Winning Strategies: Analyzing 4th Down Tactics in American Football

ISYE-7406: Data Mining & Statistical Learning
April 14, 2024

Patrick McGee, 136, pmcgee7@gatech.edu
Victor Ye, 230, vye6@gatech.edu
Sophia Oku, 755, soku3@gatech.edu
Andrew McAlister, 596, amcalister7@gatech.edu
Pranay Shah, 415, pshah426@gatech.edu

Abstract

This project explores fourth-down decision-making in the NFL, specifically focusing on the critical choices between attempting to continue the drive, punting, or kicking a field goal. While fourth down decisions historically relied on a coach's intuition, the increasing complexity and stakes in NFL games, underscored by the sport's massive viewership and lucrative media deals, suggest a more analytical approach. This project applies data mining and statistical learning to NFL play-by-play data from 2016 through 2023 to develop models that predict the optimal fourth-down decision based on various game state variables such as: yards to go, yard line, score differential, and time remaining.

The methodology integrates three modeling efforts: a "win probability" model, Probability Mass Function and Success Probability Estimation for each potential decision, and a simulation model that assesses the impact of each decision type on the win probability. The analysis uses data from seven NFL seasons for model training and validation, with the 2023 season serving as a test dataset. This approach not only allows for a thorough assessment of the decisions made during games but also tests the hypothesis that NFL teams are overly conservative on fourth down, opting to go for it less frequently than the optimal decision model suggests.

Results indicate a discrepancy between the model's recommendations and actual coach decisions, suggesting that teams could improve their game management by aligning more closely with the analytical predictions. The study confirms the initial hypothesis and proposes that a more aggressive approach to fourth-down decision-making could benefit teams, reflecting an underutilization of analytical tools in current strategies. Future work could expand the models' complexity and tailor them to individual team strengths and weaknesses, enhancing their practical application in real game scenarios.

Introduction

In fast-paced American Football, split-second decisions often determine the outcome of games, defining the line between victory and defeat. Every play matters, but none perhaps as crucially as the fourth down. The choice between going for it, punting, or kicking a field goal on fourth down is a strategic question that coaches face regularly.

Historically, coaches have relied on intuition and instincts developed over many years of experience. However, as many "Monday morning quarterbacks" will attest, it is not clear that they always make the right decision. NFL football is arguably the most popular sport in America. In 2023 the NFL clocked in at almost a trillion viewed minutes and completed multi-year media deals valued at $120 billion. Getting fourth down decisions right benefits players, coaches, and fans.

The goal of this paper is to use data mining and statistical learning techniques to develop models that provide the optimal fourth down decision in NFL games based on a given set of game state variables (e.g. yards to go, yard line, score differential, time remaining, etc). Additionally, we will compare the optimal decisions to the actual decisions made during fourth down plays of the 2023 NFL football season.

Problem Statement

Apply data mining and statistical learning techniques to historical NFL play by play data to determine an NFL team's optimal 4th down decision for a given game state. Possible decisions are (1) go for it, (2) punt, or (3) kick a field goal.

Use the developed models to assess how often NFL teams decided to go for it on fourth down relative to the optimal choice.

Hypothesis

NFL teams will decide to go for it on 4th down less often than they should.

Data Sources

Every NFL play-by-play data from 2016 through 2023. Data from 2016 through 2022 is used for training/validation and data from 2023 is used for analysis. Before filtering, there are 336,324 rows of data for 2016 through 2022 and 49,665 rows of data for 2023. Each row represents a play. Filtering was done to exclude plays not relevant to our analysis.

For the 2016 through 2022 training/validation data set, any plays that were not a run, pass, field goal, or punt were filtered out. Also, any plays that were not on 3rd or 4th down for run and pass plays were filtered out (3rd down play-calling often mimics 4th down play-calling in that the play-callers are looking to get all yards needed for a first down). For the 2023 test data set, any plays that were not a run, pass, field goal, or punt were filtered out. Additionally, any plays that were not on 4th down, were under two minutes remaining in either half, or were in overtime were filtered out. That left us with 3,605 test data points.

Each play contains 384 features. The number of features was initially filtered to 40 features. This was done because many of the features were either specific to a particular team (not needed because analysis focused on league averages) or derivatives of other variables. From these 40 features, we arrived at a final 12 features during the modeling process. Data can be accessed through the nfl_data_py Python library.

Proposed Methodology

Our methodology involves development of three interrelated models that support arriving at the optimal fourth down decision. For our analysis, we define the optimal fourth down decision to be the one with the greatest increase to the team's "win probability."

The overall methodology is described immediately below followed by the detailed methodology associated with the development of each of the interrelated model types.

Overall Methodology

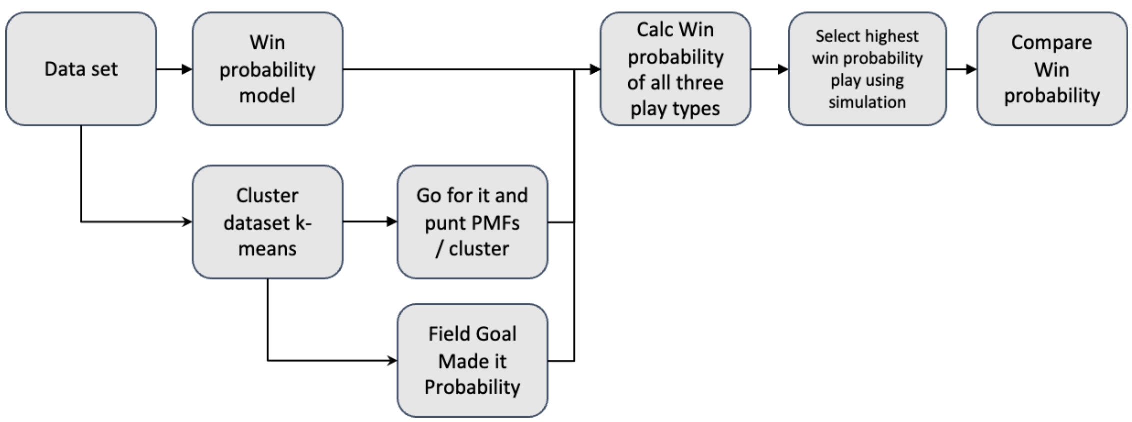
Task 1: Develop a "win probability" model, which calculates the probability that the offensive team will win the game given the current game state (down, yard line, score differential, time remaining, etc.).

Task 2: Develop Probability Density Functions (PDFs) and/or Probability Mass Functions (PMFs) or Success Probability that are associated with each potential 4th down decision: go-for-it, kick a field goal, or punt. PDFs and PMFs will take into account the game state at the time of the decision.

Task 3: Develop a simulation model, which uses the PDFs / PMFs to simulate running each play type a large number of times and uses the win probability model to determine the expected impact on win probability from each decision.

Task 4: Using the win probability model, PDFs/PMFs, and simulation model, calculate the win probability before a play and the expected win probability associated with running each possible fourth down play type (go for it, kick a field goal, or punt). Perform this calculation for most of the fourth down plays in the 2023 NFL football season to perform analysis and report results.

The diagram below provides a visual representation of the overall methodology.



Task 1 - Win Probability Model

Background

In football, a "win probability" model predicts a team's likelihood of winning a game based on several in-game factors such as time remaining, score, field position of the possession team, yards to go for a first down, etc. A win probability model enables evaluation of play decisions and player selection by assessing how much a given play increases or decreases the probability of winning the game relative to the league average performance for a given game state. There have been multiple attempts to develop a win probability model that can accurately predict the chance of winning in the NFL. Burke developed a widely used model that takes historical data and simplifies the data by grouping each play based on conditions that lead to a win (Burke, 2014). Then the model provides an estimate of winning based on score and time remaining. In 2018, Yurko et al. proposed multinomial logistic regression to estimate the expected play points, integrating a generalized additive model (GAM) to estimate the win probability of each play, and then developed a win above replacement statistic (WAR) that uses multilevel models for each individual player on the offensive side.

Approach

We attempt to replicate the model developed by Yurko and then explore multiple other approaches in an attempt to improve upon the Yurko model. We will build models using: random forests, boosting (AdaBoost, GBM, and XGBoost), logistic regression, and Generalized Additive Model (GAM). Model inputs will include relevant game state variables. The model output will be the probability that the offensive team wins the game. All of the selected models are capable of outputting a probability.

Model Validation

K-fold cross validation will be used to tune parameters and to assess model performance. The best performing models will be used to generate win probabilities associated with every NFL play in 2023.

Of the eight NFL seasons of data, seven seasons (2016 through 2022) are used for training and validation and the 2023 season is used for testing. Each of the 7 seasons is used to represent a fold, so 6 seasons are used to train and the 7th season is used for model assessment. Yurko refers to this approach as Leave One Season Out CV (LOSO CV).

Error Measurement

Model evaluation includes use of a tailored performance metric and cross validation based on this performance metric. Since the focus of the model's output is on the probabilities associated with winning rather than classification, traditional classification metrics such as misclassification rate, accuracy, precision, recall, and F1 are not good measures of a model's performance. As a result, we will implement a method of evaluating performance that was developed by Yurko, Ventura, and Horowitz (2018) for use on their win probability model. Their metric was based on the method used by Lock and Nettleton (2014) for NFL win probability models.

Using this method, the predicted probabilities are compared to observed probabilities by creating bins that aggregate similar predictions together. The win probabilities are binned in five percent increments leading to 20 possible bins total. All of the predicted probabilities with a value that falls within the range of a bin (e.g. 0.15 to 0.20) are aggregated together and the mean predicted probability of this aggregated group is calculated (e.g. 0.185). The observed probability of this bin is calculated as the actual number of times the binned observations won divided by the number of observations in the bin.

The error associated with a bin is the absolute value of the difference between the predicted probability and the observed probability.

$$e_b = |\hat{}P_b(Y=1) - P_b(Y=0)|$$

The mean absolute error is the weighted average of the mean absolute errors of the bins.

$$e = \frac{1}{n}\sum_b n_b * e_b$$

Evaluation of Model Stability

Graphs of predicted probabilities versus actual probabilities will be created for candidate models in order to assess stability across the range of probabilities from 0 to 100% and throughout the game from the first quarter through the fourth quarter.

Task 2 - Probability Density Estimations

Overview

We plan to develop Probability Density Functions (PDFs) and/or Probability Mass Functions (PMFs) that can be used to simulate each potential 4th down decision: go-for-it, kick a field goal, or punt. PDFs/PMFs. The “go for it” distribution provides yards gained for a given game state (e.g. yards to go, yard line, score differential, time remaining, etc) at the time of the fourth down decision. The “punt” distribution provides net punt yards for a given game state. The “field goal” distribution provides probability of the field goal being good for a given game state.

Since there are a very large number of combinations of game states, the problem of developing density estimations that are dependent on game state is very complex. We reduced the complexity of the problem through two methods:

- Performed k-means clustering to develop clusters of similar game scenarios and developed Probability Mass Functions for each cluster.

- Used domain knowledge of football to group game scenarios. We assumed plays with the same ‘yards to go’ would have similar Probability Mass Functions (PMFs). We assumed that teams are trying to gain a minimum of the yards to go when they are going for it on 4th down, and that plays would be similar across different yard lines, game times, and other game scenarios.

Task 3 - Simulation Model

Overview

In order to increase the utility of our win probability model, and be able to apply it to specific game situations, we needed to create a simulation model that would combine the individual play type models with the win probability model to identify the best decision to make on a critical 4th down play. The simulation model allows for quickly simulating outcomes of a potential play situation to more robustly evaluate both the play and win probability models. This helps reduce some of the variability inherent to the result of a football play.

The simulation model begins by identifying a 4th down game state that will be used to identify an ideal play decision. A custom created game state can be created, but to simplify our process we chose to query the 2023 play by play dataset to find 4th down plays that fit certain game situations we wanted to simulate. For example you could query for a 4th down and 5 yards to go play from the opponent's 23 yard line while down 4 points.

After identifying a game state, that data is loaded into the win probability model to generate an initial win probability. This serves as the baseline value to compare the future win probabilities for each play type decision. Next, the game state is run through the probability models for each play type (a traditional run/pass play, kicking a field goal, or punting to the other team). This involves selecting the relevant PMF for a play's game state, converting the PMF into a CDF, generating a random number between 0 and 1, and using inverse transform to find the appropriate yards gained from the PMFs.

The play type models return a simulated result for that play (yards gained, field goal % change, and punting yards). The simulation model takes the result of the play and updates the game state to reflect the simulated result. This is a complex process as the simulation model has to determine if the simulated play resulted in points scored, a turnover in possession, and what yardline the other team will take over possession (if applicable).

Once the game state has been simulated through each of the play type models and the new game states are created, they are each run through the win probability model to determine the updated win probability for each play type. The simulation model runs this process through the selected number of iterations requested and the mean win probability for each play type is recorded. Finally the model outputs the original game state, original win probability, and the mean win probability for each play type that was simulated (go for it, field goal, and punting). What results is a comparison of win probabilities of each potential decision leading to an "ideal" decision to go for a traditional play, kick a field goal, or punt to the other team based on which play win probability is the highest.

Simulation Model Verification

We will verify that the simulation model performs properly by having it re-create a given PMF input model. Additionally, we will verify that the model's calculated win probabilities for each of the three potential decisions look reasonable for a sample of data points.

Analysis and Results

Our methodology involves development of three interrelated models that support arriving at the optimal fourth down decision. This section provides analysis and results at both the individual model level and at the overall project level.

Analysis and Results for Task 1 - Win Probability Model

Logistic Regression, Random Forest, and Boosting models, including AdaBoost, Gradient Boosting Machine, and XGBoost, were developed to estimate the conditional probability that the offensive team would win the game given the existing game conditions (e.g. score, time remaining, yard line, etc.). Additionally, a win probability model based on a generalized additive model (GAM) that was detailed in an academic research paper by Yurko, Ventura, and Horowitz (2018) was built.

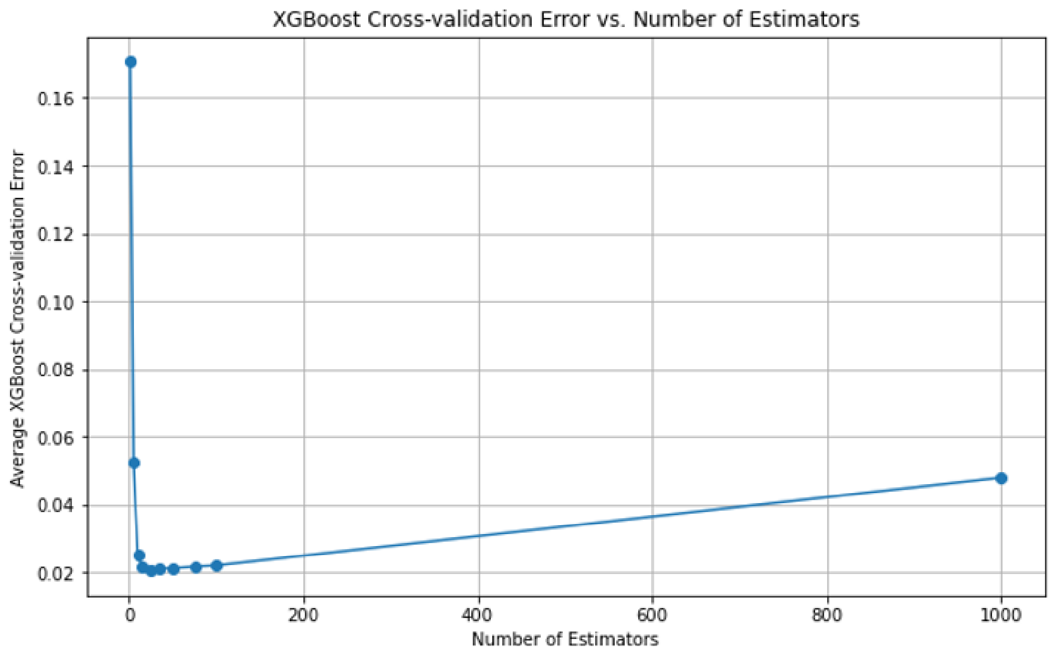
As shown in the table below, XGBoost resulted in the best performing model both in terms of accuracy and speed. It had a mean absolute error of 0.02 percentage points and outperformed the GAM model. Logistic Regression performed almost as well as XGBoost and also outperformed the GAM model. Random Forest had long run times and performed significantly worse despite considerable effort at tuning parameters. AdaBoost had by far the worst performance.

Cross Validation Error Measurements

Model Type	**CV Mean Absolute Error**	**Comment**
Logistic Regression	0.0209	No improvement with parameter tuning.
Random Forest	0.053	Long run times. Increasing n_estimators showed modest improvement.
AdaBoost	0.2333	Poor performance overall.
Gradient Boosting Machine (GBM)	0.0226	Long run time.
Extreme Gradient Boosting (XGBoost)	**0.0207**	**Tuned number of estimators (see below). No improvements from tuning other parameters.**
Generalized Additive Model (GAM)	0.0220	Well performing model developed by Yurko, Ventura, and Horowitz (2018). Has a second logistic regression model that provides “expected points” as an input.

Variable Selection: Initial variable selection was based on domain knowledge of features most likely to be relevant. A modest amount of experimentation on variables did not improve model performance. Given the large number of features and long run times for some of the models, the initial selected features were used for all of these models. The variables used in the models included: 'score_differential', 'Down', 'half_seconds_remaining', 'yardline_100', 'goal_to_go', 'ydstogo', 'UnderTwoMinutes', 'game_seconds_remaining', 'h2', 'OT', 'Posteam_timeouts_remaining', 'defteam_timeouts_remaining'

Parameter Tuning: Other than GAM, efforts were made to improve performance of all models through parameter tuning. The table below shows the tuning of the "n_estimaters" parameter in the XGBoost model.



Performance on Test Data, NFL 2023 Season: The logistic regression model, the XGBoost model, and the GAM model were re-trained on all of the training data (2016 through 2022 seasons) and then used to predict "win probabilities" for every play in the 2023 season. These predictions were compared to actual probabilities (as described in the methodology section above). The results are included below. The performance of the logistic regression model, the XGBoost model, and the GAM model all improved relative to the cross validation mean absolute error.

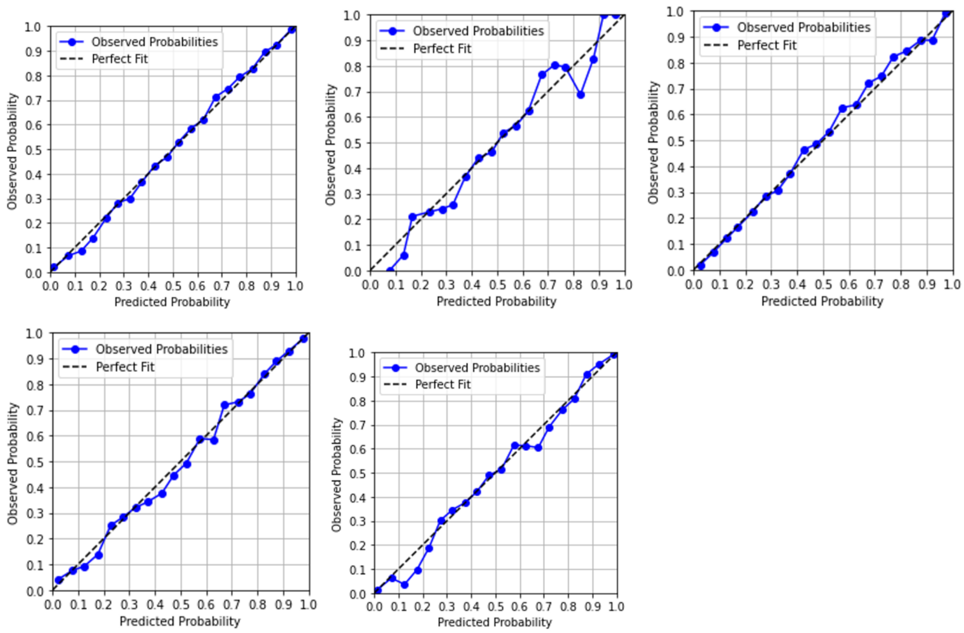
NFL Season 2023 Test Data
Error Measurements of Win Probabilities

Model Type	Mean Absolute Error	Comment

Logistic Regression	0.0147	Improvement over the cross validation mean absolute error
Extreme Gradient Boosting (XGBoost)	**0.0127**	**Improvement over the cross validation mean absolute error. Still best performing model**
Generalized Additive Model (GAM)	0.0138	Improvement over the cross validation mean absolute error

Model Stability / Calibration Across Range of Probabilities and Across Four Quarters: In addition to using mean absolute error to evaluate the models, we developed the plots below showing the observed probability versus the predicted probability. Ideally, the plotted points should fall on the dotted lines. These plots are intended to show that the models are well-calibrated across the range of predicted probabilities from 0 to 1, and are well-calibrated across the entire game. XGBoost plots are below.

XGBoost Probability Plots on Fitted 2023 Data
Plots are in the following order: Entire Game, First Quarter, Second Quarter, Third Quarter, Fourth quarter



Analysis and Results for Task 2 - Probability Density Estimations

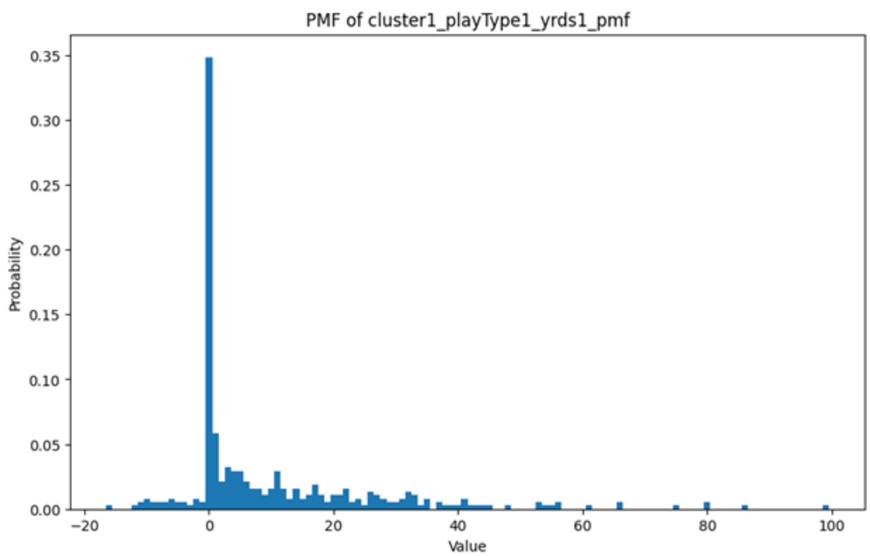
We plan to develop Probability Density Functions (PDFs) and/or Probability Mass Functions (PMFs) that can be used to simulate each potential 4th down decision: go-for-it, kick a field goal, or punt. The “go for it” distribution provides yards gained for a given game state (e.g. yards to go, yard line, score differential, time remaining, etc) at the time of the fourth down decision. The “punt” distribution provides net punt yards for a given game state. The "field goal" distribution provides probability of the field goal being good for a given game state.

Probability Mass Functions - Clustering

We developed both continuous probability density functions and discrete probability density functions for fourth down decisions.

The probability mass function is specifically designed for discrete random variables. It assigns probabilities to individual outcomes in a sample space. For a discrete random variable, the PMF provides the actual probabilities of specific outcomes. If (X) is a discrete random variable, the PMF (P(X = x)) gives the probability of (X) taking the value (x). For these reasons and better accuracy of our model, we decided to use the PMFs. Details of the PDF development are in the Appendix.

Figure – Sample Probability Mass Function Distribution on play type 'goforit'

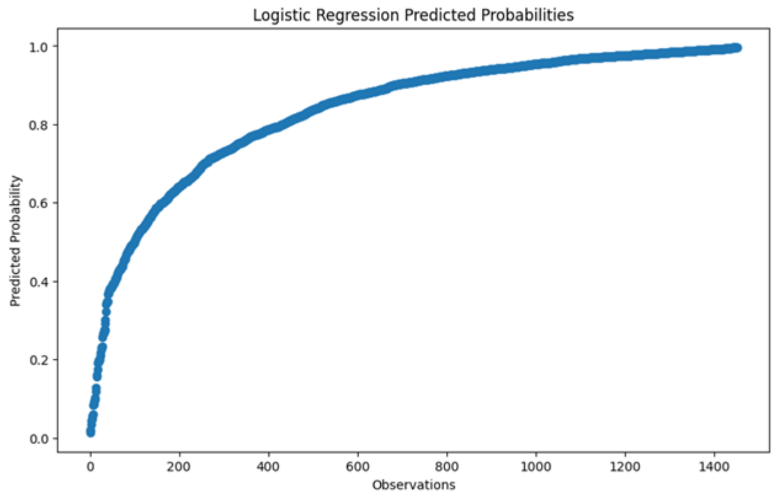


Logistic Regression

Further analysis of the 'field goal' play type led us to use logistic regression. The result of field goals is not based on distance the ball traveled but rather the play resulted in a made or missed field goal. In this case, we did another switch to our data gathering of ensuring that we don't use a PMF but rather calculate the success probability.

In our first iteration, we simply calculated the success probability as total field goals made divided by the total field goal attempted. This was a great start but doesn't provide any value when it comes to play by play analysis. Ultimately, this led us to create an additional logistic regression model with the sole purpose in predicting the success probability of making a field goal. The model was simple where we would take in all game state features as relative features. For this model, all in-game information was important for considering field goal success.

Figure – Sample Predicted Success Probabilities for Field Goal Play Type

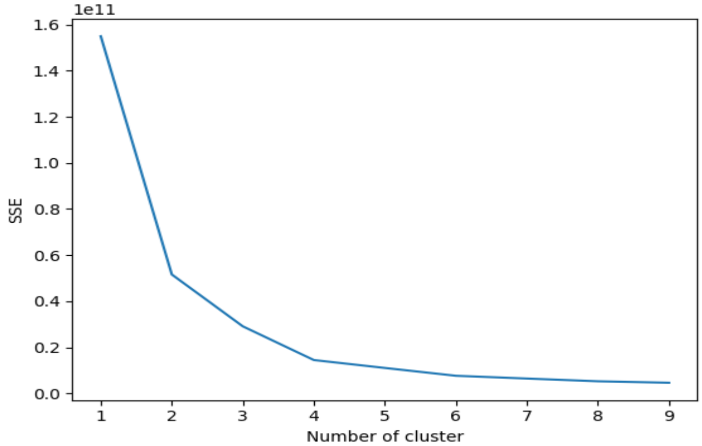


Unsupervised K-means Clustering

The play-by-play dataset we gathered was large with, before filtering, 336,324 unique rows (i.e. plays) for seasons 2016 through 2022 and 49,665 rows for the 2023 season. The initial plan was to ensure that we gather play type distributions by different game states. However, even with the many rows of game states, it's very rare for a game state to duplicate. The variables generally will be different. However, even though they are not exactly the same, they are similar in some ways. After several attempts to see the repeating game state, it was concluded that there is a very limited amount of repeated game state. In addition, it's unreasonable to establish a distribution per game state, but rather, we should find a way to gather similar game states together for the analysis. Our exploration led us to employ the K-means clustering algorithm. This approach allows us to identify similarities among all the data points in an unsupervised manner.

To create our K-means model, we filtered the data to include only plays categorized as 'goforit', 'punt', and 'field_goal'. Instead of arbitrarily selecting a cluster count, we conducted statistical analysis to identify the optimal number of clusters based on our dataset. This determination was made using the Sum of Squared Errors (SSE) from the center of the clusters.

Figure - Elbow graph different K in K-Means Model



Based on the elbow graph above, the elbow point on the k-means plot was at 4 clusters. This means that at k = 4, the majority of the variance will be captured by the model.

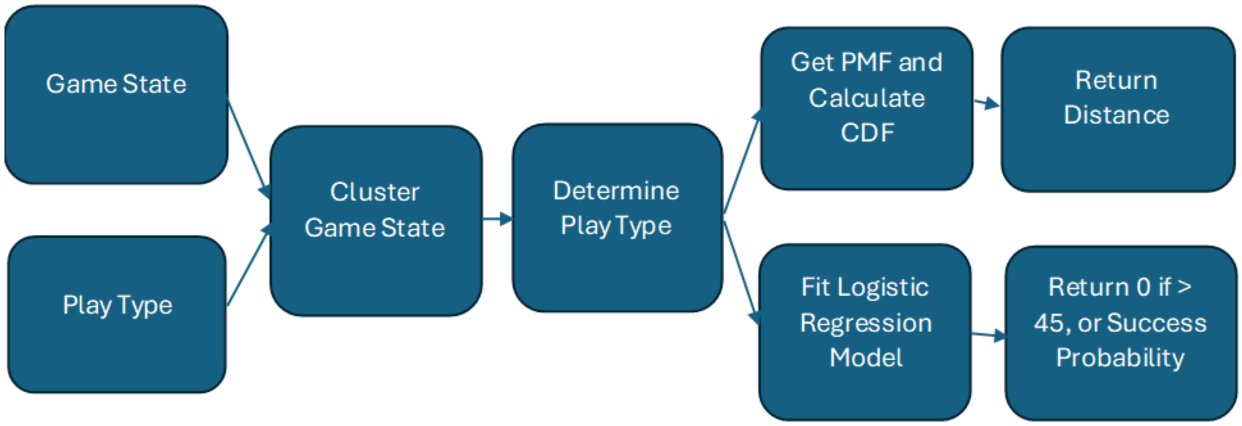
Play Type Model and Creation

Now that all components are established, we will combine all parts and scenarios to create our play type model. We start by running a k-means clustering algorithm to cluster our data into 4 separate clusters. This similarity labeling will be an additional feature to our dataset.

With play type 'goforit' and 'punt', we will take out probability mass functions of each type and clusters. For play type 'goforit', we will create the function based on distance gained on each play. This equates to distance travel by passing and running down distance. On the other hand, for play type 'punt', this is derived from kick distance. Together, they made up the new feature total distance gained. To add on more emphasis on the differential of game state, we separated the PMFs into further subdivision into discrete distance left on the play, 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10, where the 10 represents 10 or more yards to go. The PMFs for all combinations of the 4 clusters, 2 play types and 11 yards to go were created and saved into a separate csv. The decision to save the PMFs was made for the sole purpose of computation speed for the next steps. Once we have created the PMFs, we can use this information to create Cumulative Distribution Functions (CDFs) for use by the inverse transform function of the simulation model.

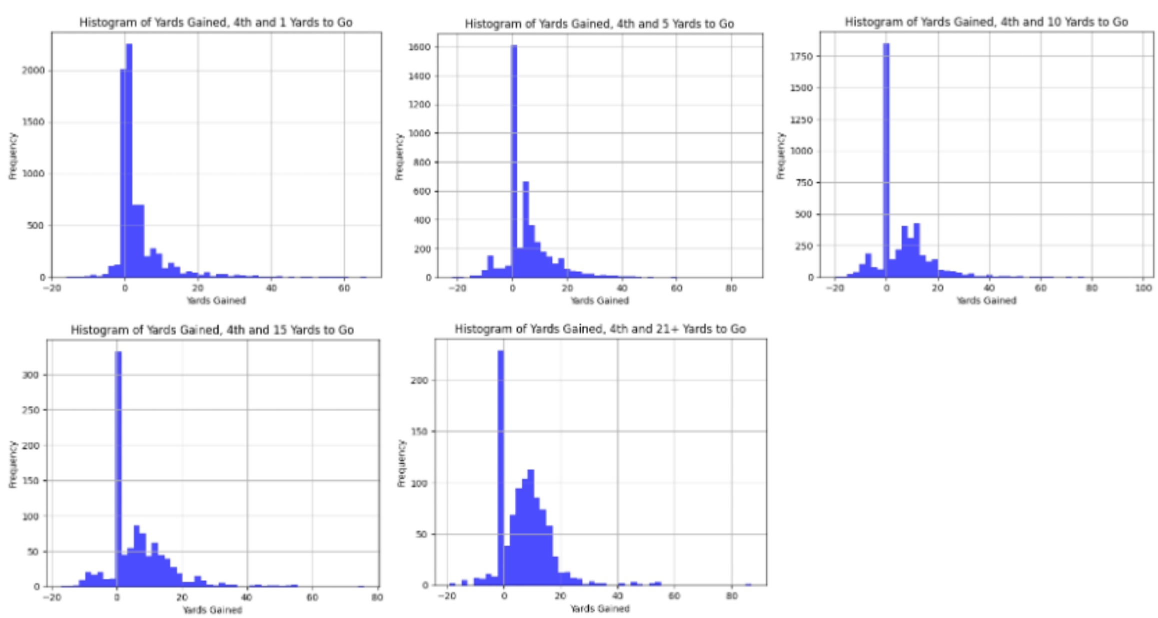
With play type 'field goal', we continue from our logistic model of success probability. Since the creation of the original model was based on our full data set, we will use the same model to predict the success probability by simply running the predicted functionality of the model. We did add in an additional rule, where if the yard line to the end field is greater than or equal to 45 yards, we will automatically assume a success rate of 0.
The combination of all the components makes up our play type model. To ensure that we can use this for our simulation model, the input of the model will include the game state, the play type of either 'goforit', 'punt', or 'field goal'. As a result, you'd expect a yardage to return if play type 'goforit' or 'punt' is inserted, or a success probability if 'field goal' was selected.

Figure - Play Type Model



Alternate Probability Mass Functions based on “Yards to Gain”

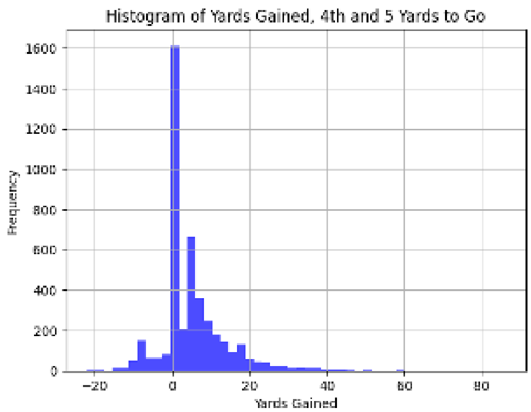
Alternate PMFs of ‘yards gained’ were developed for the play type ‘go for it’. Density estimations were based solely on the “yards to go” for a first down (broken down in this case from 1 to 21+). Assumption is teams are trying to gain a minimum of the yards to go on 4th down. Representative PMFs are shown below.

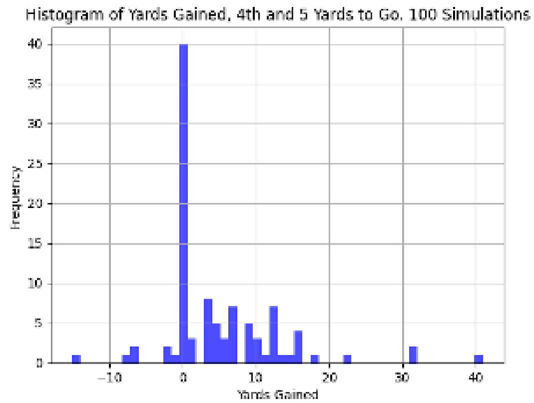


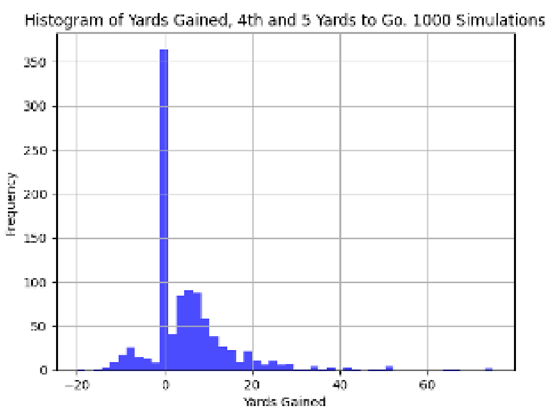
Analysis and Results for Task 3 - Simulation Model

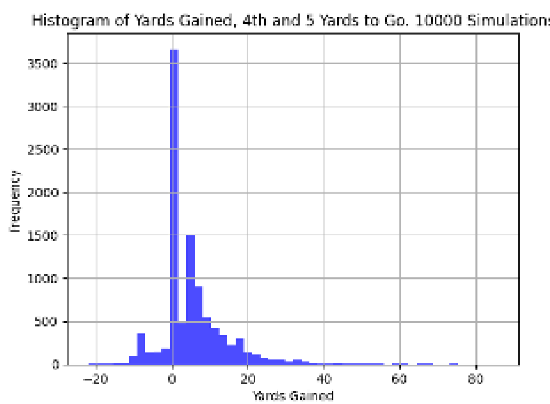
Model verification

We verified that the simulation model performed properly by having it re-create a given PMF input model. It can be seen below that the more simulations run, the more closely the simulation histograms mirror the original PMF's histogram. The original PMF is the histogram in the upper left corner below.









Model Results

In order to test the viability of the simulation model we chose to query specific 4th down scenarios to see how the initial play win probability compared to the expected win probability of the 3 potential play decisions. Below is a grid showing the results of this sampled run:

Yards to Go	Yard line	Score Diff	Seconds Left	Pre-Play WP	Go for it WP	Punt WP	Field Goal WP	Actual Play

1	Opp 1	-7	469	23.8103%	**45.2391%**	20.711%	23.324%	Went for it (S)
2	Opp 2	+9	1322	89.1523%	**89.8171%**	82.554%	89.1323%	Field Goal (S)
2	Opp 3	+10	1574	89.9327%	**90.323%**	80.3911%	89.5799%	Field Goal (S)
3	Opp 4	+5	742	87.6067%	**84.7405%**	76.639%	83.7926%	Field Goal (S)
5	Opp 5	+2	988	78.5068%	70.7461%	54.8554%	**74.6469%**	Field Goal (S)
5	Opp 6	0	1329	61.7418%	**61.9047%**	47.9004%	56.4378%	Field Goal (S)
7	Opp 7	-5	433	32.0297%	**35.1145%**	14.235%	29.3585%	Field Goal (S)
8	Opp 8	-5	170	34.7943%	**38.9609%**	11.8125%	27.7822%	Went for it (F)
9	Opp 9	-7	808	33.1454%	**27.1697%**	16.659%	17.7709%	Field Goal (F)
5	Opp 10	-14	439	5.2441%	**5.2134%**	1.6621%	4.6171%	Went for it (F)
4	Opp 20	+2	374	76.6415%	75.437%	66.9776%	**78.1895%**	Field Goal (S)
10	Opp 30	-23	133	0.0956%	0.1508%	0.0166%	**0.2469%**	Went for it (F)
5	Opp 40	+23	690	99.6764%	99.8146%	**99.8304%**	99.828%	Punt

Simulation results. Bold cells indicate best play choice based on expected win probability for that game situation. S next to Actual Play indicates play was successful and F indicates failure

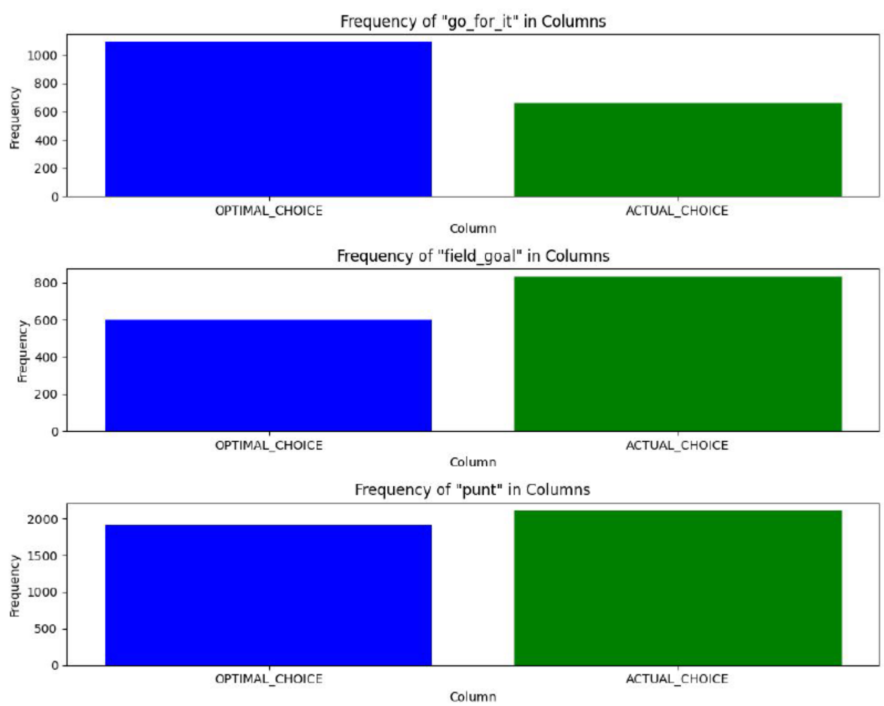
We can see from the results grid that in most of our game situations, since they are relatively close to the end zone and generally late game situations, the model recommends the team runs a play to attempt to score or get the first down (if applicable). Interestingly we can see from the actual play result column that most teams chose to kick the field goal. This indicates that generally the NFL is more conservative with play calling but it's important to note that in many of these cases the field goal win probability is very close to the "go for it" probability. All things equal, many coaches would choose to kick a field goal, feeling it's more likely to result in points and a move in a positive direction for the overall game. We do see that in specific game situations where our "go for it" win probability shows a significantly higher win probability than punting or kicking a field goal that the actual team did in fact go for it. This is likely due to the play situation of time, score, and yardline more or less forcing a team to go for it as the only viable option.

Overall Project Analysis and Results

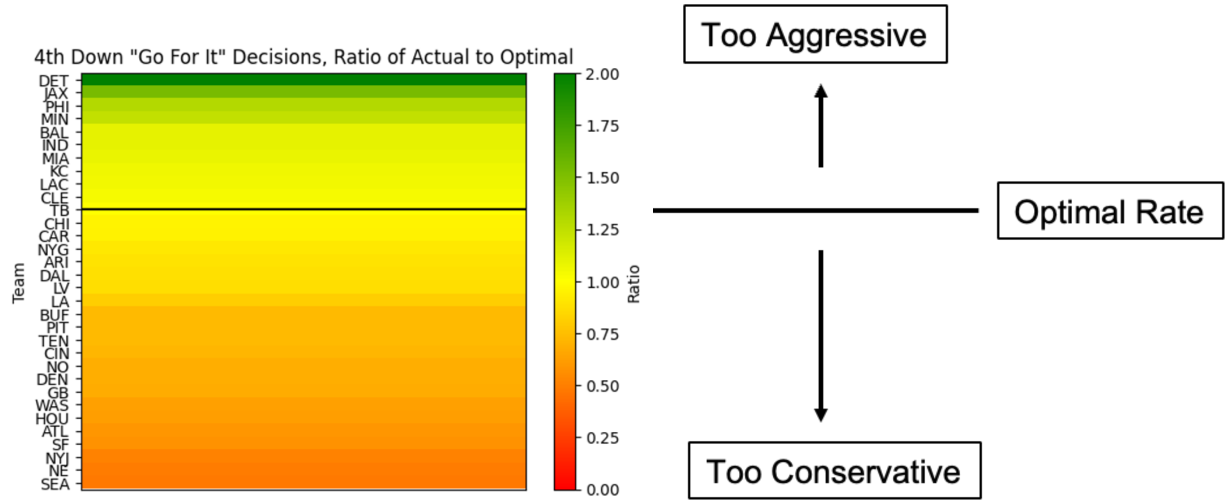
We analyzed 3,605 fourth down plays in the 2023 season. These 3,605 plays consisted of all fourth down plays in the season after filtering out edge cases that our win probability model either could not handle or handled poorly. Filtered plays included those where there was under two minutes remaining in either half, overtime plays, and play types that were not run, pass, field goal, or punt. In addition to the data shown in the model results table above, we captured the offensive and defensive teams on each play.

We assessed how often NFL teams decided to go for it on fourth down relative to the optimal choice through the use of several visualizations.

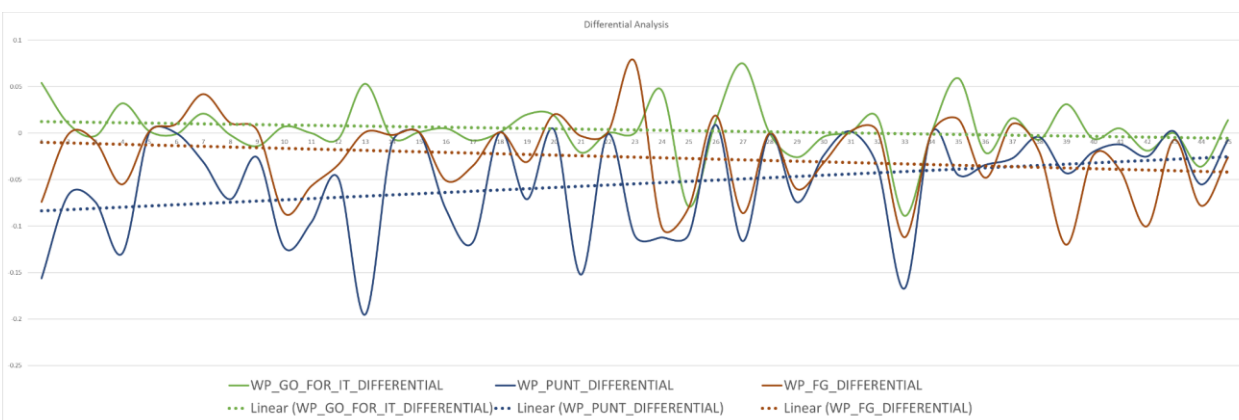
The first set of graphs below illustrate the frequencies of when the optimal fourth down choice was to “go for it”, kick a field goal, or punt versus the actual choice the teams made. As shown in these graphs, NFL teams are less aggressive in going for it on fourth down in real-life compared to the frequency that our model suggests that they should go for it. If we look at the first two columns in the top-most chart, we can tell that the optimal choice would be to “go for it” ; however , in reality, teams only chose that choice approximately half the time. Teams would rather attempt a field goal, a safer option with less of a point reward associated with it. This is shown by the second column chart where teams choose to attempt a field goal 30-40% more often than they should based on the optimal choice which will increase their win probability. Teams are not as aggressive as they should be. Lastly, we have punts where according to our analysis and modeling, teams punt slightly more than they should.



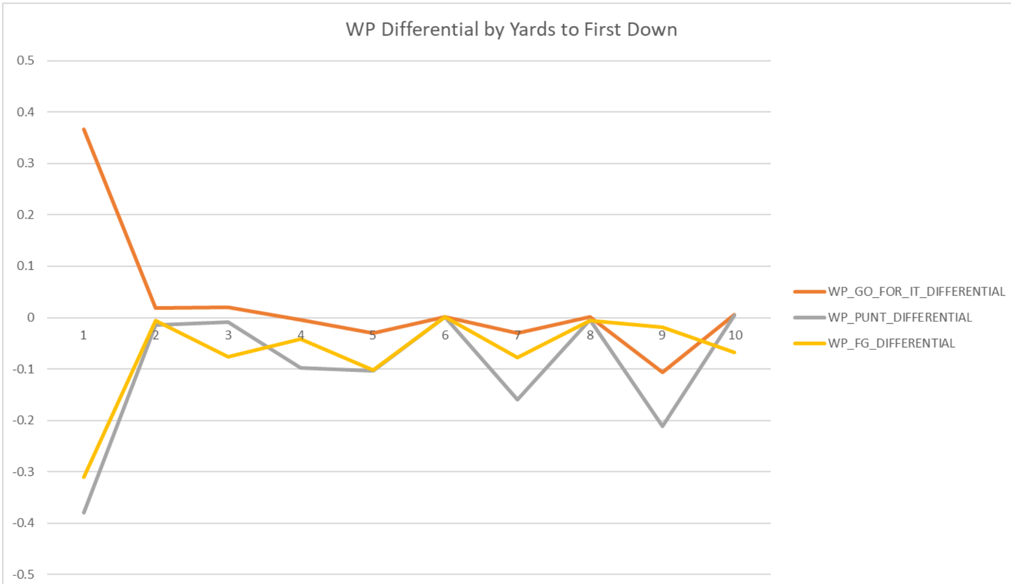
The second visualization was a heat map that compared the ratio of actual to optimal ‘go for it’ decisions by team. Ten teams went for it too often, Tampa Bay went for it at the optimal rate, and 21 teams went for it too little.



A third way of attempting to understand tendencies of all league-wide teams is to take a look at a data-wide sample of plays where the team had a given win probability for a certain yard-to-endzone. We then take the before probability and subtract it from the post-play probability of "go for it", "punt", and "field goal". As we can tell based on the trendlines, the closer we are to the end zone, the more sense it makes to go for it, then it makes sense to score a FG, and lastly it makes sense to punt. As we move farther away from the endzone - it makes more sense to punt depending on other parameters than it does to attempt a field goal. The need to go for it is dependent on other factors such as yards to first down. There's an inherent risk in going for it for teams that already have a high winning probability. In those circumstances, a field goal is the better option.



In the final figure below, we see a different win probability differential analysis. This is similar to above, but is broken down by yards to first down. We've used the standard 1-10 yards although teams can have more than that in a single fourth down. As we notice, the win probability jumps tremendously when a team is one yard from the endzone and they decide to "go for it". The punt and field goal do fluctuate below the "no change" line however they're probably more of a function of how far you are from the endzone (see figure above). It is to be noted that these visuals are based on sampled data from our large dataset and do not have aggregated values such as averages or medians.



Conclusions

Our hypothesis, "NFL teams will decide to go for it on 4th down less often than they should", appears to be supported. Teams tend to be more risk-averse than they should be based on visuals shared above - our analysis shows that teams would perform optimally if they were less hesitant to go for it. It seems that avoiding risky plays is the league-norm.

Overall, our methodology to solving the problem seems like a sound approach - we were able to generate output data that seemed reasonable for a decent number of plays.

Caveats

It's important to keep in mind that our analysis uses league wide averages, so our data provides the optimal decision for an average offensive team, with an average kicker playing against an average defensive team. In reality, there is variability in team and player performance. What is optimal for the average team may be suboptimal for a team that is not average. A team with a highly reliable field goal kicker capable of making 60 yard field goals is going to have a different probability of success than a team with an average field goal kicker. One example that showed up in our analysis was the Kansas City Chiefs, the Super-Bowl winners for the period of our test dataset, consistently appear amongst the teams that face the highest dip in win probability after making a decision.

Future Work

This work represents a basic implementation of a tool for NFL teams to make optimal fourth down decisions. In the future the win probability model, the PMFs, and the simulation model could be expanded in several different areas.

Win Probability Model and PMFs: While the win probability model generated good data and led to macro-level results that seemed reasonable, there were more results that seemed off than we had expected given the relatively low error rate of the win probability model during cross validation (i.e. error of +/- 1.27 percentage pts). The win probability model could be updated to better handle edge cases and to be accurate over all game states.

Simulation Model: Could be expanded to include a more convenient front end user interface to allow a general user to select game situations or enter their own and to interact with the model in a more user-friendly way. This model could also be extended to more easily analyze an entire set of game states such as an entire season to be able to compare and contrast performance of coaches/teams, their play decisions, and their outcomes relative to what the models suggest was the ideal decision.

Finally, the methodology could be expanded to take into consideration the unique attributes of the teams playing in the matchup in order to make it into a tool that could be used in real life.

Lessons We Have Learned:

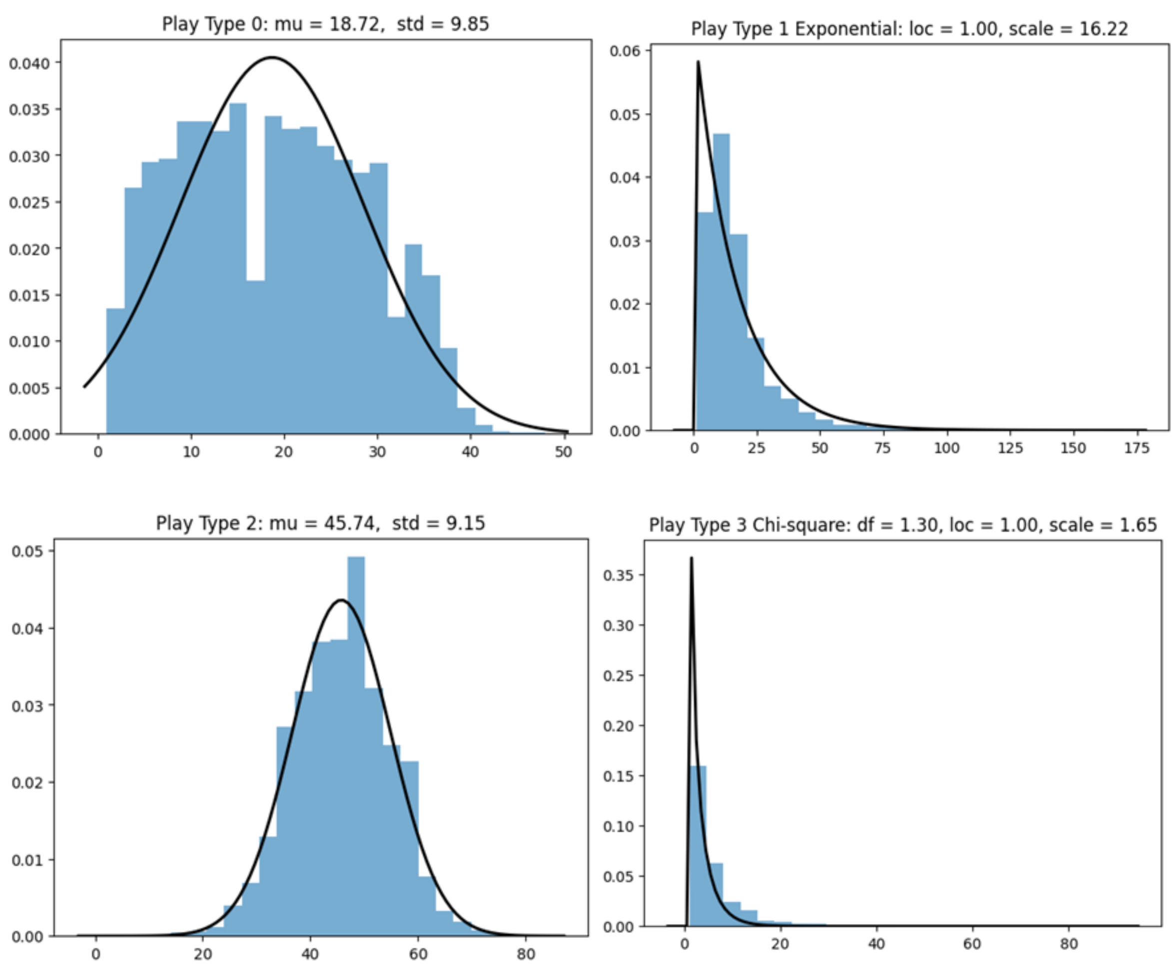
- Creating data models and abstract visualizations prior to development to ensure proper integration from one model to another with minimal hiccups
- Ensuring data loading and development environment are easy-to-use and everyone on the team is familiar with it
- Ensuring that we use a combination of domain knowledge on the game of American Football along with the data mining principles we’ve learned in class
- We chose to make this project wider by trying four different models in a sequential modeling pipeline however we could’ve also made the project deeper by trying to fine-tune a given model and getting it to perform excellently
- Given more time, we could have trained our model using different models than the ones we used and also used additional data going even before 2016

Appendix

Analysis of Probability Density Functions

The objective is to comprehend the distribution of game state results based on different play types. For play types, we have established that run, pass, punt, and field goal were our focus. At the start of the analysis, the distribution of the different play types was analyzed. The intent is to use the same distribution to make random state prediction of possibilities of occurrence. This led us to construct a probability density function (PDF) specifically for successful plays against the distance gained on the play. The usage of successful plays was selected with the assumption that only selecting successful plays instead of common plays would lead to better win probability. This was later proven wrong. For our analysis of PDFs, we continued on this path, and assessed the fit of each play type distribution to various theoretical models: Normal, Exponential, and Chi-square.

Figure – PDF sample by play type



Based on the analysis, the play type 'punt' and 'field goal' have distribution closer to normal, whereas play type 'run' and 'pass' provide a distribution closer to exponential. To ensure that we aren't just visualizing the naked eye, the Kolmogorov-Smirnov (KS) test was used to ensure this hypothesis was correct. For the normal distribution, the KS test compares the fit of the empirical cumulative distribution function (ECDF) of our play type data to the cumulative distribution function (CDF) of a normal distribution. For p-values that were all less than 0.05, we will reject the null hypothesis. Similarly for the exponential distribution, the KS test compares the fit of the ECDF of our play type data with the theoretical exponential CDF (1 - exp(-x)). Again, for p-values that were all less than 0.05, we will reject the null hypothesis. For our play type distribution analysis, 'punt' and 'field goal' were tested against normal distribution and 'run' and 'pass' were tested against exponential distribution.

Figure – Sample Kolmogorov-Smirnov (KS) Test Result

	Group	KS Statistic	P-Value
0	(0, 0, 1.0)	0.739496	2.395666e-67
1	(0, 0, 2.0)	0.689555	9.369665e-51
2	(0, 0, 3.0)	0.670831	1.049974e-58
3	(0, 0, 4.0)	0.781297	9.860105e-90
4	(0, 0, 5.0)	0.713664	1.273043e-74
...	...	...	...
155	(3, 3, 6.0)	0.924051	7.289819e-89
156	(3, 3, 7.0)	0.938108	4.714370e-91
157	(3, 3, 8.0)	0.981132	8.211243e-92
158	(3, 3, 9.0)	0.968750	9.363353e-97
159	(3, 3, 10.0)	0.982759	0.000000e+00

160 rows × 3 columns

References

Statista. (2020). Most watched sports leagues in the USA 2019. https://www.statista.com/statistics/1430289/most-watched-sports-leagues-usa/

Burke, B. (2014). Win Probability Prediction Model. Advanced Football Analytics. https://www.advancedfootballanalytics.com/2008/08/win-probability.html

Yurko, R., Ventura, S., & Horowitz, M. (2018). nflWAR: a reproducible method for offensive player evaluation in football. Journal of Quantitative Analysis in Sports, 15(3), 163-183. https://doi.org/10.1515/jqas-2018-0010

Lock, D., & Nettleton, D. (2014). Using random forests to estimate win probability before each play of an NFL game. Journal of Quantitative Analysis in Sports, 10(2). https://doi.org/10.1515/jqas-2013-0100

Sports Illustrated. (2017, June 28). Bengals Hire James Urban as Wide Receivers Coach. https://www.si.com/nfl/2017/06/28/nfl-james-urban-cincinnati-bengals-wide-receivers-coach

NFLFASTR. (n.d.). NFLFASTR. https://www.nflfastr.com/

NFL Data Py. (n.d.). PyPI. https://pypi.org/project/nfl-data-py/

NFLverse. (n.d.). NFLverse. https://nflverse.nflverse.com/

The Huddle. (2023, March 2). 2023 NFL Coaching Change Tracker. https://thehuddle.com/2023/03/02/2023-nfl-coaching-change-tracker/