A Submodelling-Based Approach to Expected Points in North American Football using Scikit-Learn

Baxter Madore

August 11, 2023

1 Introduction to Expected Points

Expected points is a metric that aims to quantify the value of a specific situation in a sports event, usually in order to compare it to a different situation in the same sport. The metric has seen use in soccer, hockey, basketball, and Australian Rules football. Not all scoring efforts are taken in the same conditions, and thus not all scoring efforts have the same likelihood of leading to a score. Using expected points, we can quantitatively describe the difference between scoring efforts, including whether one attempt, if repeated often enough, would likely result in more points for the scoring team. There is no meaningful distinction between "probability of scoring on a given scoring attempt" and "number of points expected from a given attempt" in sports like hockey and soccer, where all scores are equal and points are scored one at a time. These two sports, in particular, call the expected points metric "Expected Goals", often shortened to xG. In sports like Australian football and basketball, differences in scoring shots can result in different numbers of points for the attacking team,

and thus, expected points has an additional factor to consider over expected goals, namely, the number of points that a successful attempt will score. In this chapter, we will examine, recreate and extend an existing expected points model for the National Football League, with the ultimate goal of adapting one of them for use in the Canadian Football League.

1.1 How Are Expected Points Calculated?

Expected points are calculated using previously observed data. Each data point has multiple inputs, and one output. As an example, Robert Younger's 2016 model of expected points for the Australian Football League[13] takes as inputs:

- The distance from goal.
- The angle of the shot.
- The type of shot taken.
- How the shooter got possession of the ball.
- How long the shooter had possession before they shot.

The output of each individual data point is the number of points that the shot scored, which in Australian football can be 0, 1, or 6[1]. With all of these data points for each shot over the prior four seasons, approximate formulae can be constructed that use the parameters surrounding a shot to predict the average number of points that the shot scores.

1.2 Expected Points in Gridiron Football

In gridiron football, whether Canadian or American, most individual plays, while technically attempting to score points, are not designed specifically for an

[1] In Younger's case and most Australian football research, shots that fail to score any points do not have sufficient data, so the output space is restricted to $\{1, 6\}$, but ideally shots that score no points would be included.

immediate score. Points are not scored from "shots" with a binary outcome of hit/miss, but of drives consisting of many plays chained together. Still, most of the utility provided by an expected points measurement are still applicable in these football contexts.

Long, sustained drives consisting of many plays until a touchdown are worth the same amount of points as one extremely successful play leading to a touchdown on its own, so both of them should have the same result when modelled. This is why every expected points model in gridiron football that we have examined uses the scoring of the current drive as output, with some looking even further than that.

1.3 History of Expected Points Modelling in Football

One of the first expected points models for gridiron football was made by then-NFL quarterback Virgil Carter and his colleague Robert Machol in 1970. Carter looked at all 1^{st} and 10 plays from the first half of the 1969 NFL season. He divided the field into 10 equal chunks, and each play that he studied had one input, namely, the number of yards to the endzone rounded to the middle of the chunk of field that the ball started in, and it had one output, namely, the outcome of the drive [6][2]. Due to the limitations of the technology available in 1970, he could not perform analysis on downs other than the extremely common 1^{st} and 10, nor could he be more granular in the exact placement of the ball to come up with an expected points value for any specific field position.

Carter's work was adapted to the Canadian game by Peter Bell in 1982 [3], following the same methodology of splitting the field into 10-yard chunks. Bell's work was more comprehensive than Carter's, taking into account the chances of getting a first down on the drive, worth more points since the ball advances

[2]The output space was -7 (Defensive Touchdown), -2 (Safety Against), +3 (Field Goal), +7 (Offensive Touchdown), and the negative of the expected points from whichever chunk of field the defense took over the ball.

up the field if the offense gets a first down. Bell also considers the ensuing possession following a score. After a rouge, touchdown, or field goal, the team that gave up the score normally gets the ball back. For example, In [3], a drive that scores a 1-point rouge is treated as if it were only worth 0.695 points, since an offence starting at their own 35-yard line as they would after a rouge or field goal is scored[3] has an expected points value of 0.305 points, essentially discounting the value of the rouge and field goal by 0.305 points.

In 1987, Bob Carroll, Pete Palmer, and John Thorn developed an expected points model in their book "The Hidden Game of Football", and they also described a linear relationship between the distance that the offensive team was from their opponents' end-zone and the number of points that the next score would be. The authors of [4] took into account the full value of the next score, no matter when or what it was, as opposed to [6], whose analysis only included plays that happened within one change of possession and excluded defensive field goals and possession team safeties. Page 105 of [4] indicates that the authors "came to basically the same conclusions" as [6]. The results are similar even though the authors of [4] took more scoring possibilities into account than Carter and Machol and used data from over a decade later than [6]. The book indicates that the "point potential" (their term for expected points) of a drive can be given by the formula $0.08X - 2$, where X is the distance between the start of the drive and the offensive team's end zone, in yards. This is the first model to our knowledge which did not split the field into chunks, instead using values from each individual yard line. They also only attempted to model 1^{st} down situations.

The next model to our knowledge regarding Canadian football is by Keith Willoughsby [12] from 2001, who related the initial field position of a drive

[3]This was changed before the 2022 season. Offenses now start at their own 40-yard line after a rouge or field goal against, though they can elect to receive a kickoff after conceding a field goal.

to the next score, so long as the next score occurred within one change of possession. He obtained a linear fit between the two, determining that the expected points on a drive starting X yards away from the offence's own end zone was $0.054X - 1.738$, meaning that a drive starting from a team's 35-yard line was expected to be worth 0.152 points. The model in [12] only used data from the 1998 Saskatchewan Roughriders to fit the model, a considerably smaller sample than used by any previous model.

In 2002, David Romer published a paper [11] attempting to find the optimal play-calling balance on fourth downs in the NFL. He used a long-run approach, taking into account the score differential over an infinite game to determine the expected points from a specific situation. Romer was, in our examination, the first to model expected points non-linearly, using quadratic splines to smoothly model expected points across the field, using 8 "knot points" [11]. Romer also attempted to mitigate the effects of time remaining and score differential by using only first-quarter plays in constructing the dataset used to find the first down expected points.

The first expected points model we consider in this chapter that accounts for non first-down plays is Brian Burke's 2009 model [5], which models 1^{st} and 10 in the same way as its predecessors. Burke innovated on prior models by estimating the value of 2^{nd} and 3^{rd} downs with anywhere from 1 to 15 yards to go, instead of only estimating the value of 1^{st} and 10. Burke, like Romer, excludes plays to mitigate score and time effects, this time excluding all plays in the second and fourth quarter, as well as excluding plays where the score difference is greater than 10. Similarly to [3], Burke also discounts the value of a kickoff from any score. 1^{st} down, 2^{nd} down, and 3^{rd} down are all modelled separately, with no interdependence between one down and the other. 4^{th} down data is scarce as teams in close games where time is not a factor generally kick

the ball on 4^{th} down and do not attempt a conventional offensive play, so to estimate the probability of a team converting a $4th$ down, Burke uses data from 3^{rd} down.

In 2018, Ronald Yurko, Samuel Ventura, and Maksim Horowitz made a public expected points model as part of the larger nflscrapR project [14]. nflscrapR takes in more inputs than previous models, including time remaining in the half, whether the offence is in a "goal-to-go" situation and whether the two-minute warning in either half had already occurred. Additionally, this model uses the natural logarithm of yardage required for a first down as opposed to other models using the untransformed yardage required for a first down.[14] Unlike [5] or [11], this model does not exclude any plays from their analysis but instead weights plays less the larger the score differential is. Since time is an input to the model, plays are not excluded on the basis of there being a small amount of time remaining in the half. Plays are excluded from analysis only if the play is a quarterback kneel, whose sole purpose is to deplete the clock and scores no expected points, positive or negative.

In 2021, Ben Baldwin of OpenSourceFootball created an expected points model as one of many nflfastR models. Baldwin's study, [1] used the machine learning technique xgboost to fit NFL play-by-play data to a model. nflfastR uses more types of input than nflscrapR, including the NFL season that the game took place in, the stadium type, and whether the possession team is at home. [1] makes no mention of excluding or weighting plays for any reason. nflfastR also includes era adjustments to account for gradual changes in strategies and rules over the years.

Some models (like [5] and [11]) fully count the next score even if that score happens after many changes of possession, while the more modern nflscrapR applies a weighting penalty to scores beyond the current drive. Additionally,

unlike the other mentioned sports, each of which have free-flowing possession punctuated by a shot attempt, football is segmented into discrete plays. Most expected points investigations in football are unconcerned with modelling while play is live, preferring instead to use the state of the game prior to the play.

2 The nflscrapR Model

While the entire model description can be found in Section 3 of [14], the main advancement that Yurko, Ventura, and Horowitz made was to use multinomial logistic regression to predict the probabilities of each score given enough prior data, rather than just using a single value averaging the amount of points typically scored after that position. This is useful for expected points, as the probability of each event occurring can be multiplied by its value and added together to get a single number of expected points for a given scenario. It is also modular, with each scoring event able to be modified independently. One event could be "field goal attempt", without regard to the number of points that the kick actually scored. Instead of that being built into the rest of the model, there could be an outcome and associated probability for "field goal attempt", with a different model used for the actual field goals, perhaps one which takes into account the improvement of professional kickers over the years; see, for example [9].

2.1 Explaining Multinomial Logistic Regression

Logistic regression is a machine learning prediction technique that aims to estimate the probability of a dependent variable belonging to one or more output classes, based on one or more independent variables. There are two primary types of logistic regression: binary logistic regression, used when there are exactly two output classes, and multinomial logistic regression, when there are

more than two output classes. In this study, the independent variables consist of game state components, such as down, time remaining, and score differential. The output classes represent the various types of scoring events that can potentially occur next. The model is trained on a substantial amount of data, which establishes a relationship between the input data's parameters and the likelihood of belonging to a specific class. After the training phase, the model obtains optimal coefficients for all independent variables. These coefficients are then utilized to predict the "log-odds," denoted as ℓ, for a similar set of independent variables through the relationship,

$$\ell = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_N X_N,$$

where β_0 through β_N are the coefficients found by the fitting process for N independent variables, and X_1 through X_N are the values of the independent variables. [8] The log-odds are subsequently transformed into predicted probabilities using the logistic function: ($\hat{p} = \frac{1}{1+e^{-\ell}}$"). Because the log-odds are found using a linear function, logistic regression is sometimes known as the log-linear classifier.

To mitigate the effects of possible overfitting, models can be fit with a regularization term, which penalizes higher magnitudes of β. Infinite regularization corresponds to each β being zero, meaning that any predicted probability for an outcome is simply the mean of the probability for that outcome across the whole training dataset, regardless of the values of the independent variables of the trial being predicted. Zero regularization corresponds to unrestricted values for each β.

2.2 Training-Testing Split

Multinomial logistic regression, like any other machine learning algorithm, is prone to overfitting, and although regularization can help by making sure none of the coefficients or parameters are too extreme, testing the performance of a model on data that it wasn't trained with can be used to evaluate and compare different models to see which would be best at predicting data in the future. The 2018 season, as the middle season in the data, was selected as the test season, to evaluate the models against each other on data they were not trained with.

3 Constructing the Model for the CFL

3.1 Why a New Model

Canadian football and American football have many similarities in gameplay and tactics. This makes adapting an existing American football model to Canadian football feasible. Despite these similarities, there are several key differences in the rules and strategies used in each type of football that impact optimal strategies and scoring rates. For example, the Canadian rule that all defenders must start a play at least one yard behind the line of scrimmage as opposed to the American rule allowing defenders as close as 11 inches behind the line of scrimmage makes 2^{nd} or 3^{rd} down with a yard or less to go much easier to convert in Canadian football than American football, and as such, the possessing team's scoring probabilities would be higher in the Canadian game. Other rule and game differences (larger field, more players, 3 downs instead of 4, goalposts at the front of the end zone, etc.) prevent the use of American football play-by-play data in modelling the Canadian game. The possibility space for scoring in Canadian football is also larger than that of the American game,

with the one-point rouge being a possibility in addition to all possible scoring plays in American football. Though [3] has already created a model for CFL expected points, strategies, tactics, and rules have changed in the nearly half-century since 1978. Additionally, NFL expected points modelling is much more sophisticated than it was in the days of Carter and Bell, taking in much more data than before, evaluating values on downs other than 1^{st}, and discarding the assumption that expected points must be linear with respect to field position.

3.2 Our Model

Our model is largely based on nflscrapR, but with some additional tweaks.

- The model is fit with more recent data based partially on games that had not been played when [14] was written.
- The "Under Two Minutes" Variable has been changed for an "Under Three Minutes" Variable. The CFL has a three-minute warning at the end of each half, and even when strictly dealing with NFL data, predictions were (very slightly) improved by this change.
- Because NFL overtime has vastly different strategies and is quite rare, as well as being largely inapplicable to the CFL's "shootout" overtime format, overtime plays were excluded from analysis.
- During testing, the probabilities of the next score being a safety for either team are usually low enough to always fall into the first (0% to 4%) bucket and as such, the model is not evaluated on how well it predicts the probabilities of safeties from either team.
- Different input parameters are used, based on the prediction accuracy, including a term for the score differential, and using the raw yards-to-first parameter instead of its logarithm

- After attempting to weight the training data, we observed worse predictions on the test data. As a result, the models presented here do not weight the input data.
- The primary modification to nflscrapR that we implement is the introduction of many sub-models for different game states. Instead of solely relying on a single monolithic model, smaller sub-models are created to enhance prediction accuracy. This approach maintains the reliability and power of the main model while better handling specific game states.

The multinomial model was created using the Python machine learning library `scikit-learn` [10] with the following settings:

- Regularization: None. The data set for all models and sub-models was large enough that underfitting was more of a problem than overfitting, and adding regularization led to less accurate predictions. This also makes the penalty setting irrelevant, as no penalties are ever applied.
- Solver: lbfgs,
- Maximum Iterations: 30,000,
- Tolerance: 10^{-4},
- Class: Multinomial,
- Sample Weights: All ones,
- All other settings: scikit-learn 1.3.0 defaults.

4 Testing the Model

To train the model, it was fit to data from the 2015 to 2017 NFL seasons and the 2019, 2021, and 2022 NFL seasons. Due to the unusual scheduling and personnel

of the 2020 season, it was excluded from the model training. Though the CFL has an application programming interface (API) which includes play-by-play data, its use is restricted and we were unable to gain access to it, and as such, the various models were tested with NFL play-by-play data obtained using the `nfl_data_py` package [7]. Should we at some point in the future gain access to CFL or other Canadian football data, the code can be easily repurposed, with multinomial models fit to the Canadian football data.

To test each model's accuracy, it was tasked with predicting probabilities of each type of score (excluding safeties) or no score for every play in the 2018 NFL season, which was excluded from the training set.

4.1 The Accuracy Metric

To test the accuracy of the model, probability predictions were made for each play of the test season. The model assigned a probability to each possible "next score" outcome. Those outcomes are, in ascending order of value:

- Possession team concedes touchdown.
- Possession team concedes field goal
- Possession team concedes safety
- Scoreless (i.e. there is no more scoring in the half)
- Possession team scores safety
- Possession team scores field goal
- Possession team scores touchdown

25 bins were made for each of the possible outcomes, excluding safeties. Each of these 25 bins represented a 4% probability interval, with the first bin for

each outcome containing plays with a chance lower than 4% of resulting in that outcome, the next bin containing plays with a chance between 4% and 8% of resulting in that outcome, and so on.

For each bin, the average estimated probability of the selected outcome for each play in the bin was compared to the observed proportion of plays in that bin which ended with the selected event. The difference between those two numbers was multiplied by the number of plays in the bin, then multiplied by the overall proportion of plays in the training data that ended in that outcome. These multiplications amplify errors on more common events, making them more important to predict correctly.

4.2 An Example Play

As an example, in Super Bowl LIII, with 669 seconds remaining in the second half, the Rams faced a 1^{st} and 20 from their own 33 yard line, in a tie game against the Patriots. The model's predicted probabilities for the next score in such a situation are:

- Patriots touchdown (-TD): 18.7%
- Patriots field goal (-FG): 13.1%
- Half expires with no further score (scoreless): 10.4%
- Rams field goal (+FG): 28.3%
- Rams touchdown (+TD): 29.2[4]%

This play would then go into the fifth bucket for opponent touchdown probability, the fourth for opponent field goal probability, the third for scoreless

[4]Listed probabilities do not add to 100.0% because of both rounding and the exclusion of safeties

probability, the eighth for possession field goal probability, and the eighth for possession touchdown probability.

Plots were made for the probability of each common type of score given the model's expected probability. In this plot, the size of each dot indicates the number of plays in that bucket, and the dot's proximity to the blue line (which represents the line where predicted probability and actual probability are equal) shows the predictive accuracy of the model over plays in that bucket. Figure 1 shows the prediction accuracy graph of the possession team scoring a touchdown, and Figure 2 shows the prediction accuracy graph of the possession team conceding a touchdown.

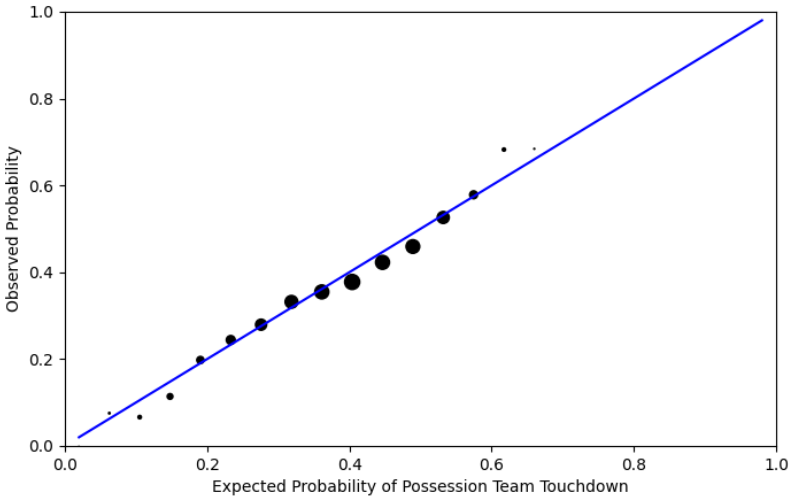

Figure 1: Model Accuracy Graph, showing the actual probability of the possession team scoring a touchdown given the model's predicted probability, in 4% buckets.

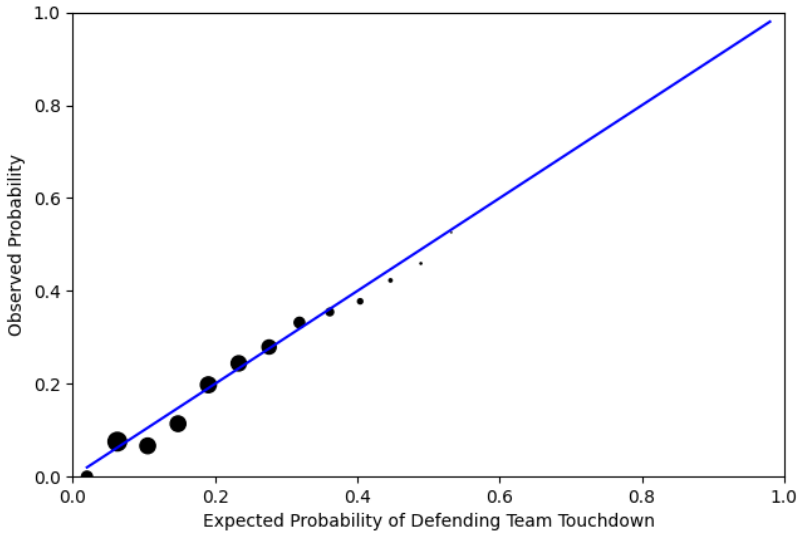

Figure 2: Model Accuracy Graph, showing the actual probability of the possession team conceding a touchdown as the next score in the same way as Figure 1

4.3 Comparing Different Models

Many different models were tested, taking in different types and numbers of parameters, and their prediction errors are listed below (lower is better). "Distance" refers to the distance to a first down, while "field position" refers to the distance from the end zone that the offense is attacking. All values are rounded to the nearest tenth:

- Down, Distance, Field Position: 811.4
- Down, Distance, Field Position, weighted by score differential and drives to score: 1278.3
- Down, Distance, Field Position, Time: 763.5
- Down, Distance, Field Position, Time, Last Down?: 709.3

- Down, Distance, Field Position, Time, Last Down?, Goal-to-go?: 655.8
- Down, ln(Distance), Field Position, Time, Last Down?, Goal-to-go?: 1023.5
- Down, Distance, Field Position, Time, Score Differential, Goal-to-go?: 586.1
- Down, Distance, ln(Field Position), Time, Score Differential, Goal-to-go?: 1460.4
- Down, Distance, Field Position, Time, Score Differential, Goal-to-go?, Last Down?: 595.3
- Down, Distance, Field Position, Time, Score Differential, Half, Goal-to-go?, Last Down?: 627.1
- Down, Distance, Field Position, Time, Score Differential, Goal-to-go?, Under Three Minutes?: 540.1

5 Sub-Models and Different Game States

5.1 Kickoffs and Expected Points Adjustments

After a score, the team that was scored against normally gains possession of the ball, whether that's from a scrimmage at their own 40 following a rouge or field goal against, or from receiving a kickoff after a field goal or touchdown against. The exception to this is the safety, where the scoring team gains possession of the ball through either a kickoff or a scrimmage at their own 40. The possession following this kick is worth something, and has to be taken into account for any expected points model. The multinomial approach is used again to determine the value of a kickoff, given the yard line. [5], score differential, and time

[5] This is almost always the 30 yard line for touchdowns and field goals, but was the 35 yard line before the 2022 season, and can change when penalties occur or after a safety.

remaining in the half. Including or removing the terms for score differential or yard line make little difference to the accuracy of the predictions, but they have been incorporated. Because the largest factor in expected points from a kickoff is the time remaining, Figure 3 shows the expected points of a team receiving a kickoff, as a function of the time remaining in the half.

5.2 Kickoff Expected Points Time Adjustment

Since scoring takes time, and time is the biggest factor in the points expected from a kickoff, to properly make the adjustments to the expected points model taking the kickoff into consideration, there needs to be an estimation of how much time will pass until the next score. We use the average time until next score of all plays that had a next score. This is a simple way to factor the subsequent kickoff into the expected points calculations without assuming a constant value for the expected points from a kickoff or assuming that subsequent scoring within a half is instant. The value of the expected points from a kickoff is calculated based on the time remaining, and the absolute value of touchdowns and field goals are lessened by that amount, with the absolute value of safeties being increased by that amount.

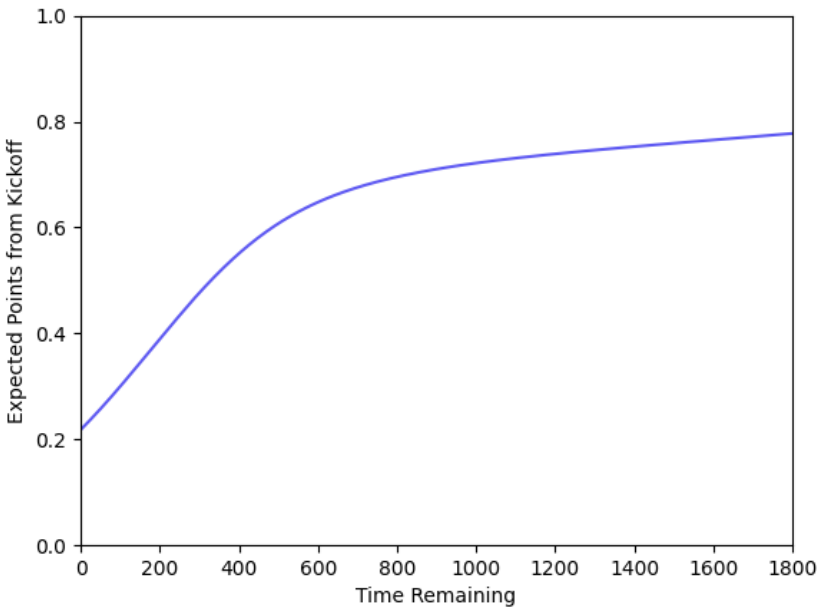

Figure 3: Expected Points from a Kickoff, given the time remaining in the half. Kickoffs are assumed to be from the 35 yard line in a tie game.

5.3 The Endgame

In the final minutes of a game, tactics change drastically, as the score difference and the time on the clock become much bigger factors than in the previous 55 minutes of the game. Using only plays from the last 5 minutes of the game should create a dataset that is better tuned to those specific game states, given the same inputs as the main model. When tested on all plays from the final five minutes of games in 2018, the error for the main model was 356.5, and the error for the "Final 5" model was 117.5, indicating that tailoring a model to the endgame results in better predictive power when used.

5.4 Goal-To-Go

Since, in the main model, the "Goal-To-Go" binary variable causes a huge jump in the expected points when set, indicating that just being in a goal-to-go scenario is very valuable, even more so than being in the same spot on the field without being in a goal-to-go scenario, a separate sub-model was made for just goal-to-go scenarios, taking in most of the same parameters as the main model. The two exceptions to this are the parameters for distance to first down, and goal-to-go. The distance to first down is always the same as the distance to the end zone, so it is omitted. Similarly, the goal-to-go parameter is true for all plays by definition, so it is also eliminated. Interestingly, over the 5 seasons used for training, at no point was a team in a goal-to-go situation with the next score being a conceded safety, so the model assigns it zero probability in all cases. However, since conceding a safety, especially after multiple changes of possession, is still possible, the main model is used to predict the probability of the possession team conceding a safety in these scenarios, and the probabilities are all normalized so they sum to 1.

5.5 Final Down

When testing the main model on NFL data, the largest error in prediction was "possession team field goal". To mitigate this issue, a separate 4^{th} down model was built. This model takes into account distance to first down, distance to goal, time remaining, score differential, whether the offense is in a goal-to-go situation, and whether the offense is 37 yards away from goal or closer, in "field goal range".

5.6 Why the Large Jumps?

The binary variable for field goal range leads to a sharp jump in field goal chances and expected points, while leading to sharp declines in the probabilities of all other major scores, as seen in Figure 4. Without this extra variable, the 4^{th} down model predicts a non-negligible chance that a field goal is attempted from 60+ yards away from goal, while the longest field goal made in NFL history was attempted on a play starting from 49 yards away from goal, and attempts from further are extremely rare. Including an independent variable deciding whether a team is in field goal range significantly enhances prediction accuracy on fourth down as it the model no longer predicts unprecedented field goal attempts when this variable is included. Out of all integer yard lines in the range 30 - 40, the one that resulted in the lowest prediction error for next score was 37, indicating an effective field goal range of 54 yards, at least in 2018. This distance has likely gotten longer since then, as Benjamin Morris' kicking analysis has found a roughly linear improvement in NFL kickers' accuracy over time since the 1970s [9].

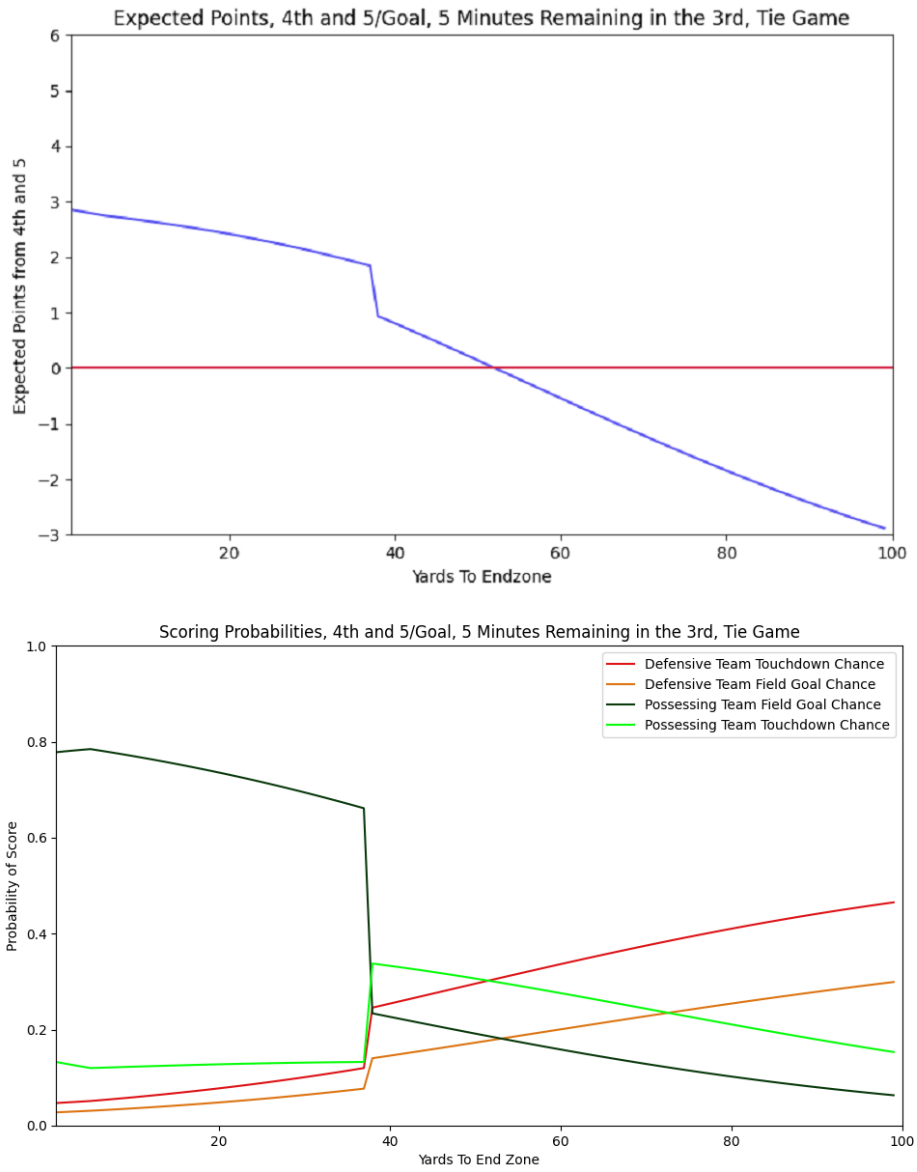

Figure 4: Expected Points and next score probabilities on 4th and 5. Notice the large jump in expected points when a team is considered to be in range and the drastic shifts in probabilities. These are unnatural, but increase predictive capability

5.7 Decision Making

Because multinomial regression uses output classes, it is possible to make the output classes categorical as opposed to numerical. This allows for the creation of a model to predict the probabilities of the play type on 4^{th} down, whether they are field goal attempts, punts, or conventional offensive plays. The play type classification in the data may not fully represent the offense's choices. As an example, if a team chose to punt, but the punter fumbled the ball, this would be recorded as a run play, and would be categorized as a conventional offensive play. No accuracy testing was performed on the decision analysis, though the log-loss metric would be well suited for such an application. The decision probability predictions take in the same parameters as the expected point predictions, including the "field goal range" variable, which also results in sharp changes in the probability of punts and field goal attempts at the 37 yard line.

5.8 The Conventional Play Model

After including all of the sub-models described above, the main model is trained with redundant data that it would not be called on to predict; those plays would be handled by the sub-models. To better predict plays that the main model actually attempts to predict, the main model is replaced by a (still very large) model that uses the same input parameters, with the exception of the binary goal-to-go indicator, which would always be zero due to the inclusion of the submodel to handle those scenarios introduced earlier. For the 2018 season, filtered to exclude plays handled by a submodel, the conventional play model outperformed the main model 424.4 to 447.9 on applicable plays. Even this model shows a faster than normal increase in expected points and touchdown probability when inside the 10 yard line, due to the fact that distance to first

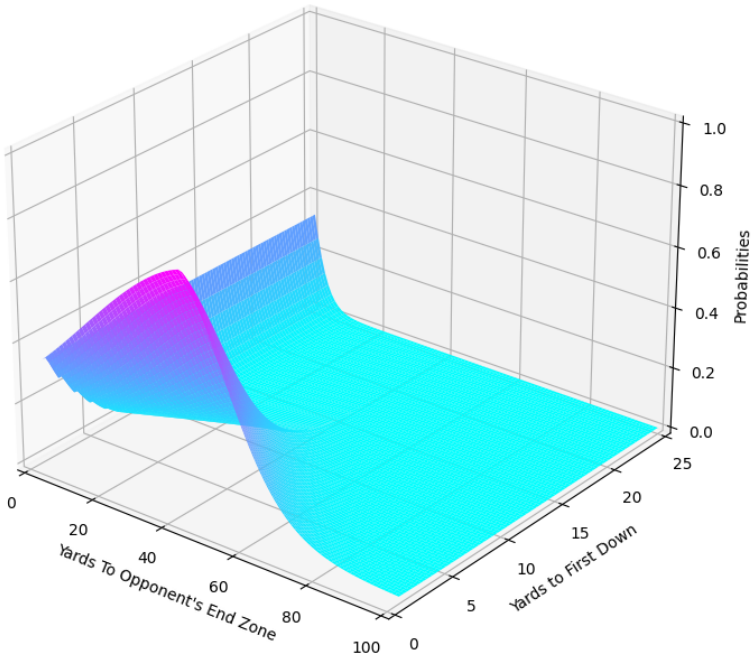

Figure 5: Probability that a team selects a run play or pass play on 4^{th} down, by distance to end zone and distance to first down.

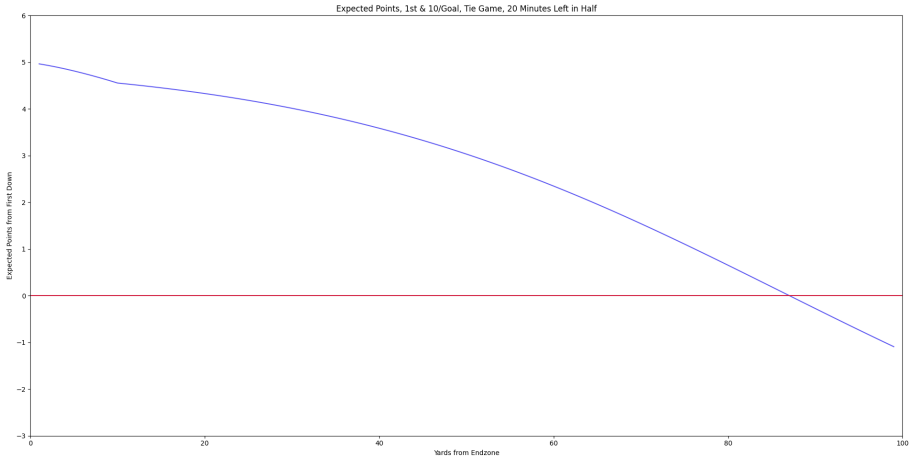

Figure 6: The Expected Points From 1st Down, as predicted by the "Normal Model" which does not take into account whether the offense is goal-to-go. Notice the transition in the line's shape starting at the 10 yard line, where the offense transitions from first and 10 to first and goal (less than 10)

begins decreasing at that point.

6 Attaching the Models

Given that the main model, the 4^{th} down model, the kickoff model, the goal-to-go model, and the final-5 model are all separate, in order to make a single cohesive model, they need to all be pulled together. Using the python `pickle` library, the fitted coefficients for each model can all be saved to files, and those models can all be called in the same script. Some of the sub-models may overlap, in which case, priority is in the following order.

- Final 5 Minutes
- Goal-to-Go
- Fourth Down

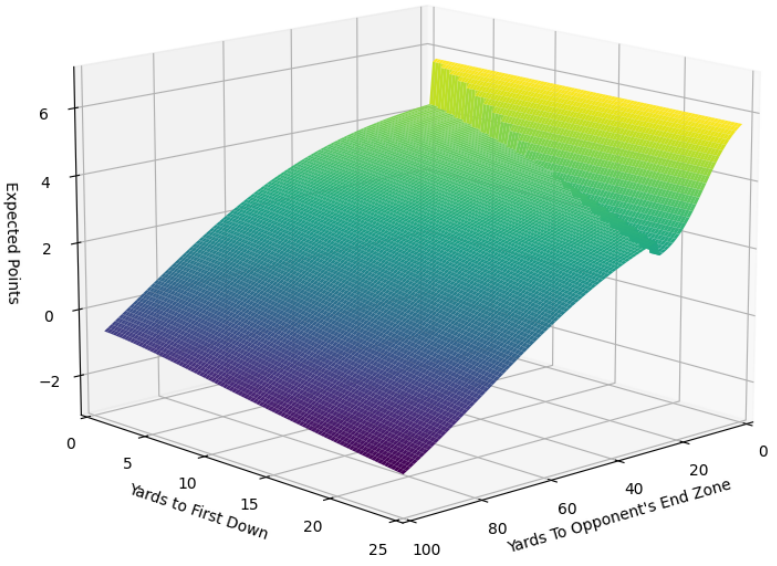

Figure 7: The Expected Points From 1st Down, across the normal model and the goal-to-go model, based on distance to goal and distance to first down. Higher points/darker colours indicate a higher point expectation.

7 Future Work and Improvements

- Make the models transition into each other more smoothly, instead of sharp discontinuous jumps as appear in the currently implemented model. There's nothing special about the 37 yard line that makes teams' field goal probability shoot way up, even though that predicted the test season very well. One possible solution would be to split the final down submodel into two, one for either side of midfield. The very longest field goals made in both the NFL and CFL were both on plays starting within a yard of midfield, though teams do not often attempt field goals of that distance. The final down decision probabilities should change smoothly with more emphasis on actual kick distance rather than the arbitrary concept of "field goal range".

- See if total score rather than just score difference can affect future scoring probabilities. Are plays in games with an already high scoring rate more likely to result in further scoring?

- Take a deeper look at converts, instead of assuming that each convert is worth 0.97 points. What game factors lead teams to decide to go for 2, what makes them decide to kick, are there game state factors that make 2-point conversions more or less successful?

- Incorporate the Morris field goal model into the extra point and field goal averages to reflect the fact that kickers are continually getting better, and perform a similar analysis for CFL kickers.

- Use the Expected Points model to determine the expected points from different strategies or decisions (e.g., choosing to go for it on third down, runs vs. passes, different risk levels in play calls, running a kick out of the

end zone or conceding a rouge), in order to make the models much more useful for coaching and decision-making.

- Similarly to the progression that Yurko et. al. made in [14], use the expected points model to create a CFL win probability model, which uses the current game state, including the expected points of the current situation, to predict the probability of either team winning the game.

- Remove arbitrary cut-offs, or smoothly transition models into each other, perhaps using a weighting system, so that there are not large unnatural changes at 300 seconds. A technique like LOESS (as used in [5]) could be used to smooth the transition between the final 5 minute model and the normal model.

- Find out what time range teams' strategies start changing to end-game strategies on offense and defence, and whether changing strategies earlier or later is more likely to result in winning the game, similarly to analysis on optimal goaltender pulling strategies in hockey [2].

- Use the multinomial approach in a similar way to determine the probabilities for the terminal outcome of a given offensive drive. What factors make a drive more likely to end in a lost fumble, or a field goal attempt, etc?

- Properly cross-validate and evaluate the models, instead of having just one test season, the models would be trained on each possible combination of five seasons out of the six in the dataset, and tested against the sixth.

References

[1] Ben Baldwin. *Open Source Football: nflfastR EP, WP, CP xYAC, and xPass models*. 2021. URL: https://www.opensourcefootball.com/posts/2020-09-28-nflfastr-ep-wp-and-cp-models/ (visited on 08/01/2023).

[2] David Beaudoin and Tim B Swartz. "Strategies for pulling the goalie in hockey". In: *The American Statistician* 64.3 (2010), pp. 197–204.

[3] Peter Bell. "Analysis of strategies in the Canadian football league". In: *INFOR: Information Systems and Operational Research* 20.2 (1982), pp. 116–125.

[4] John Thorn Bob Carroll Pete Palmer. *The Hidden Game of Football*. University of Chicago Press, 1988. ISBN: 9780226825878.

[5] Brian Burke. *Expected Points (EP) and Expected Points Added (EPA) Explained*. URL: http://www.advancedfootballanalytics.com/2010/01/expected-points-ep-and-expected-points.html (visited on 06/29/2023).

[6] Virgil Carter and Robert E Machol. "Operations research on football". In: *Operations Research* 19.2 (1971), pp. 541–544.

[7] cooperdff. *NFL-data-py*. URL: https://pypi.org/project/nfl-data-py/ (visited on 07/04/2023).

[8] IBM. *What is Logistic Regression?* URL: https://www.ibm.com/topics/logistic-regression (visited on 08/04/2023).

[9] Benjamin Morris. *Kickers are Forever*. URL: https://fivethirtyeight.com/features/kickers-are-forever/ (visited on 07/11/2023).

[10] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[11] David H Romer. *It's fourth down and what does the Bellman equation say? A dynamic programming analysis of football strategy*. 2002.

[12] Keith A Willoughby. "The return of a missed field goal in Canadian football". In: *Chance* 14.3 (2001), pp. 29–33.

[13] Robert Younger. *A Model to Predict and Rate Shots by Quality*. URL: https://www.figuringfooty.com/2016/08/04/a-model-to-predict-and-rate-shots-by-quality/ (visited on 07/05/2023).

[14] Ronald Yurko, Samuel Ventura, and Maksim Horowitz. "nflWAR: a reproducible method for offensive player evaluation in football". In: *Journal of Quantitative Analysis in Sports* 15.3 (2019), pp. 163–183. DOI: 10.1515/jqas-2018-0010.