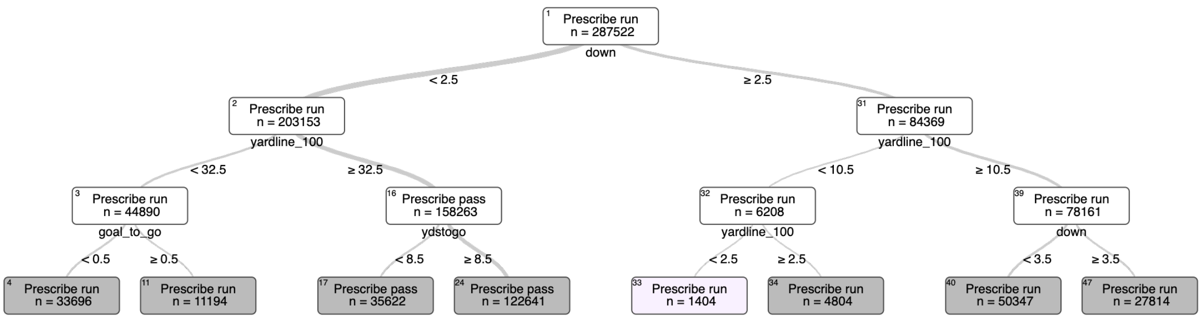

Figure 5: OPT for Down Four Late

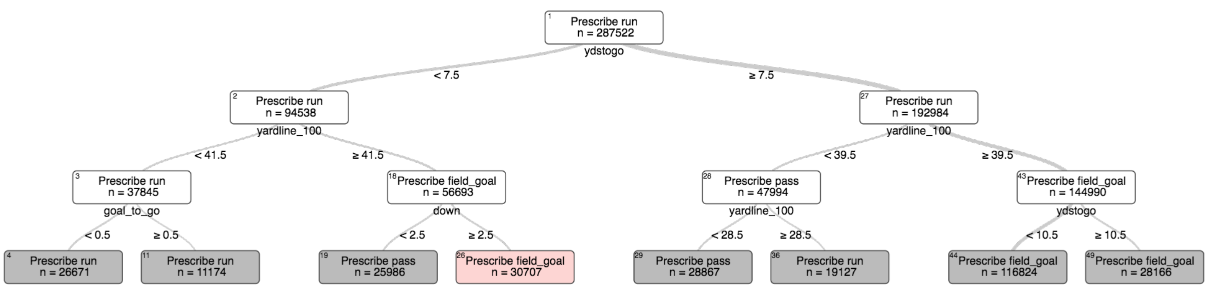

Figure 6: OPT for Down Three Late

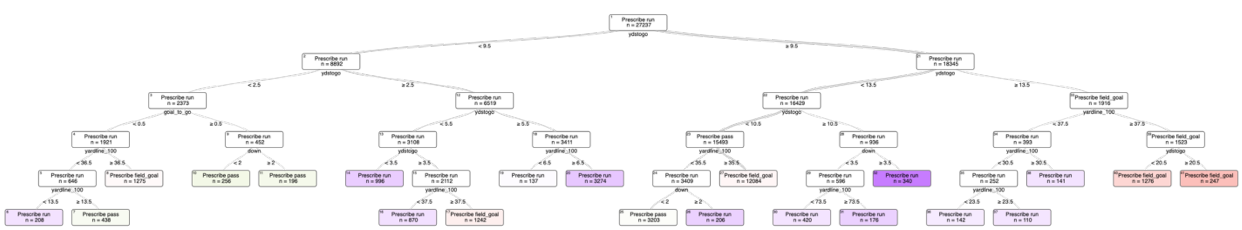

Figure 7: OPT for Good QB - Drive Outcome (Naive)

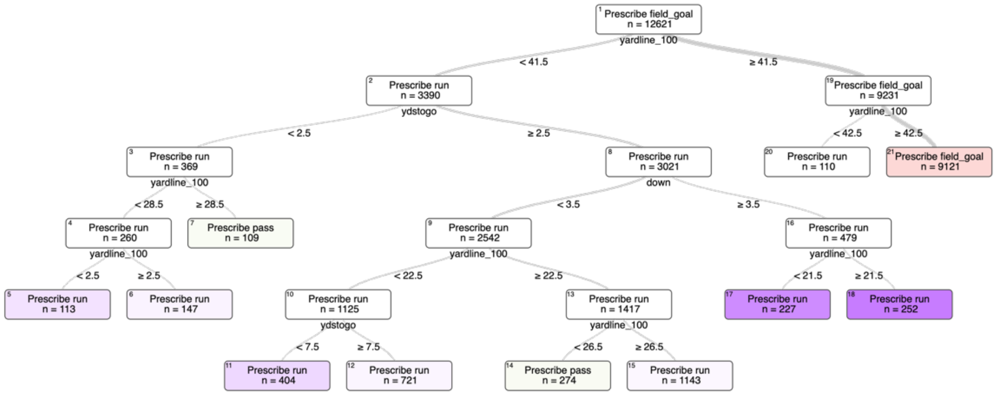

Figure 8: OPT for Bad QB - Drive Outcome (Naïve)

Figure 9: OPT for Good QB - Base Case

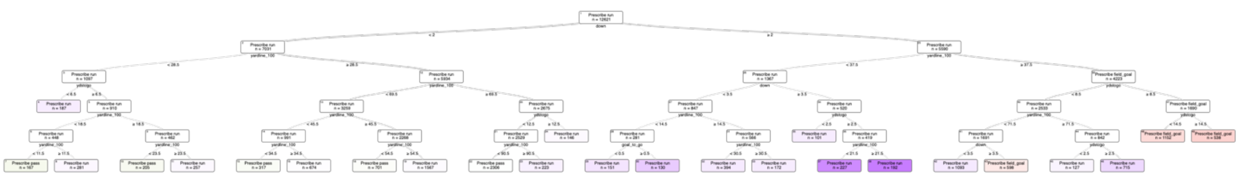

Figure 10: OPT for Bad QB - Base Case

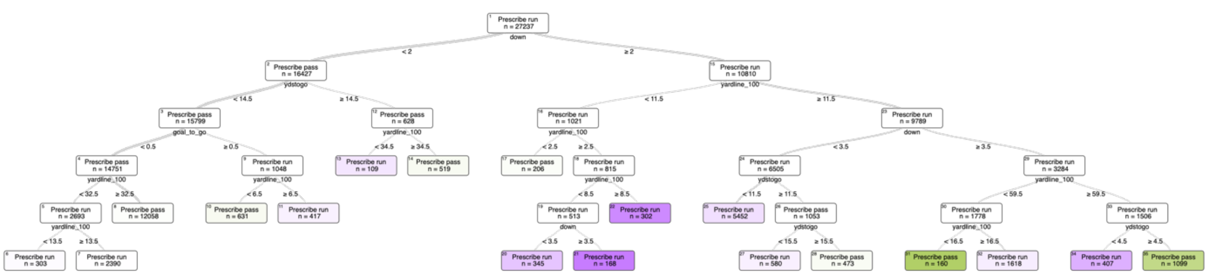

Figure 11: OPT for Good QB - Down Four Late

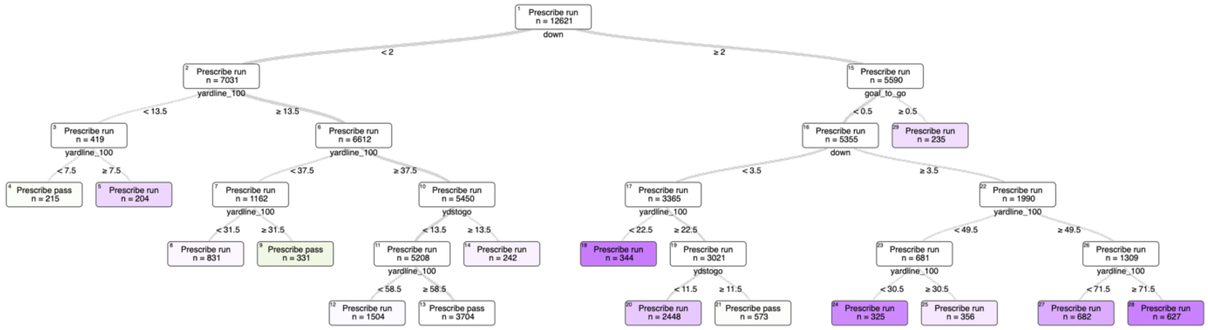

Figure 12: OPT for Bad QB - Down Four Late

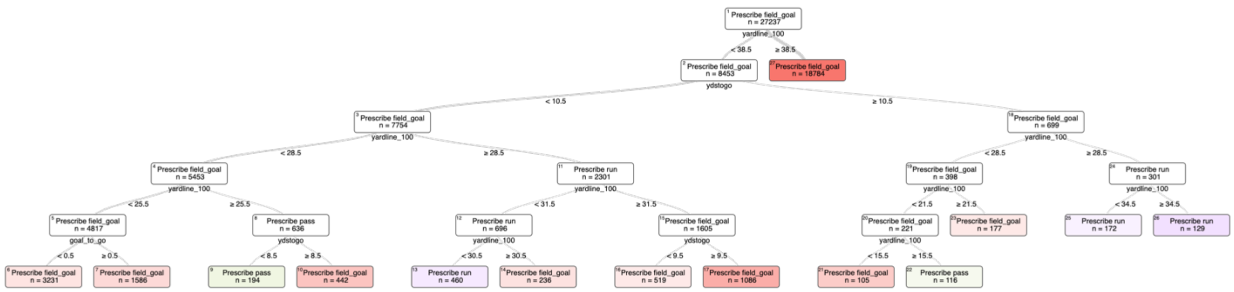

Figure 13: OPT for Good QB - Any Score

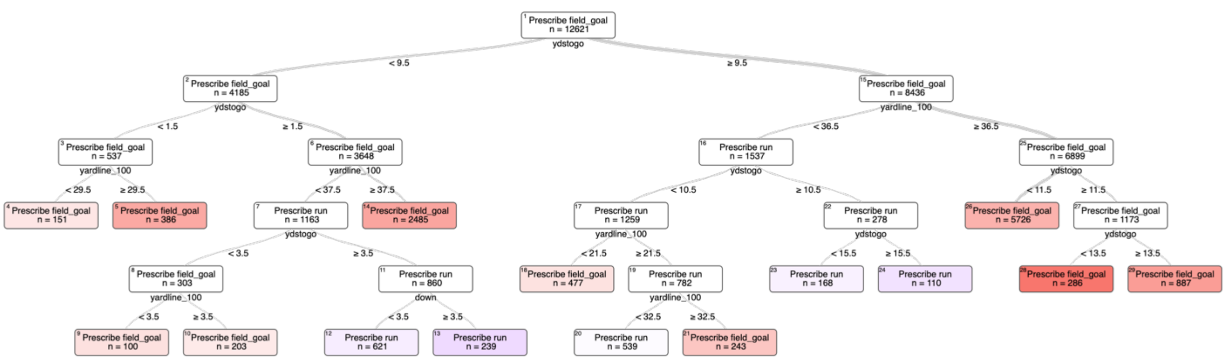

Figure 14: OPT for Bad QB - Any Score

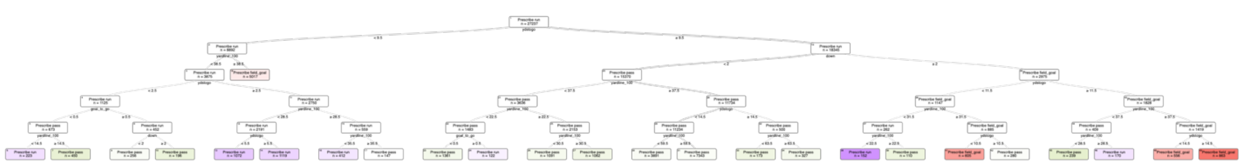

Figure 15: OPT for Good QB - Down Three Late

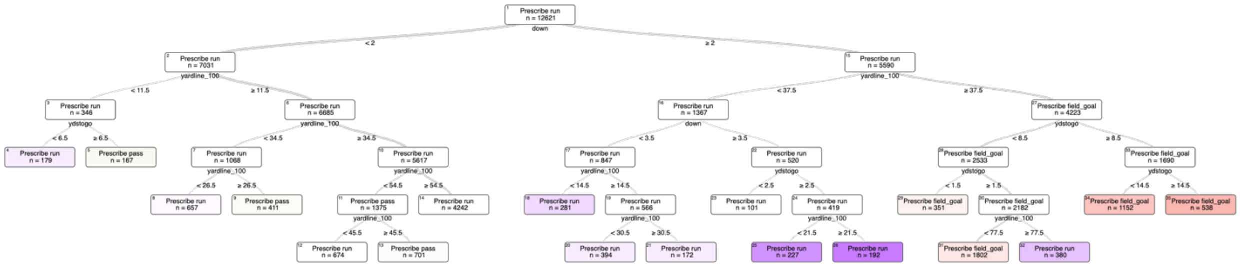

Figure 16: OPT for Bad QB - Down Three Late