

Learning to Go for It: Contextual Bandits, Policy Optimization, & Reinforcement Learning in NFL Fourth-Down Scenarios

Ardak Baizhaxynova, Mahima Batheja, Lucas Huynh, Mburu Kagiri

Andrew IDs: abaizhax, mbatheja, lqh, akagiri

Emails: abaizhax@andrew.cmu.edu, mbatheja@andrew.cmu.edu, lqh@andrew.cmu.edu,
akagiri@andrew.cmu.edu

Carnegie Mellon University, M.S. PPM- DA and HCAIT

From Data to Action

Peter Zhang

Heinz College of Information Systems and Public Policy

October 10, 2025

This final project report is submitted in fulfillment of the 94867 Final Project requirements.

Attribution

Ardak Baizhaxynova: PTO+ Bandit model evaluation, Authored files: model_evaluation.ipynb, Presentation slides 14-15, Final Report (Analysis section 4.2.4, 4.2.5)

Lucas Huynh: project formulation, proposal page, proposal slide, data pre-processing, feature engineering, PTO (multinomial logistic regression, ridge) + Greedy/LinUCB Bandit framework, designed IPS and DR evaluation, implemented Streamlit and artifact loading pipeline, created the GitHub repository

- Authored files: data(decisions_2016_2024.csv), data_clean_2016_2024.ipynb, behavior_2016_2024_epa.ipynb, behavior_2016_2024_wpa.ipynb, app.py, demo_det_gb.mov, team_logos, team_stadiums, artifacts (arm_models_epa.joblib, arm_models_wpa.joblib, inference.py, metadata.json, preprocessor.joblib, test_infer.py), GitHub README markdown file, requirements.txt
- Final presentation slides: 1, 8-13, 16, 29, 30
- Final report: report structure, cover page, executive summary (first 3 paragraphs), list of abbreviations, table of contents, analytics formulations (section 2), analysis (section 4.2; last three paragraphs of section 4.2.2; section 4.2.3; section 4.2.6), Appendix A (A.1-A.13) , Appendix A.1

Mburu Kagiri: Project proposal, Code evaluation, Presentation slides 1-7, Final Report (Analytics Formulation, Data Summary, Analysis Review, Document Preparation)

Mahima Batheja: feature engineering, inverse reinforcement learning model, evaluation

-Authored files: IRL.ipynb

-Final presentation slides: 16-26, 30-31

-Final report: analytics formulations (section 2.1 and 2.3), analysis (section 4.1.1; section 4.3), Appendix A (A.14-A.15)

Executive Summary

Over the past decade, analytics have shifted fourth-down strategy from heuristics toward quantified decision rules. This project aims to quantify and optimize those decisions (i.e., go for it, punt, attempt a field goal) through the lens of operational research topics including a machine learning-driven predict-then-optimize framework, multi-armed bandits (MABs), and reinforcement learning to improve decision-making under uncertainty. The project frames fourth-down decisions as a contextual bandit problem, where each possible action represents an “arm,” and the objective is to maximize either Expected Points Added (EPA) or Expected Win Probability added (WPA) given game context. Logged play-by-play data representing actual coaching behavior serves as the behavior policy from which new strategies are evaluated and learned. This off-policy framework systematically assesses when aggressive or conservative calls better align with long-term winning outcomes.

The dataset is drawn from nflverse’s play-by-play (PBP) data (via `nfl_data_py` API) covering detailed situational features for every fourth-down play between 2018 and 2024. After cleaning and filtering for relevant plays, we constructed a feature matrix of ~32,000 observations with key engineered variables encompassing game state, team-specific performance metrics, special teams context, and environmental factors. These features not only capture the static context of a play but also dynamic team form and environmental uncertainty, allowing the model to condition recommendations realistically.

The analytical workflow proceeds in three stages. First, behavior policy estimation models coaching decisions using multinomial logistic regression with ridge regularization to quantify the probability of each action given contextual features. Field position emerges as the dominant determinant, with the model achieving approximately 86% classification accuracy. Next, policy learning via contextual bandits applies ϵ -greedy and LinUCB algorithms to balance exploration and exploitation. LinUCB, in particular, estimates context-specific expected rewards (EPA/WPA) and yields interpretable coefficients that indicate when aggressive play-calling is analytically favorable. Finally, Inverse Propensity Scoring (IPS) and Doubly Robust (DR) estimators are employed for off-policy evaluation, reweighting historical data to simulate the performance of alternative strategies. Both bandit policies convert historically neutral or negative fourth-down situations into consistently positive-EPA outcomes, achieving roughly +1 EPA per play relative to existing coaching norms. In terms of win probability, LinUCB delivers the largest improvement (+0.05 WPA) but exhibits higher variance due to smaller effective sample size. In addition to the multi-year analysis, we ran a 2024 holdout evaluation using the same IPS + DR methodology, but focused on WPA for interpretability. The short-term results show a positive, but not statistically significant lift (~+0.006 WPA per play), indicating that the model’s decisions remain directionally effective in the 2024 season yet require more data to confirm reliability under current conditions.

The inverse reinforcement learning model uses multinomial logistic regression to define a reward function and to understand trade-offs made by coaches while making fourth down decisions. Coaches are more likely to undertake aggressive actions such as go-for-it when desperate in must score scenarios or when in offensive field position with high conversion rate such as 1 or less yard to go. Team abilities such as good recent offense performance and availability and trust in short range kickers also encourage coaches to shrug heuristics and make aggressive bets. However, scenarios such as short lead and playing in own territory increase likelihood of coaches exercising the safer punt option. While our model accurately predicts field goals and punts it finds it hard to predict “go for it decisions” particularly in tricky situations such as must score situations, short yardage in own territory or the fuzzy 45-55 yard range. These scenarios depend on coach and team specific factors that are hard to quantify.

List of Abbreviations

Abbreviation	Term
CI	Confidence Interval
DR	Doubly Robust
EDA	Exploratory Data Analysis
EPA	Expected Points Added
ESS	Effective Sample Size
FG	Field Goal
IPS	Inverse Propensity Scoring
IRL	Inverse Reinforcement Learning
LinUCB	Linear Upper Confidence Bound
MAB	Multi-Armed Bandit
NFL	National Football League
NGS	Next Gen Stats
OPE	Off-Policy Evaluation
PBP	Play-By-Play
PTO	Predict-Then-Optimize
UI	User Interface
WPA	Win Probability Added
ϵ -greedy	Epsilon Greedy

Table of Contents

Attribution.....	1
Executive Summary.....	2
List of Abbreviations.....	3
1. Problem Statement.....	4
2. Analytics Formulations.....	5
2.1. Feature-Engineered Metrics.....	5
2.2. PTO and Bandit Formulations.....	6
2.3. Inverse Reinforcement Learning Formulation.....	7
3. Data Summary.....	7
4. Analysis.....	8
4.1. Initial EDA.....	8
4.1.1 EDA for Inverse Reinforcement Learning.....	8
4.2. PTO and Bandit Algorithms.....	8
4.2.1. Predict Step.....	8
4.2.2. Additional Analysis of Driving Features.....	9
4.2.3. Optimize Step and Model Evaluation.....	11
Table 1. Summary and diagnostic data for EPA off-policy evaluation.....	12
Figure 6. Average policy value versus time step in an offline setting for bandit algorithms (EPA)....	
12	
Table 2. Summary and diagnostic data for EPA off-policy evaluation.....	13
Figure 7. Average policy value versus time step in an offline setting for bandit algorithms (WPA)....	
13	
4.2.4. Holdout Evaluation of 2024 Data.....	13
Table 3. Results of 2024 hold-out season evaluation (WPA).....	14
4.2.5. Comparison of historical and short term evaluation approaches.....	14
4.2.6. Inference, Limitations, and Assumptions.....	15
4.3 Inverse Reinforcement Learning.....	16
4.3.1. League Level Model.....	16
Table 4. Feature Importance: Feature Weights from Field Goal model, Go for It model, Punt model; Absolute average feature weight of all actions; feature weights for Go for It model trained on 2016-21 data and change in coefficients.....	17
4.3.2. Model Performance Evaluation.....	18
4.3.3. Temporal Validation.....	18
4.3.4. High Confidence Error Evaluation.....	18
Figure 11. Game situations where model accuracy is limited.....	19
4.3.5. Inference, Limitations, and Assumptions.....	19
Appendix A. Supplementary Material.....	21
Appendix A.1.....	27
A.1.1 Source Code, Detailed Framework, and Dataset.....	27
A.1.2 LLM Use.....	27

1. Problem Statement

Fourth-down decisions are among the most strategically consequential moments in an NFL game, often determining possession, field position, and ultimately game outcome. Yet despite their critical impact, these decisions have historically been guided by intuition, tradition, and conservative heuristics such as punting in “no man’s land” or “taking the points” with long field goals. Such momentum-based reasoning often undervalues the long-term expected gains of maintaining possession or pursuing high-leverage scoring opportunities.

A defining example occurred in Super Bowl LII, when the Eagles faced a late second-quarter 4th-and-goal against the Patriots. Conventional wisdom dictated a field goal attempt, a decision that would have preserved a roughly 45% win probability. Instead, analytics revealed a superior risk-reward profile for attempting a touchdown, given the down, distance, and the Patriot’s defensive assumptions. The now-iconic “Philly Special” play capitalized on this insight, resulting in a touchdown and boosting the Eagles’ win probability to nearly 70%.

This instance illustrates a broader reality: advanced analytics can transform descriptive insight into prescriptive confidence in decision-making. This can enable coaches to align tactical decisions with probabilistic models of success. This project aims to bridge the gap between data availability and its systemic integration into real-time decision making to maximize expected value given game context. The goal, therefore, is to formalize what traditionally has been instinct into a reproducible, quantitative foundation for strategic fourth-down decision-making in the modern NFL.

2. Analytics Formulations

2.1. Feature-Engineered Metrics

The following formulas use key notation: t (current week index in a season; integer), i (past week within a rolling window), p (current play index within a game), EPA_i (average offensive EPA per play in week i for the possessing team), and EPA_i^{opp} (average EPA allowed per play in week i by the opponent’s defense).

$$\text{off_epa_4w}_t = \frac{1}{n} \sum_{i=t-4}^{t-1} \text{EPA}_i, \quad \text{def_epa_4w}_t = \frac{1}{n} \sum_{i=t-4}^{t-1} \text{EPA}_i^{(opp)}, \quad n = \min(4, t - 1)$$

Equation 1. off_epa_4w_t captures how efficiently a team’s offense has performed over its past four games, reflecting short-term momentum and drive quality. A higher value suggests offensive strength.

and supports more aggressive “go” recommendations. def_epa_4w_t captures how the opposing defense has performed recently in preventing points or giving up field position. n number of weeks is constrained up to 4 if available. A higher value indicates a weaker defense and greater likelihood of conversion, tilting model toward aggression.

$$\text{fg_pct}_{\text{bin},t} = \frac{1}{n} \sum_{i=t-16}^{t-1} \text{made}_i, \quad \text{bin} \in \{\text{short, mid, long}\}$$

Equation 2. $\text{made}_i = 1$ if a field goal attempt in week i and distance bin was successful, else 0; n = number of recent field-goal attempts in the distance bin (up to 16 weeks). $\text{fg_pct}_{\text{bin},t}$ represents a 16-week rolling

field-goal make percentage for short (≤ 39 yds), mid (40-49 yds) and long (≥ 50 yds) attempts. It aims to capture special-teams reliability, with high values making the field-goal option more rewarding, especially when line of scrimmage corresponds to a kickable range.

$$\text{punt_net_4w}_t = \frac{1}{n} \sum_{i=t-4}^{t-1} \text{net_yards}_i$$

Equation 3. $\text{net_yards}_i = \text{start_yardline}_i + \text{next_yardline}_i - 100$ (clipped to [0,80]); start_yardline_i = distance from opponent’s end zone when the punt occurs; next_yardline_i = yardline where the receiving team’s next drive begins. punt_net_4w_t represents the 4-week rolling average of net punt yardage to illustrate field-position efficiency of the punting unit. Higher values favor punting in marginal

“no-man’s-land” areas around midfield where flipping field position is valuable.

$\text{def_share}_p = \frac{\text{cumulative defensive time}_p}{\text{game time elapsed}_p}$ *Equation 4.* cumulative defensive time_p = total seconds the defense has been on the field up to play p; game time elapsed_p = total elapsed game time up to play p; def_share_p represents the fraction of the game that a defensive unit has been on the field to capture defensive fatigue. A high value means the defense has been on the field longer, making conversions on “go” attempts more likely late in drives.

$\text{is_q4_or_later} = \begin{cases} 1, & \text{if quarter} \geq 4 \\ 0, & \text{otherwise} \end{cases}$ *Equation 5.* quarter = current quarter of the game (1-4). Is_q4_or_later represents the indicator variable for late-game or end_game situations to reflect game urgency. In the 4th quarter, maximizing win probability becomes critical, often shifting the decision boundary toward aggression (“go for it” instead of punt).

$\text{in_own_territory} = [\text{yardline}_{100} > 60]$ *Equation 6.* in_own_territory = current field position more than 60 yards from opponent’s endzone. Field position explains why certain fourth down decisions may have been chosen. Go for it is far more risky when team is deep in their own territory.

$\text{very_short} = [\text{ydstogo} \leq 1]$ *Equation 7.* very_short = binary flag indicating whether a team needs 1 yard or less a first down. This is considered “go for it” territory with a high conversion rate.

$\text{close_game_late} = [|\text{score_differential}| \leq 7 \wedge \text{qtr} \geq 3]$ *Equation 8.* close_game_late = when the score difference is 7 points or less in the second half. Close games are high-pressure situations where failure due to a risky play may result in the coach being blamed in the media narrative and risking career.

$\text{must_score} = [\text{score_differential} < 0 \wedge \text{time_remaining} < 300 \wedge \text{qtr} = 4]$ *Equation 9.* must_score = when the team is behind on points and with less than 5 mins left in Q4. It may require coaches to take unconventional fourth down decisions and do whatever it takes to not lose.

$\text{protecting_lead} = [0 < \text{score_differential} \leq 7 \wedge \text{time_remaining} < 300]$ *Equation 10.* protecting_lead = when the team has a lead of 7 points or less and with less than 5 mins. A slim lead may require coaches to go for safer decisions. Risk aversion may be a product of loss aversion bias and preventing criticism.

$\text{home_field_advantage} = [\text{posteam} = \text{home_team}]$ *Equation 11.* home_field_advantage = team with ball possession is the home team. A home crowd can boost morale. Coaches may not want to disappoint home fans and may feel pressured to take a specific decision.

$\text{is_outdoors} = [\text{roof} = \text{"outdoors"}]$ *Equation 12.* roof = the stadium has no roof. Playing in an open stadium results in weather affecting fourth down strategies.

2.2. PTO and Bandit Formulations

$\pi_b(a | x) = \frac{\exp(\beta_a^\top x)}{\sum_{a' \in A} \exp(\beta_{a'}^\top x)}$ *Equation 13.* $a \in A = \{\text{Go, Punt, FG}\}$: discrete fourth-down decisions; x = feature vector (yardline, distance to go, score differential, time remaining, team form, special teams metrics, weather, etc.); $\pi_b(a | x)$ = estimated probability of taking action a in context x ; β_a = coefficient vector for action a , learned via maximum likelihood estimation. The multinomial logistic regression models the probability that a coach selects a particular fourth-down action given contextual features. The fitted coefficients capture how each feature influences likelihood of selection of the potential actions. It serves as the behavior policy required for computing propensity score in off-policy evaluation (IPS, DR).

$\hat{r}_a(x) = x^\top \theta_a$ where $\theta_a = \arg \min_{\theta} \|y_a - X_a \theta\|^2 + \lambda \|\theta\|^2$ *Equation 14.* y_a = vector of observed rewards (EPA or WPA) for action a ; X_a = feature matrix for plays where action a was chosen; θ_a = learned parameter vector for action a ; $\lambda = 5.0$, ridge regularization penalty controlling coefficient shrinkage; $\hat{r}_a(x)$ = predicted reward for action a in context x . Ridge regression estimates expected reward for each action as a linear function of contextual features and is trained per action type, penalizing large coefficients to reduce overfitting and smooth noisy outcomes. Estimates form the foundation for bandit optimization.

$a^*(x) = \arg \max_{a \in A} \hat{\mu}(x, a)$ *Equation 5.* $\hat{\mu}(x, a)$ = predicted expected reward from ridge regression; $a^*(x)$ = selected action maximizing $\hat{\mu}(x, a)$. Greedy policy (pure exploitation) always selects the action with the highest predicted expected reward, representing a fully deterministic strategy that optimizes expected value without accounting for uncertainty. Useful as a baseline comparison against exploration-based approaches.

$\text{Score}(x, a) = \hat{\mu}(x, a) + \alpha \sqrt{x^\top V_a^{-1} x}$ *Equation 16.* $\hat{\mu}(x, a)$ = predicted reward for action a ; $V_a = X_a^\top X_a + \lambda I$ = regularized covariance matrix for action a ; α = exploration parameter controlling optimism; x = contextual feature vector. LinUCB adds an uncertainty bonus to predicted reward, allowing exploration of actions with

uncertain value. Allows for risk-aware decision-making under limited information. $\alpha = 0.8$ and $\lambda = 0.5$ used.

$$\hat{V}_{IPS} = \frac{1}{N} \sum_{i=1}^N \frac{\pi(a_i | x_i)}{\pi_b(a_i | x_i)} r_i$$

Equation 17. $\pi(a_i | x_i)$ = probability of action a_i under the target policy; $\pi_b(a_i | x_i)$ = probability of action a_i under behavior policy; r_i = observed reward for play i ; N = total number of 4th down plays used for evaluation. IPS used to estimate the average reward under a new policy by reweighting historical plays based on likelihood of new policy taking each action. High variance seen when target/behavior policies differ.

$$\hat{V}_{DR} = \frac{1}{N} \sum_{i=1}^N \left[\frac{\pi(a_i | x_i)}{\pi_b(a_i | x_i)} (r_i - \hat{\mu}(x_i, a_i)) + \sum_{a \in A} \pi(a | x_i) \hat{\mu}(x_i, a) \right]$$

Equation 18. Variables same as IPS with additional $\hat{\mu}(x_i, a)$ = predicted reward from ridge regression for action a in context x_i . DR estimator combines IPS weighting with model-based predictions to achieve bias-variance robustness. It is unbiased if either the behavior model (π_b) or the reward model ($\hat{\mu}$) is correct. It is more stable than IPS, particularly for rare actions.

$$ESS = \frac{\left(\sum_{i=1}^N w_i \right)^2}{\sum_{i=1}^N w_i^2}, \quad w_i = \frac{\pi(a_i | x_i)}{\pi_b(a_i | x_i)}$$

Equation 19. $\pi(a_i | x_i)$ = target policy probability; $\pi_b(a_i | x_i)$ = behavior policy probability; w_i = importance weight for play i ; N = number of evaluation samples. Effective Sample Size (ESS) measures the representativeness of logged data under a new policy. Smaller ESS indicates fewer “effective” samples and higher variance, signaling low overlap between old and new policies.

2.3. Inverse Reinforcement Learning Formulation

$$R(s, a) = \theta_a^T \phi(s) = \sum_{j=1}^{17} \theta_{a,j} \cdot \phi_j(s)$$

Equation 20. $R(s, a)$ = estimated reward given state-action pair; a = fourth down action - go for it, field goal, punt; s = game situation before fourth down action; $\phi(s) \in \mathbb{R}^{17}$ = feature vector representing state s ; $\theta_a \in \mathbb{R}^{17}$ = learned weight vector for action a , j is feature index

3. Data Summary

Play-by-play data was collected using the open-source nfl_data_py API (a wrapper around the NFLverse repository). All regular-season plays from 2016-2024 were imported, resulting in a raw dataset containing 416,321 play-level observations, with each row representing a unique play and over 300 contextual variables. Our analysis focuses on 4th-down decision context, treating each valid play as a single “bandit round.” After filtering, the final decision-level dataset contains 31,849 plays spanning all 32 NFL teams and nine seasons.

Data cleaning, filtration, and type conversion steps were implemented to ensure integrity and applicability. Improper outcomes and outlier events that could derail the modeling were omitted, and associated actions were grouped into our 3 outcomes {punt, go, fg} for multivariate logistic regression. The final dataset distribution, which contained no null values, was 17,783 (55.8%), 8,388 (26.3%), and 5,678 (17.8%) rows of punt, field goal, and go, respectively.

To better model team tendencies and situational context, several rolling and cumulative metrics were constructed (see section 2.1) for recent offensive/defensive efficiency, field goal conversion rates, net punt averages, and fatigue/urgency indicators. These engineered features were merged back into the main play-by-play DataFrame using season and team keys.

The final modeling table (see Appendix A, Table 1) used for contextual bandit modeling contains 58 total variables across context, team form, special teams, fatigue context, and outcome/reward signals. Each row therefore encodes the full decision context x , observed action $a \in \{\text{go, punt, fg}\}$, and resulting reward r (EPA/WPA) as the foundation for behavior policy and reward models.

4. Analysis

4.1. Initial EDA

Analyzing the data corroborated conventional expectations of field behavior and pragmatic decision-making. As Figure 1 highlights, field positioning influences the probability of taking any particular action, with punting options dominating when a team is deep in its own half, and conversely field goal attempts deep in the opponent's half. The data provided the field contexts resulting in any particular action, such as the time remaining to play the game, the field positioning, the distance to a first-down marker, the score differential, the expected points added, and the win probability added. Additionally, to enhance feature engineering, the data also included metrics on a 4-week rolling average of a team's efficiency, the historical performance of a team's special teams (the unit traditionally tasked with 4th down punt or field goal execution), the time-specific cumulative fatigue and situational stress, and the raw game state. Together, these datapoints offered a holistic snapshot of the relevant factors influencing the decisions.

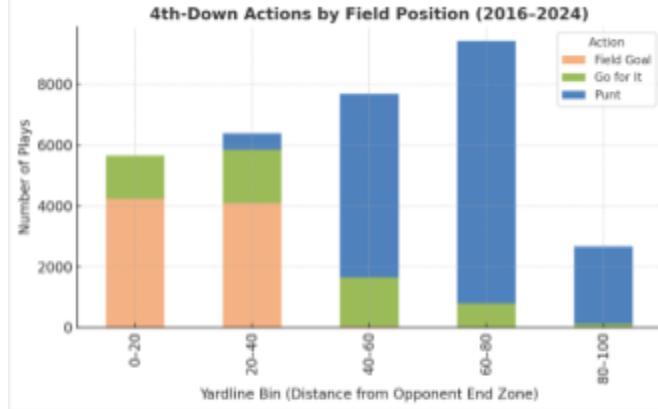


Figure 1. 4th-down actions by field position; punts dominate in own territory, while field goals and go for it attempts increase closer to endzone

4.1.1 EDA for Inverse Reinforcement Learning

To select appropriate features as candidate reward features for the IRL model. We undertook spearman correlation analysis and multicollinearity test using variance inflation factor. The correlation analysis helped us filter out qtr, game_seconds_remaining and defteam_timeouts_remaining. These features were correlated with themselves and other features. They also had high multicollinearity. We did not eliminate yardline_100 despite $r=-0.8$ relation with in_own_territory as the two variables capture two different things. Yardline_100 is a mechanical feature that is critical for field goal decisions. On the other hand, in_own_territory is a psychological feature that represents field position safety. This affects the decision to punt or go for it. Another reason to keep both the features was that we did not find high multicollinearity between them. We are assuming standard L2 regularization settings ($c=1$) should handle the strong negative correlation. Is_outdoors was found to have a high variance inflation factor (potentially with temperature and wind) and was eliminated from the feature matrix (see annexure A Figure A14 and A15).

4.2. PTO and Bandit Algorithms

4.2.1. Predict Step

Fourth-down decisions are first modeled as behavioral probabilities using a multinomial logistic regression with ridge regularization (see section 2.2. for more detailed formulation) and a shared preprocessing pipeline (i.e., median-impute, standardizing numeric features, one-hot encoding categorical features via ColumnTransformer, etc.). Inputs include field positions, yards-to-go, clock and score state, rolling team form, special-teams indicators, fatigue/drive context, timeout, and basic weather and field conditions. The model outputs the logged policy $\pi_b(a|x)$ needs for IPS/DR and captures historical tendencies and classes accurately. The in-sample accuracy of 0.857 means that the model is predicting the observed action of coaches about 86% of the time, and the features being fed in explain most of the

variance in play-calling. The distribution of logged probabilities shows high confidence on most snaps (mean $p_b \sim 0.786$) with occasional low-probability “surprise” calls.

Coefficient magnitudes (mean $|\beta|$ across actions) show field position as the dominant driver in historical coaching decisions (Figure 12). As a team gets further away from the opponent end zone, the log-odds shift strongly toward punt ($\beta = 3.05$) and away from FG ($\beta = -2.95$) with a small negative effect on go. This exactly matches the traditional logic of decisions for long FG and protecting field position. Clock and distance are the next most important factors. More time remaining promotes conservative actions (punt = 1.19, FG = 0.27, go = -1.46) and a longer distance-to-go for a successful conversion depresses go (-1.23) in favor of FG (0.69) and punt (0.54). We see the same pattern with score differential, as with a lead, coaches protect possession (punt = 0.45, FG = 0.20) rather than chase conversions (go = -0.65).

Fatigue/drive context has a measurable but smaller influence in comparison. More defensive time on the field is associated with more conservative calls in games where the offense is already managing a lead or playing longer sequences. Goal-to-go situations in short-yardage near the endzone flips that, as go (0.53) rises and punt drops (-0.57). Obvious environmental effects factor in as well. Outdoors reduces FG odds (-0.28) and nudges toward punt (0.46), as FG kickers tend to struggle with high winds for example. Finally, some team-specific one-hot encoded features show deviations from reference level. Particularly, PHI (Philadelphia Eagles) shows relative aggressiveness (go = 0.34) as seen with the “Philly Special.” While these metrics are descriptive rather than causal, they provide a summary that the reward and OPE models build upon.

Expected rewards are then modeled per action using ridge regression on the same design matrix. There is one linear model for each arm $a \in \{\text{go, punt, fg}\}$ with L_2 penalty $\lambda = 5$ as a constant fallback for scarce samples. Regularization allows for variance control, as shrinking coefficients towards zero reduces the effects of noise and prevents unstable coefficients that could later inflate the DR estimator. It also handles multicollinearity since correlated features (i.e., yardline, yards-to-go) are common. L_2 stabilizes the joint effects and improves conditioning of the inverse. Using a moderate lambda provides a smooth coefficient across actions. A lambda close to zero approaches OLS (low bias, high variance) while a limit towards infinity collapses to the intercept (high bias, low variance). The fallback for scarce samples is used (i.e., rare arms such as mid-field FGs) where if the number of training rows for an arm falls below a minimum threshold of $n = 50$, the model for that arm is set to a constant equal to the arm’s sample mean reward. This allows for protection against overfitting noisy data with a trade-off of small bias in return for large variance reduction. The DR estimator (discussed later) is robust as long as the behavior model is well specified.

Both EPA and WPA use an identical training setup with the same features, preprocessing, model class, and hyperparameters. They only differ in the target vector so that downstream policy optimization (Greedy/LinUCB) and OPE (IPS/DR) operate with consistent rewards.

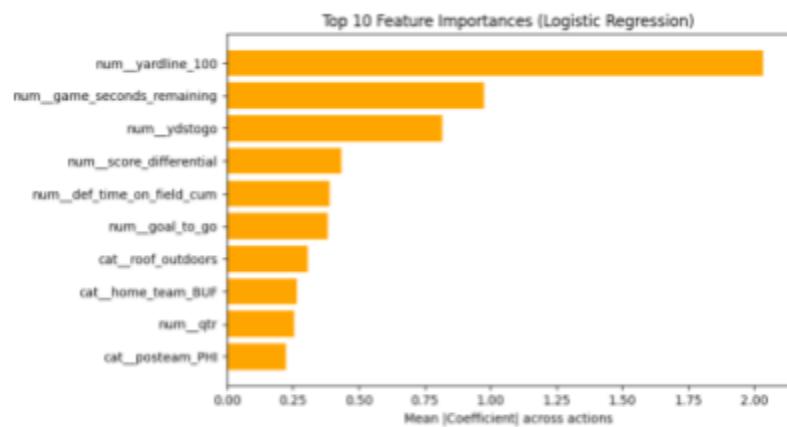


Figure 2. Multivariate Logistic Regression Model Key Determinants of Real-Time Decision Making

4.2.2. Additional Analysis of Driving Features

Feature engineering also revealed the key determinants of decision making, isolating field positioning, the time remaining on the clock, and yardage to a first-down marker as the most pertinent considerations (Figure 2). Importantly, this insight suggested that 4th-down strategies

were situational, rather than team specific; coaching behavior was less influenced by the opposing team.

Given the outsized role of field positioning on influencing decisions, we performed further analysis to see how this influenced the EPA and WPA.

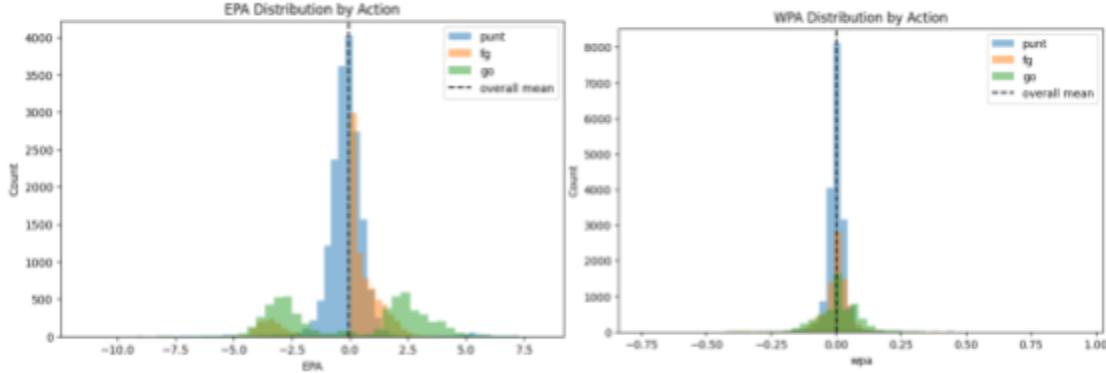


Figure 3. For both EPA and WPA, “go for it” instances show greater variance, indicating higher risk-higher reward outcomes

We observed varying risk profiles associated with each action. Generally, attempted field goals within range resulted in greater EPA additions with comparatively minimal risk. However this strategy, although a safer option, did not lead to sizable WPA gains (Figure 3). Conversely, go attempts indicated greater risk fluctuations (wider variance), with a multimodal distribution indicating higher EPA gains upon successful conversions, and costly EPA losses in failed conversions. Expectedly, punt attempts did little to influence WPA and marginally reduced EPA points.

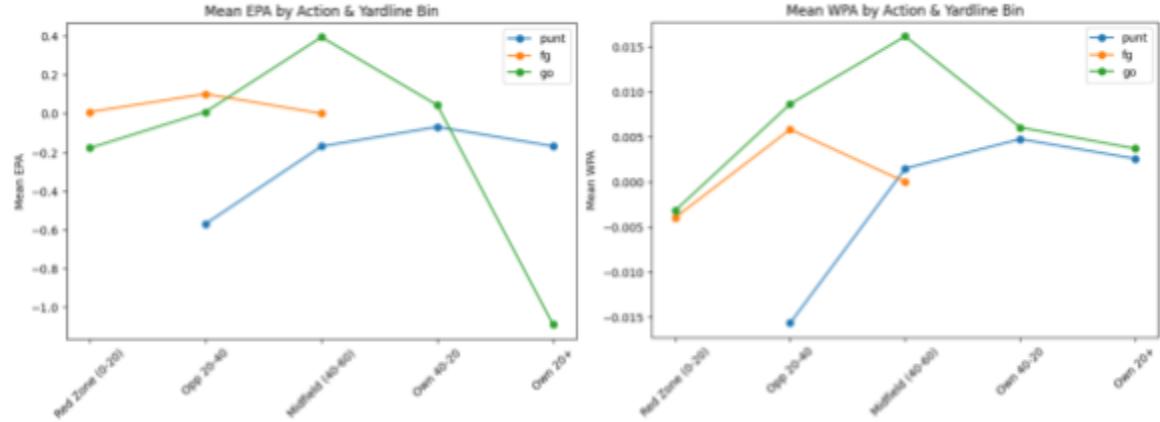


Figure 4: “Going for it” maximizes both EPA and WPA near the the midfield.

In keeping with expected results, go attempts near the midfield area yielded the greatest WPA and EPA values (Figure 4). Interestingly, go attempts near the opponent’s endzone had an on-average negative EPA, which reveals deeper insight into Figure 3’s variance. Despite the multimodal distribution, the cost profiles were drastically different. Failed go conversions held a heaving penalty on EPA and WPA

probabilities than successful conversions, explaining the leftward skew in values that Figure 4 illuminates. It is also noteworthy that although go attempts have noticeable losses in EPA when deep into one's territories, go and punt strategies had near-identical gains (marginally low to begin with) in impacting WPA. This is a surprising finding, since the two strategies lead to vastly different positioning upon successful conversion.

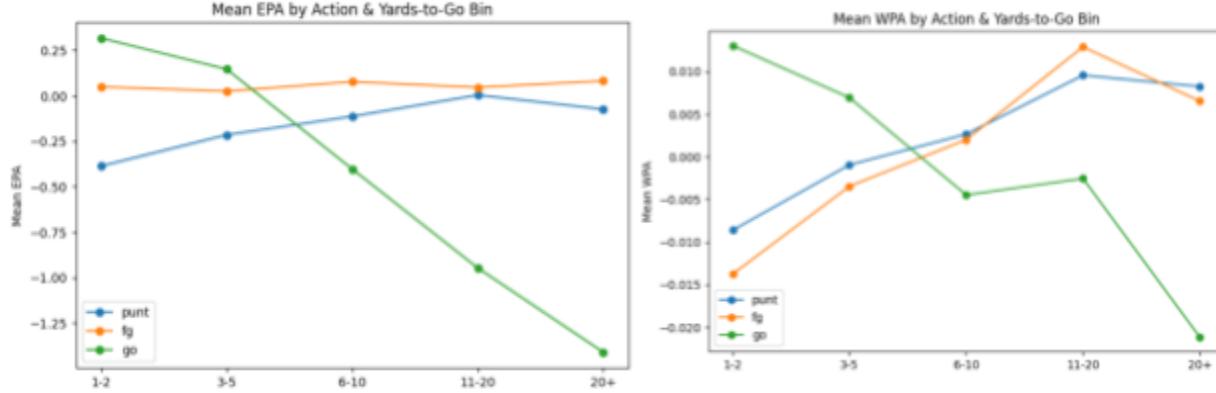


Figure 5: Distance to a 1st-Down marker directly influences the EPA and WPA gains for “going for it”

Positioning is tangentially related to the distance away from a first down; 5 yards away from a first down deep into one's territory carries distinguishable risks from 5 yards away from a first down deep into the opponent's territory. As expected, the success rates of go attempts are highly dependent on the distance from a first down, with a general linear correlation between distance and EPA/WPA gains (Figure 5). More in-depth decision tables and numerical metrics for the figures above can be found in the appendix (A.1-A.4 for EPA, A.5-A.8 for WPA).

4.2.3. Optimize Step and Model Evaluation

Two target policies are derived from the per-action reward models (ridge, $\lambda = 5$) for decision optimization: (1) Greedy that selects the highest expected reward arm and (2) LinUCB which adds an uncertainty bonus to the predicted reward using $\alpha = 0.8$ (large enough to surface high-value but under-sampled contexts without overturning greedy preferences). Per-arm parameters come from the ridge fits, and Greedy ignores uncertainty. LinUCB, on the other hand, trades some bias for robustness in regions with scarce data by favoring exploration of actions whose context has been poorly explored historically.

Policies are then evaluated off-policy with IPS and DR. IPS reweights logged outcomes and DR adds a model-based control variate to reduce variance and remain consistent if the behavior model is correctly specified (see section 2.2. for a more formal representation). The behavior policy comes from the multinomial logistic regression, and a standard sanity check confirms DR's calibration in that $\pi_e = \pi_b$ reproduces the same logged mean reward. ESS reported for IPS quantifies how much independent information the weighted sample contains and explains interval width.

95% confidence intervals are obtained via $B = 100$ bootstrap resamples of plays with replacement, as the DR estimate is recomputed per resample with fixed π_b and $\hat{\mu}$, and percentile bounds at 2.5% and 97.5% are reported.

EPA policy values per 4th down show that both Greedy and LinUCB learned policies convert historically neutral/negative situations into positive EPA decisions:

	DR EPA Est.	EPA 95% CI	IPS EPA	ESS
Greedy	0.938	[0.472, 1.681]	1.00	33.6
LinUCB	0.915	[0.509, 1.518]	0.995	29.0

Table 1. Summary and diagnostic data for EPA off-policy evaluation.

DR estimates for both policies are near +1.0 EPA per play compared to the logged mean of -0.05 EPA. IPS and DR agree closely, which indicates good overlap between recommended and historical actions. CIs exclude zero, indicating statistically meaningful improvement over the logged mean. They remain fairly wide, however, due to ESS $\sim 29\text{-}34$ being small. This is typical when an evaluation policy deviates from history and reweighting concentrates on fewer comparable plays.

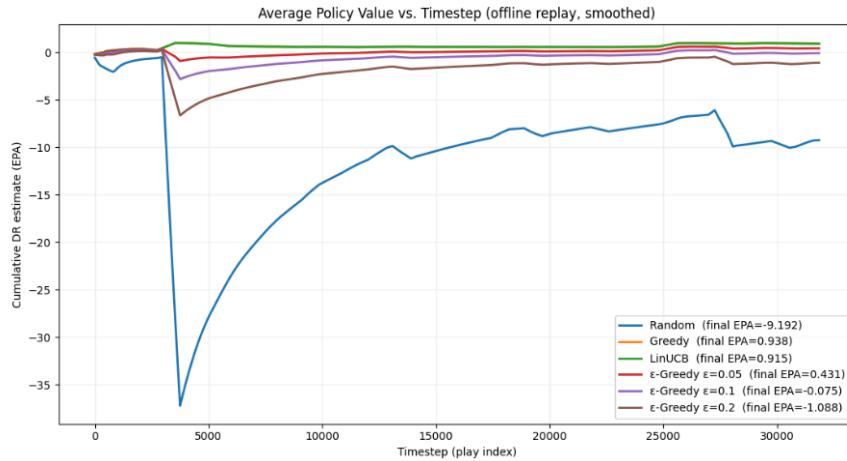


Figure 6. Average policy value versus time step in an offline setting for bandit algorithms (EPA).

The timestep graph for the EPA model corroborates our previous findings. Greedy and LinUCB steadily outperform random policies, and added offline exploration (large ϵ) degrades value in offline data by forcing mismatched actions. Injected randomness appears to lead to systematically worse play calls and slower convergence. Since pure exploitation (Greedy) performs about as well as Lin UCB's confidence-based exploration, it suggests that in 4th-down contexts, the best action is usually already clear.

The EPA lift is driven by systematic choices toward “go” at midfield (40-60 yards away from the opponent’s endzone) and short yardage (1-5 yards), both regions where EDA shows that “go” dominates and routine punts are costly. Since both Greedy and LinUCB policies concentrate mass in these high-value “pockets,” their EPA improvements are similar.

In terms of WPA, LinUCB’s exploration offers greater long-run upside with higher short-term uncertainty. Improvements compared to the baseline of 0.003 WPA are as follows:

	DR WPA Est.	WPA 95% CI	IPS WPA	ESS
Greedy	0.0215	[0.0111, 0.0294]	0.0157	314
LinUCB	0.0597	[0.0037, 0.1562]	0.0609	43.5

Table 2. Summary and diagnostic data for EPA off-policy evaluation.

While WPA gains are smaller in magnitude compared to EPA, they are directionally consistent. In interpreting these values into context, following the greedy policy would increase win probability by ~2% for every 100 plays and LinUCB would increase win probability by ~6% for every 100 plays. Greedy has a smaller CI and high ESS ~ 314 , which signals strong overlap with historical endgame behavior. LinUCB, in contrast, has a much wider CI and smaller ESS ~ 43.5 , meaning that the higher variance is due to the bandit departing more from historical coach decisions in high-stakes but rare contexts.

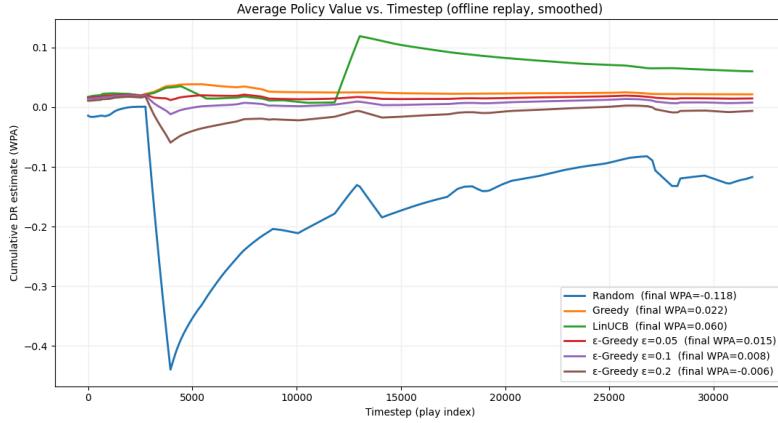


Figure 7. Average policy value versus time step in an offline setting for bandit algorithms (WPA).

The offline curve for WPA demonstrates similar trends to the EPA, with one key difference. LinUCB achieves the best overall value, with a noticeable early surge followed by slight regression as play mix shifts. Limited, structured exploration with LinUCB outperforms pure exploitation (Greedy) in identifying high-value plays.

4.2.4. Holdout Evaluation of 2024 Data

We also performed 2024 holdout evaluation, by excluding year 2024 for testing, instead of all historical data as in the abovementioned analysis. This evaluation followed the same methodology of using Doubly Robust (DR) method, which combines two complementary ideas: (1) propensity scores (how likely the logged behavior was to choose each action) and (2) direct reward models (predicted win probability added, WPA). For short term analysis only WPA was considered as it is easier to interpret and is the ultimate goal for the coach.

Estimator	Mean Estimate	95% Confidence Interval	Notes
Direct Method (DM)	0.019459	[0.019037, 0.019943]	Relies only on reward models; optimistic but stable.
Inverse Propensity Score (IPS)	-0.003502	[-0.033645, 0.018612]	High variance; negative mean suggests worse than baseline.
Self-Normalized IPS (SNIPS)	-0.002789	[-0.021899, 0.017847]	Variance-reduced IPS; similar negative tendency.
Doubly Robust (DR)	0.006058	[-0.026982, 0.029934]	Most reliable; indicates small positive lift, but not significant.
Effective Sample Size (ESS)	24.7 / 3634	—	Very low ESS, limiting confidence in estimates.

Table 3. Results of 2024 hold-out season evaluation (WPA)

Our results show that the new ϵ -greedy policy achieved a DR estimate of $\sim +0.006$ WPA/play, about $+0.004$ higher than the baseline behavior policy, though the 95% confidence interval included zero, so the overall lift is not statistically significant yet. The Effective Sample Size (ESS) was only ~ 25 out of 3,634 plays, indicating that although the dataset is large, only a very small fraction of plays provide meaningful weight for the evaluation. Slice analysis revealed credible gains between the opponent's 20–40 yard line and near the red zone, but instability (low ESS) when deep in their own territory. This low ESS substantially reduces the precision of the estimates. As a result, the observed improvements should be interpreted with caution, as they may be driven by limited effective data rather than consistent underlying effects.

In simple terms: we tested whether our decision model makes better 4th-down calls than real coaches, using a method that double-checks itself. On average it looked a bit better (about $+0.4\%$ more win probability per play), especially in midfield situations, but the evidence isn't strong enough yet to declare it a clear improvement everywhere.

4.2.5. Comparison of historical and short term evaluation approaches

To validate our 4th-down decision policy, we conducted Off-Policy Evaluation (OPE) using the Doubly Robust (DR) estimator on both the full 2016–2024 dataset and an independent 2024 holdout season. The full evaluation shows the average historical gain across eight seasons, while 2024 hold-out shows the current-season generalization and reliability. Together, they demonstrate both historical strength and forward validity, which is the primary goal of reviewers.

Across all seasons, Greedy and LinUCB policies achieve roughly $+1$ Expected Point Added (EPA) and $+0.02$ – 0.06 Win Probability Added (WPA) per 4th-down, with confidence intervals excluding zero, which confirms a statistically significant improvement over historical coaching decisions. The large-sample analysis thus establishes strong average performance and shows that structured exploration (LinUCB) can outperform pure exploitation (Greedy).

Our 2024 holdout evaluation reproduces the same methodology on unseen data to test temporal generalization. Using ϵ -greedy and greedy variants, we obtain a DR estimate of $+0.006$ WPA per play ($\approx +0.6$ percentage-point win-probability lift). While the confidence interval includes zero (not statistically

significant), the model still directionally improves field position matches suggested by multi-year results. Namely, positive impact from midfield to the opponent’s red zone and unstable estimates deep in their own territory (low ESS). This consistency supports the robustness of our learned policy under modern-season conditions.

Together, these evaluations show that analytics-driven 4th-down strategies not only outperform legacy heuristics across eight years of data but also generalize to the 2024 NFL season, maintaining directionally positive effects on win probability.

4.2.6. Inference, Limitations, and Assumptions

To prepare the bandit models to be translated from back-end to front-end, saved artifacts for the preprocessor, per-arm EPA and WPA ridge models, and feature list metadata were loaded. For a given single play context, a one-row DataFrame was built with the expected columns, preprocessing during training was applied, and per-action rewards for each arm a were predicted. Due to time constraints and for the sake of simplicity, the recommended action follows the Greedy recommendation (arg-max of chosen objective: EPA or WPA), and the function returns both EPA and WPA scores.

Several assumptions were made to apply feasibility constraints to the model before selecting an action. For example field goals beyond a reasonable distance are suppressed, with an allowable range of 17-65 yards away from the endzone. The following adjustments are made for possible FG range:

1. In stadiums featuring domes, the distance cap is increased by 3 to quantify the diminished role of weather conditions.
2. In stadiums without domes, the presence of strong winds (≥ 15 mph) and/or very cold temperatures ($\leq 20^{\circ}\text{F}$) decreased the cap by 3 yards each.

Punts are marked as infeasible in two common cases:

1. If within 35 yards of the endzone, punting is rendered obsolete given the possibility of a FG.
2. If the time remaining in the game constitutes 5 or less minutes and the team is losing, punting is essentially an affirmation of guaranteed loss and is thus not a viable option.

Infeasible arms are masked by setting predicted values to an incredibly small sentinel (i.e. 1×10^{-9}) so they cannot be selected while raw scores are still reported.

The Streamlit front-end application collects game state and automatically updates the team-week features from the cleaned dataset (i.e., rolling offensive/defensive EPA, FG% bins, punt net average). Venue defaults (roof/surface) and stadium imagery are populated from team maps but can be overridden by the user. The app converts user inputted field position to distance from the endzone, derives time remaining in the game, maps timeouts by possession, and passes the full context to the scorer. The user is also able to see metrics representing the current form of the team currently in possession of the ball. The UI then surfaces the top recommendation (for EPA or WPA), the margin over the next-best feasible option, and any exclusions (e.g. messages include “FG out of realistic range” and “Punt not considered here”). The user is able to choose on the sidebar which metric to optimize.

Some assumptions and limitations are also made in the implementation of the bandit model. If team-week metrics are missing (i.e., in the first four weeks of the season), the model defaults to neutral historical averages (short FG accuracy = 88%, medium FG accuracy = 75%, log FG accuracy = 55%), net punt = 42 yards). In terms of weather defaults, if the game is played in a dome, temperature and wind

default to 70°F and 0 mph, respectively. Otherwise, they default to 60°F and 5 mph. Neutral fatigue placeholders are also provided when not collected live (i.e., plays in the drive set to 3, defensive share of time on field set to 50%). Finally, after masking infeasible arms (i.e., infeasible range FGs and analytically dominated punts), feasible actions are ranked and the top option is recommended. If the initial arg-max was masked for any reason, the model recommends the highest-valued feasible and alternative and the user is made aware of any exclusions.

Detailed visualizations of the application can be viewed in the appendix (Figure A.9-A.13). The input fields are designed to model an end-of-game 4th-down scenario in an actual game played between the Baltimore Ravens and Buffalo Bills earlier this year. Interestingly enough, the Ravens rebuffed the advice of their analytics team by deciding to punt the ball away near midfield in a high-scoring affair. They sent their defense back on the field to attempt to stop Josh Allen and the Buffalo Bills offense that had been rolling all night. The Bills drove down the field into FG position and kicked a last-minute game-winning FG.

4.3 Inverse Reinforcement Learning

4.3.1. League Level Model

We averaged the coefficients after running the multinomial logistic regression model with default L2 regularization of 1.0 on each fourth down action. Our mean coefficients revealed that the coach's fourth down decision is influenced most by desperation. Desperation as reflected by 'must_score' makes coaches more likely to opt for the aggressive "go for it" action that could possibly help the team win. Even if the action were to fail in generating more points, coaches would not be blamed for not taking steps to avoid a loss. Interestingly, lower levels of desperation demonstrated by 'close_game_late' or situations with no desperation demonstrated by 'score-differential' do not influence coach's decision making.

When in a scenario where the team has a slim lead, coaches are more likely to play it safe by opting for a punt. Log odds of punting increase by 0.83x.

Feature	FG	GO	PUNT	Avg	2016-21 (GO)	Change
in_own_territory	-0.879	+1.460	-0.582	0.973	+1.462	0.002
yardline_100	-0.124	-0.039	+0.163	0.108	-0.039	0
very_short	-1.507	+2.431	-0.924	1.621	+2.311	0.120
score_differential	+0.014	-0.050	+0.036	0.034	-0.050	0
must_score	+0.260	+2.307	-2.567	1.711	+2.299	0.008
protecting_lead	-0.152	-0.678	+0.830	0.554	-0.483	0.195
close_game_late	+0.241	-0.217	-0.025	0.161	-0.279	0.062

posteam_timeouts_remaining	-0.222	-0.195	+0.417	0.278	-0.233	0.038
temp	+0.005	-0.006	+0.001	0.004	-0.007	0.001
wind	-0.008	+0.005	+0.003	0.006	+0.002	0.003
off_epa_4w	-0.508	+0.745	-0.237	0.497	+0.876	0.131
def_epa_4w	-0.216	+0.477	-0.261	0.318	+0.424	0.053
fg_pct_short	+1.042	+0.744	-1.786	1.191	+0.773	0.029
fg_pct_mid	+0.140	-0.088	-0.052	0.093	-0.001	0.087
fg_pct_long	+0.172	+0.067	-0.105	0.115	+0.031	0.036
punt_net_4w	-0.005	+0.006	-0.001	0.004	+0.001	0.005
home_field_advantage	+0.025	+0.045	-0.070	0.046	+0.092	0.047

Table 4. Feature Importance: Feature Weights from Field Goal model, Go for It model, Punt model; Absolute average feature weight of all actions; feature weights for Go for It model trained on 2016-21 data and change in coefficients.

Similarly specific field positions like short yardage (very_short) and in_own territory have greater sway on fourth down decision choices than the generic field position measure ‘yardline_100’. When yardage is less than 1, coaches have high odds of going for it. They are highly unlikely to attempt a field goal or punt. Similarly, when the team is deep in its own territory most coaches usually punt (see Figure 8) but a small set take an aggressive approach and “go for it”.

Go decisions in own territory: 570
Action distribution IN own territory:
action
punt 0.928972
go 0.071028
Name: proportion, dtype: float64

Figure 8. Proportion of in own territory decisions that were punt or go-for-it

Coaches are more selective in factoring team strengths in their fourth down action. When the team has demonstrated good offensive capabilities in prior games coaches opt for aggressive strategies. The odds of ‘go-for-it’ becomes 0.745 times. Similarly, having a field goal kicker with good short distance accuracy increased field goal log odds by 1.042 times. However, having a well performing punter did not increase log odds of punt action.

Interestingly, temperature and wind do not affect fourth down decisions of coaches. However, they do affect field goal and punting accuracy in real life. Coaches aren’t more aggressive at home despite crowd support. This contradicts intuition about home-field psychology.

4.3.2. Model Performance Evaluation

With 88% accuracy, the inferred reward function explains the observed behaviour well. Our model was able to predict field goals and punting with high accuracy demonstrated by the precision, recall

***** MODEL PERFORMANCE METRICS EVALUATION *****				
Overall Accuracy: 88.0%				
Action	Precision	Recall	F1-Score	Support
fg	84.0%	90.9%	87.3%	5342
go	73.4%	55.8%	63.4%	3611
punt	93.1%	96.4%	94.7%	11850
Macro Avg	83.5%	81.1%	81.8%	
Weighted Avg	87.3%	88.0%	87.4%	

Figure 9. Evaluation of Model Performance

and F1 scores. However, the model has a low recall of 56%. The model missed 44% of the go-for-it decisions. One of the reasons for this could be that “go-for-it” is the minority class. Only 17% of all fourth down decisions are “go-for-it”. Further, coach personality, player match-ups, momentum and many other factors could influence a coach’s decision. This data is not accounted for in our feature set. To put it simply, our IRL models show that NFL coaching is conservative with high reliance on rules and could benefit with greater analytics integration in fourth down decision-making.

Training accuracy (2016-2021):	88.4%			
Test accuracy (2022-2024):	86.8%			
Generalization gap:	1.6 pp			
 TEST SET PERFORMANCE (2022-2024)				

	precision	recall	f1-score	support
fg	0.831	0.912	0.870	1491
go	0.787	0.510	0.619	1149
punt	0.903	0.977	0.939	3177
accuracy			0.868	5817
macro avg	0.840	0.800	0.889	5817
weighted avg	0.862	0.868	0.858	5817
 BEHAVIORAL CHANGES (2016-2021 vs 2022-2024)				

Action	Train %	Test %	Change	
fg	25.7%	25.6%	-0.1 pp	
go	16.4%	19.8%	+3.3 pp	
punt	57.9%	54.6%	-3.3 pp	

Figure 10. Action metrics on test set and action frequency changes between training and test set

4.3.3. Temporal Validation

We re-ran our IRL model with a truncated training set with data from 2016 to 2021 only. Data from 2022 - 2024 was used as our test dataset. Model accuracy drops marginally by 1.6 percentage points. Random sampling variation typically causes 1-2% fluctuation. This indicates that the model is generalizable. Further, our coefficients are stable showing little change compared to go decision coefficients from the original model. This indicates that coaching philosophy remains the same. We see a small change in action decisions between 2016-21 and 2022-24. A 3.3 pp change in Go for it decisions means that teams went for it 20% more in the last 2 years. Similarly, there was a 3.3 pp reduction in punt decisions translates to a ~65 reduction in punt decisions. This indicates a small shift in the decision boundary and coach risk tolerance.

4.3.4. High Confidence Error Evaluation

We evaluated errors where our model was more than 80% confident about prediction but was wrong. We found that for about 4% of our data, the model gets confused between go and punt decisions. This is because of an unconventional coach decision of “going for it” while still deep in their own territory. 9 out of 10 errors where the model had very high confidence, the coach went for it while still deep in their territory. In some of the cases (387) the yards to go were not small either. This indicates that some coaches/teams were more aggressive than usual.

1. HIGH-CONFIDENCE ERRORS (Model >80% sure but wrong)										

Found 871 high-confidence errors (4.2% of data)										

season	week	posteam	yardline_100	ydstogo	score_differential	must_score	action	predicted	pred_confidence	
2020	11	DEN	84.0	14.0	7.0	0	go	punt	0.999924	
2024	10	PIT	84.0	15.0	7.0	0	go	punt	0.999759	
2019	17	LA	83.0	7.0	3.0	0	go	punt	0.999645	
2016	12	BAL	77.0	8.0	7.0	0	go	punt	0.999644	
2022	9	TEN	91.0	26.0	-3.0	0	go	punt	0.999602	
2017	11	GB	66.0	1.0	-23.0	1	punt	go	0.999507	
2020	12	CIN	80.0	6.0	-3.0	0	go	punt	0.999358	
2017	13	LV	80.0	5.0	3.0	0	go	punt	0.998915	
2017	15	DAL	76.0	11.0	0.0	0	go	punt	0.998817	
2019	1	BAL	3.0	3.0	42.0	0	go	fg	0.998418	

Table 5. Top 10 errors where model had 99% confidence in its prediction

Specifically, our model had a 44.2% error rate, predicting punt or field goal equally. Our model particularly falters in tricky scenarios such as short yardage in its own territory (see table 5 GB, LV, BAL, LA). In this specific scenario, our model had 78% accuracy. Here, coach behaviour depends on their risk appetite. Analytics driven coaches would go for it as they would trust 70% conversion rate. Risk averse coaches would punt. In two other scenarios, must-score situations and marginal FG range, the model had 71% and 72% accuracy respectively. The challenge with must score situations is that they arise due to a complex interaction between clock management, field position and specific down. Coaches tend to act aggressively but some may opt for field goals others may go for it. The 45-55 yard range makes the decision even more unpredictable. Most kickers can successfully execute a field goal from 45 yards. But execution from 48-52 yards requires kickers with leg strength, weather conditions and the coach's trust in the kicker. To sum it up, training at league level made the model more cautious about "go for it" decisions and resulted in a conservative decision boundary.

3. PROBLEMATIC CONTEXTS

Short yardage in own territory: 78.2% accuracy
Sample size: 712

Must-score situations: 70.7% accuracy
Sample size: 1298

Marginal FG range (45-55 yards): 71.7% accuracy
Sample size: 2291
FG/punt decision boundary is fuzzy

Figure 11. Game situations where model accuracy is limited

4.3.5. Inference, Limitations, and Assumptions

Our model foremost assumes that the coaches are Boltzmann-rational. This refers to the fact that coaches choose high-reward actions with higher probability, but not deterministically. However, in reality coaches are irrational - they may make emotional decisions, suffer from other biases such as loss aversion, recency bias and can be influenced by criticism from the media.

Another key assumption is that features contribute independently and linearly to reward but if anything our analysis shows that factor interactions (e.g. field position and kicker ability) change fourth

down decisions. So, the true reward function might actually be non-linear. XG Boost or Random Forest could demonstrate these interactions better with better prediction rate, but we opted for multinomial logistic regression due to high interpretability of the coefficients. Having mentioned that, it is important to highlight that our model does not capture causation but rather shows association. Instrument variables need to be used to demonstrate causality.

A limitation we have already highlighted is our feature list not accounting for coach and team level effects such as personality, team dynamics, coach intuition, momentum. This makes “go-for-it” prediction extremely hard. Also, we currently assume homogeneity across fourth down decisions across coaches and teams. We already highlighted the impacts of it through our high confidence error analysis. To supplement it, we did a quick team level analysis and found a small variation in go for it rates of ~13% to 24%.

Appendix A. Supplementary Material

Category	Example Columns	Description
Identifiers	season, week, game_id, play_id, game_date	Uniquely identify each play
Teams & Type	posteam, defteam, home_team, away_team, posteam_type	Offensive vs. defensive context
Game State	qtr, game_seconds_remaining, ydstogo, yardline_100, score_differential, goal_to_go	Captures real-time decision context
Timeouts & Clock	home_timeouts_remaining, away_timeouts_remaining, posteam_timeouts_remaining, defteam_timeouts_remaining	Game management constraints
Conditions	roof, surface, temp, wind	Environmental effects
Play Indicators	play_type, rush_attempt, pass_attempt, punt_attempt, field_goal_attempt	Encodes possible 4th-down actions
Reward Signals	epa, wpa, success, yards_gained, first_down, touchdown	Learning targets for evaluation
Team Form (Rolling)	off_epa_4w, def_epa_4w	4-week rolling offensive & defensive performance
Special Teams	fg_pct_short, fg_pct_mid, fg_pct_long, punt_net_4w, kick_distance	Field goal & punting reliability
Fatigue & Drive Context	plays_in_drive_so_far, play_elapsed_s, game_time_elapsed, def_time_on_field_cum, def_time_on_field_share, is_q4_or_later	Cumulative drive effort and time on field
Action Label	action ∈ {go, punt, fg}	4th-down decision category

Table A.1. Feature schema for 4th down dataset.

	action	count	mean	std	min	max
1	go	5678	0.053588	2.961502	-11.448006	8.043193
0	fg	8388	0.053387	1.563120	-9.964310	3.524636
2	punt	17783	-0.132815	0.957115	-7.345287	8.227220

Figure A.1. Action summary for EPA. On average, go and FG yield slightly positive EPA (higher variance for go) while punts are negative EPA with a narrower spread.

	yardline_bin	action	count	mean	ydstogo_bin	action	count	mean
0	Red Zone (0-20)	fg	4241	0.008068	0	1-2	fg	961 0.049541
1	Red Zone (0-20)	go	1417	-0.178346	1	1-2	go	3038 0.317186
2	Red Zone (0-20)	punt	0	NaN	2	1-2	punt	2113 -0.386676
3	Opp 20-40	fg	4098	0.100926	3	3-5	fg	2376 0.025491
4	Opp 20-40	go	1759	0.008281	4	3-5	go	1253 0.146421
5	Opp 20-40	punt	538	-0.569908	5	3-5	punt	3676 -0.214947
6	Midfield (40-60)	fg	49	-0.000071	6	6-10	fg	3270 0.077067
7	Midfield (40-60)	go	1600	0.394744	7	6-10	go	920 -0.402676
8	Midfield (40-60)	punt	6035	-0.169821	8	6-10	punt	6299 -0.113161
9	Own 40-20	fg	0	NaN	9	11-20	fg	1634 0.046264
10	Own 40-20	go	789	0.043128	10	11-20	go	402 -0.946915
11	Own 40-20	punt	8644	-0.069112	11	11-20	punt	4872 0.003906
12	Own 20+	fg	0	NaN	12	20+	fg	147 0.081813
13	Own 20+	go	113	-1.090245	13	20+	go	65 -1.410538
14	Own 20+	punt	2566	-0.168732	14	20+	punt	823 -0.073982

Figure A.2. EPA by yardline bin (left) and by yards-to-go (right). Go is strongly positive around midfield and negative deep in own territory. FG is best in the opponent 20-40, and punts are most costly near midfield and in opponent territory. Short-yardage situations favor go with positive EPA, whereas greater distances make FG/punt safer options.

Decision table by yardline:

action	fg	punt	go	best_alt	delta_go
yardline_bin					
Red Zone (0-20)	0.008068	NaN	-0.178346	0.008068	-0.186414
Opp 20-40	0.100926	-0.569908	0.008281	0.100926	-0.092645
Midfield (40-60)	-0.000071	-0.169821	0.394744	-0.000071	0.394815
Own 40-20	NaN	-0.069112	0.043128	-0.069112	0.112240
Own 20+	NaN	-0.168732	-1.090245	-0.168732	-0.921512

Figure A.3. Decision table for EPA by yardline with difference between a method and the best alternative. Relative to the best FG/punt, go is not favorable in the redzone and deep in own territory.

Decision table by ydstogo:

action	fg	punt	go	best_alt	delta_go
ydstogo_bin					
1-2	0.049541	-0.386676	0.317186	0.049541	0.267645
3-5	0.025491	-0.214947	0.146421	0.025491	0.120930
6-10	0.077067	-0.113161	-0.402676	0.077067	-0.479743
11-20	0.046264	0.003906	-0.946915	0.046264	-0.993179
20+	0.081813	-0.073982	-1.410538	0.081813	-1.492351

Figure A.4. Decision table for EPA by yards-to-go with difference between a method and the best alternative. Go beats the best alternative for short-yardage situations but falls off sharply beyond 6 yards and becomes increasingly worse as distance grows.

	action	count	mean	std	min	max
1	go	5678	0.007348	0.088628	-0.752197	0.862577
2	punt	17783	0.002712	0.045451	-0.418035	0.944656
0	fg	8388	0.000841	0.062224	-0.683856	0.570333

Figure A.5. Action summary for WPA. Across all fourth downs, go has the highest mean WPA and the largest variance, punts are slightly positive on average, and FGs cluster near zero with moderate spread.

yardline_bin	action	count	mean	ydstogo_bin	action	count	mean
0 Red Zone (0-20)	fg	4241	-0.03972	0 1-2	fg	961	-0.013769
1 Red Zone (0-20)	go	1417	-0.003196	1 1-2	go	3038	0.013026
2 Red Zone (0-20)	punt	0	NaN	2 1-2	punt	2113	-0.008619
3 Opp 20-40	fg	4098	0.005833	3 3-5	fg	2376	-0.003471
4 Opp 20-40	go	1759	0.008645	4 3-5	go	1253	0.006922
5 Opp 20-40	punt	538	-0.015667	5 3-5	punt	3676	-0.000953
6 Midfield (40-60)	fg	49	-0.000011	6 6-10	fg	3270	0.002003
7 Midfield (40-60)	go	1600	0.016161	7 6-10	go	920	-0.004480
8 Midfield (40-60)	punt	6035	0.001484	8 6-10	punt	6299	0.002645
9 Own 40-20	fg	0	NaN	9 11-20	fg	1634	0.012868
10 Own 40-20	go	789	0.006045	10 11-20	go	402	-0.002555
11 Own 40-20	punt	8644	0.004754	11 11-20	punt	4872	0.009547
12 Own 20+	fg	0	NaN	12 20+	fg	147	0.006524
13 Own 20+	go	113	0.003696	13 20+	go	65	-0.021147
14 Own 20+	punt	2566	0.002573	14 20+	punt	823	0.008227

Figure A.6. WPA by yardline bin (left) and by yards-to-go (right). By field position, go peaks around midfield and remains positive in the 20-40 yard zones while red-zone tries are roughly neutral. Punts are least harmful deep in a team's own territory. Short-yardage (1-5 yards) favors go with positive WPA whereas longer distances tilt towards safer kicking alternatives (FG/Punt) and go is zero or negative WPA.

Decision table by yardline:						
action	fg	punt	go	best_alt	delta_go	
yardline_bin						
Red Zone (0-20)	-0.003972	NaN	-0.003196	-0.003972	0.000776	
Opp 20-40	0.005833	-0.015667	0.008645	0.005833	0.002813	
Midfield (40-60)	-0.000011	0.001484	0.016161	0.001484	0.014676	
Own 40-20	NaN	0.004754	0.006045	0.004754	0.001291	
Own 20+	NaN	0.002573	0.003696	0.002573	0.001123	

Figure A.7. Decision table for WPA by yardline with difference between a method and the best alternative. Relative to the best FG/punt, the go action underperforms near the red zone but dominates at midfield.

Decision table by ydstogo:						
action	fg	punt	go	best_alt	delta_go	
ydstogo_bin						
1-2	0.049541	-0.386676	0.317186	0.049541	0.267645	
3-5	0.025491	-0.214947	0.146421	0.025491	0.120930	
6-10	0.077067	-0.113161	-0.402676	0.077067	-0.479743	
11-20	0.046264	0.003906	-0.946915	0.046264	-0.993179	
20+	0.081813	-0.073982	-1.410538	0.081813	-1.492351	

Figure A.8. Decision table for WPA by yards-to-go with difference between a method and the best alternative. Go beats the best alternative for 1-5 yards, but FG/punt becomes superior once distance exceeds 6-10 yards.

Figure A.9. Users can select their metric to optimize for on the left sidebar and input season and week of the game, the home and away teams, and which team is currently possessing the ball.



Highmark Stadium — Buffalo Bills

Venue & weather

Buffalo Bills home stadium: **Highmark Stadium**. Defaults applied → roof: outdoors, surface: turf.
You can override below.

Roof

outdoors

Surface

turf

Temp (°F) Wind (mph)

50 5

Note: Temperature and wind affect predictions only if the model was trained with these features.

Figure A.10. The stadium for the home team will automatically display, with relevant environment conditions being autofilled where applicable. Users have the option to manually override these as fields as they deem fit.

4th Down Situation

Quarter	Minutes left in quarter	Seconds left in quarter	Yards to go
4	1	33	3
		- +	- +

Offense score Defense score

40 38

- + - +

Ball on

OWN side

Yard line (1-49)

38

- +

Example: 48 yd line, enemy territory near midfield → 'OPP side', 48. 10 yd line, deep in own territory → 'OWN side', 10.

Figure A.11. 4th-d own situation fields include quarter, time remaining, yards-to-go, game score, and field position. A visual cue is displayed for the user to help them in determining what it means to be on your own side of the field versus the opponent's side.

Timeouts Remaining

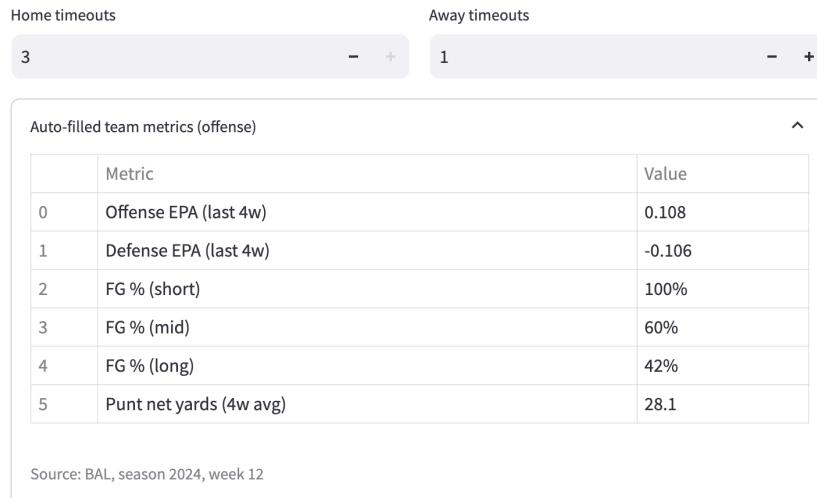


Figure A.12. Users can input the number of timeouts (the number of times each team can stop the clock) for each team. Some teams will elect to go on 4th-down knowing that the opponent does not have the ability to stop the clock, effectively allowing them to run out the clock and win the game. Rolling team metrics for the offense are also displayed to provide context for the decision the coach makes. For example, the Ravens appear to have struggled from mid- and short-distance FGs and subpar punting. The estimated FG distance from Baltimore's own 38 is 105 yards.

Recommend decision

Recommendation: GO (*optimized for WPA*)

GO improves win probability by **2.6%** vs PUNT.

Relative gains:

- GO vs PUNT: **2.6% WPA**

⚠ Excluded as infeasible: FG (field goal out of realistic range)

Figure A.13. The final step is for user to press the “Recommend decision” button. Here, in the scenario presented in Figures A.9 through A.12, making the decision to go-for-it improves win probability by 2.6% compared to punting the ball. Kicking a FG is not considered feasible in this scenario because it is beyond the range that a kicker could successfully make a FG.

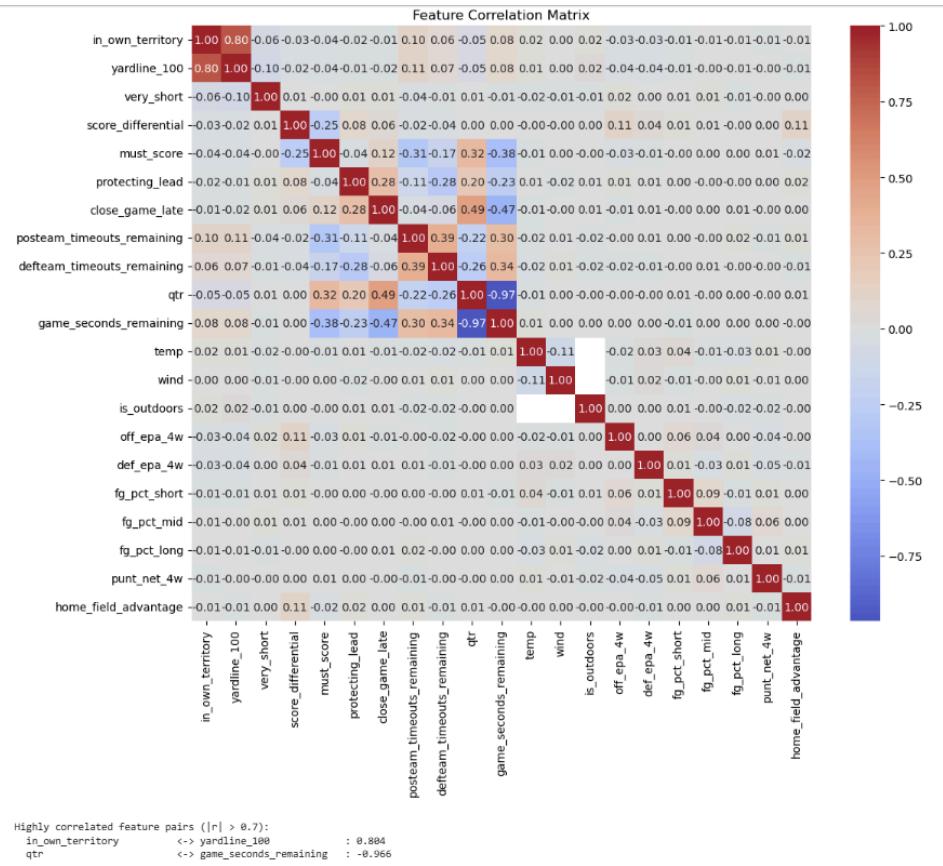


Figure A.14. Correlation matrix for IRL candidate features

	feature	VIF
13	is_outdoors	494.421894
10	game_seconds_remaining	19.815024
9	qtr	17.908079

Figure A.15. Multicollinearity check for IRL candidate features

Appendix A.1.

A.1.1 Source Code, Detailed Framework, and Dataset

The complete implementation, data-processing workflow, front-end demo, and study dataset are available at the following GitHub repository:

https://github.com/lucas-huynh/NFL_4th_MAB

A.1.2 LLM Use

Inverse Reinforcement Learning:

Claude Pro: <https://claude.ai/share/9e2a0fda-8016-453d-9fad-91dc5580baf1>