445 Presentation

Zach Goldstein Sam Rolsten Andrew Davenport Ryan Lang

2025-12-17

Contents

While watching some football we were discussing what was the right decision when it was 4th down. While our memory does not service us for what specific game or play it was, our brains were tickled by the big question we had, "Can we build a model to help decide what to do on 4th down?".

We used nflfastR to collect for play by play data. It is a package that allows for easy NFL data scraping. This play by play data began being collected in 1999 and is updated frequently. With particular commands you can even grab data from the previous day. This data included every single play from the 1999 season until the present, including well into 300 different columns. This dataset also included multiple different advanced statistics, however, we did not use them for the most part for the purposes of answering our question.

To briefly get into it, a 4th down is a situation in football where the offensive team must make it past the 1st down marker in order to continue their offensive possession. Teams can elect to punt the ball (giving the ball to the other team but further away from the endzone), kick a field goal (get 3 points), or go for the first down, risking the defensive team taking the ball for offense at the spot of the downed play.

Our first task was deciding when it was optimal to punt, kick, or go-for-it. While many situations are relatively easy to understand, many situations are far less clear. Once we make a model that handles these three options, we will look to then build an extension. When the first model says go-for-it, we then want to decide whether to run or pass the ball. We will eventually build a function to call these models. Our goal is to have the ability to watch a game live and be able to put relevant information into the function during a 4th down situation, and then compare what our model says to what the team actually does.

As mentioned previously we began with data from 1999 to 2025. This was play by play data as well. This means that every play of every game, all data is collected. While many data points will be left as N/A due to many data points that will not apply to a particular play. For example, all data that monitors when and how teams score would be left as N/A unless a scoring play actually happened. During our exploratory data analysis we did not reduce the amount of years we considered. When running models, we ended up reducing how many years we were looking at, however, this will be discussed later.

When we analyze our target, we see that there are three mutually exclusive outcomes. On any given 4th down, teams can elect to either punt, kick, or go for the first. These situations, especially late in the game, can heavily affect the outcome of the game. However, they are hard to predict. The same situation in the 1st quarter can have a vastly different outcome than a 4th down in the 4th quarter. A team down by one score will behave differently than a team down three scores. And both of these scoring examples would change how coaches call plays depending on how much time is left on the clock. Figuring out the league wide coaching behaviour would provide interesting discussion from not only a fans perspective but also other

coaches playcalling. An example, if a coach knows 5% of coaches would go for it given their situation, they may want to go for it and hope the defensive coach did not account for such a low likelihood event, and did not have a play ready.

1 EDA

Since our analysis is focused on 4th down plays that is all we will look at. The first thing we did was always remove all non-4th down plays. We also removed weird N/A values. For example if there was a 4th down play that did not include the score differential we would remove. This is because we still had plenty of data points. We also did not want to place artificial importance if any value was frequently listed as N/A when it was not supposed to be.

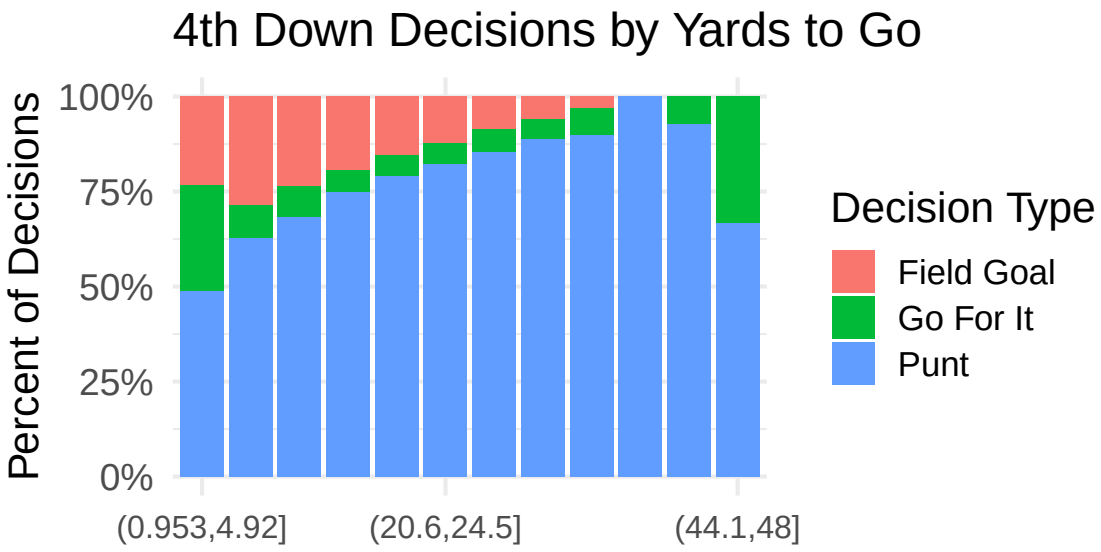
While teams can fake a punt or kick and then go for the first, we ended up deeming these fakes as whatever they were lined up for. We decided on this because there are not many variables quantifying the fake. We also would have no way of differentiating when a punt/kick breaks down (wild snap causes punter to scoop and run) versus a true planned fake. We are also mostly interested in coach behaviour since they make play calls. Looking at what play happens is inherently a study of coaching and not individual players. Fakes typically happen on purpose when the situation is very much already decided on. You wouldn't likely see a fake punt when there is a choice between kicking a field goal as well, because the fake depends on the defence being complacent in the situations they are in. The coach leans on the fact that the whole stadium except himself thinks they are not faking.

We landed on roughly 30 variables that we deemed as important. We initially went through all columns individually noting what we thought were important. We all came together and talked through everyone's list. Some notable ones include time left in the game, how many yards to the endzone and the first down, the score differential in the game, and a couple other predictive advanced stats like expected points after or epa for short (each play what are the expected points scored based on the results of that play). While there are more, these paint a good picture for what we deemed as important.

Before we moved into drafting models, we wanted to get a better feel of our data. We looked at decision rates based on various game situations. We also looked at win probabilities given 4th down situations. This helped us get a feel for what decisions were happening across the board.

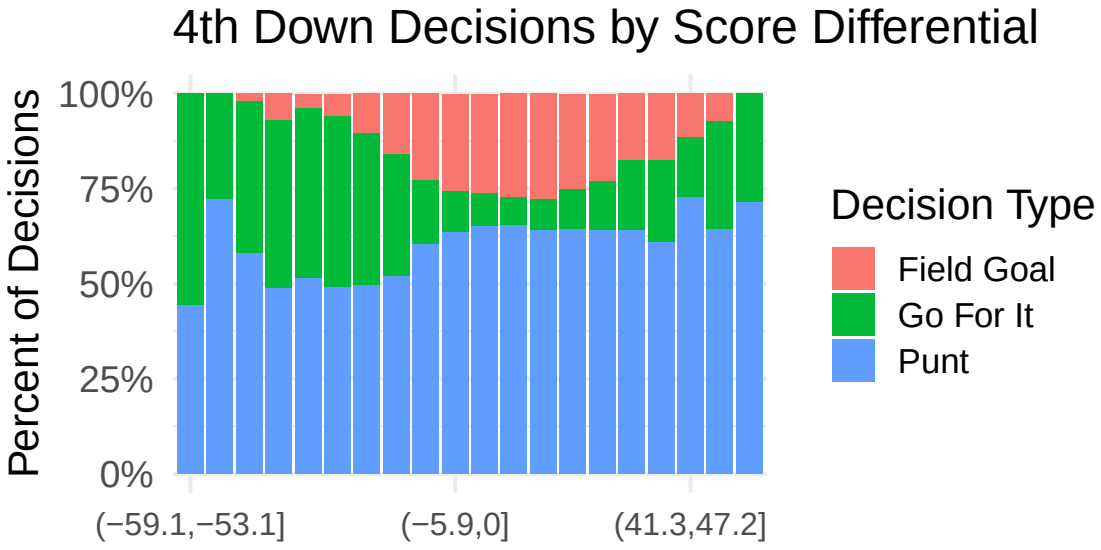
We then looked at what decision was being made based on yards to go. The more yards to go, the furthur away you are from getting a first down. We would expect to see that punting goes up the higher the yards to go, we may also see a slight boost in kicking as well. We would expect mixed results at the shorter yards

where anything is possible.



We notice within 5 yards that there are about the split you would expect, with plenty of attempts at going for it. We notice a speedy drop-off in go-for-it attempts however kicking still stays prevalent for a little bit. We do notice there is a point where there are so many yards to go that teams are no longer kicking and even opting to go-for-it. We imagine these attempts are in late game, trailing situations, most likely either last seconds on the clock, or too close to punt, but to far to kick.

Score differential is the difference in score between you and your opponent. A negative value means you are down and a positive means you are leading. We expected to see punting around the board, however we were interested to see the aggressive outcomes and how they behaved because of the score differential.



Here, we see that go-for-it attempts are spread across all score differentials, however we see more attempts happening when teams are down. We expected this to be the case, teams will get more aggresive as their

down. We do notice a zone of kicks happening, but less so at the tails. It seems in close games teams would rather get points on the board, instead of going for something risky, also why there would be more kicking and less going-for-it when teams are up.

We noticed roughly what we expected to notice from the EDA. When looking at decision versus a particular future predictor we see relationships that we expect. While we do uncover some interesting relationships with the graphs, we would generally have concluded these things on our own. We looked into a couple other variables, primarily regarding score, time, and yards. Overall, our EDA left us with a takeaway before we moved into the model work. We noticed that punting situations are rather predictable, but situationally, kicking or going for it is a tough choice for play callers. We believe this will be where our model will have the most challenge.

2 Models

2.1 Feature Selection

Next, We engineered a few additional predictors designed to capture nonlinear patterns in how coaches behave. We created score differential (score_diff = posteam_score − defteam_score) to quantify how urgency changes with game state. We also added an indicator for whether the offense is on its own half of the field (own_half), since aggressiveness is strongly tied to field position. Finally, because yards-to-go is highly skewed and coaches respond differently to short vs. long distances, We created a log-scaled distance feature (log_ydstogo = log1p(ydstogo)) to help models learn those relationships more smoothly. After engineering these features, We removed rows with missing values to ensure the final modeling dataset was clean and consistent.

Because the dataset contains many possible predictors after dummy-encoding categorical variables, We used LASSO multinomial regression as a feature selection step before fitting a more flexible final model. We split the data by season so the model is trained on older football and evaluated on modern football: training seasons were 1999–2019 and testing seasons were 2021–2025 (excluding 2020 to avoid COVID-era anomalies). Using tidymodels, We built a recipe that dummy-encoded categorical predictors and removed zero-variance columns, then tuned the LASSO penalty parameter with 5-fold cross-validation, selecting the value that minimized multiclass log loss. From the fitted LASSO model, We extracted the predictors with non-zero coefficients for at least one of the three decision classes, which produced a reduced set of the most informative features for predicting coaching decisions.

After selecting the most predictive features, We trained the main predictive model using a weighted Random Forest. This step matters because 4th-down decisions are imbalanced (punts occur most often, then field goals, with go-for-it least frequent), and a standard classifier can inflate accuracy by favoring the majority class. To address this, We computed class weights inversely proportional to class frequency and fit a Random Forest (ntree = 500) using the reduced feature set. We evaluated performance on the held-out modern seasons using a confusion matrix and found the model achieved strong overall accuracy (roughly high-80% range), with the best performance on punts and field goals and relatively lower performance on go-for-it decisions. That pattern is expected: punts and field goals typically occur in more "obvious" regions of field position and distance, while go-for-it situations occur in the most borderline cases where coaching tendencies vary the most.

To better understand where the model struggled, We created a diagnostic plot that bins situations into field position zones (own red zone, midfield, opponent red zone) and yards-to-go buckets (short/medium/long/very long), then visualizes whether the prediction was correct within each true decision category. This helped confirm that most errors occur in the ambiguous middle of the field and medium distance ranges, where real coaching behavior is less consistent. We also tested whether a simpler model could achieve comparable performance by fitting Random Forest models with increasing numbers of LASSO-selected predictors (top 3, top 6, top 9, and all selected features). Accuracy increased steadily as more predictors were included, which suggested that the additional LASSO-selected variables were genuinely adding predictive signal rather than just noise-driven complexity.

Finally, We packaged the model into a user-friendly 4th-down advisor function. The function takes a single game situation as input (yardline, distance, score differential, quarter, time remaining, season/week, roof and surface), applies the same preprocessing recipe used during training (so the new data is encoded consistently), selects the LASSO-chosen feature columns in the correct order, and returns both the predicted decision and the predicted probability for each option (punt, field goal, go for it). This makes the model usable in real time: during a live game, a user can enter the current situation and instantly get a recommendation that reflects league-wide coaching tendencies learned from historical data, while also seeing the uncertainty of the choice through the probability outputs.

2.2 Feature Selection Results

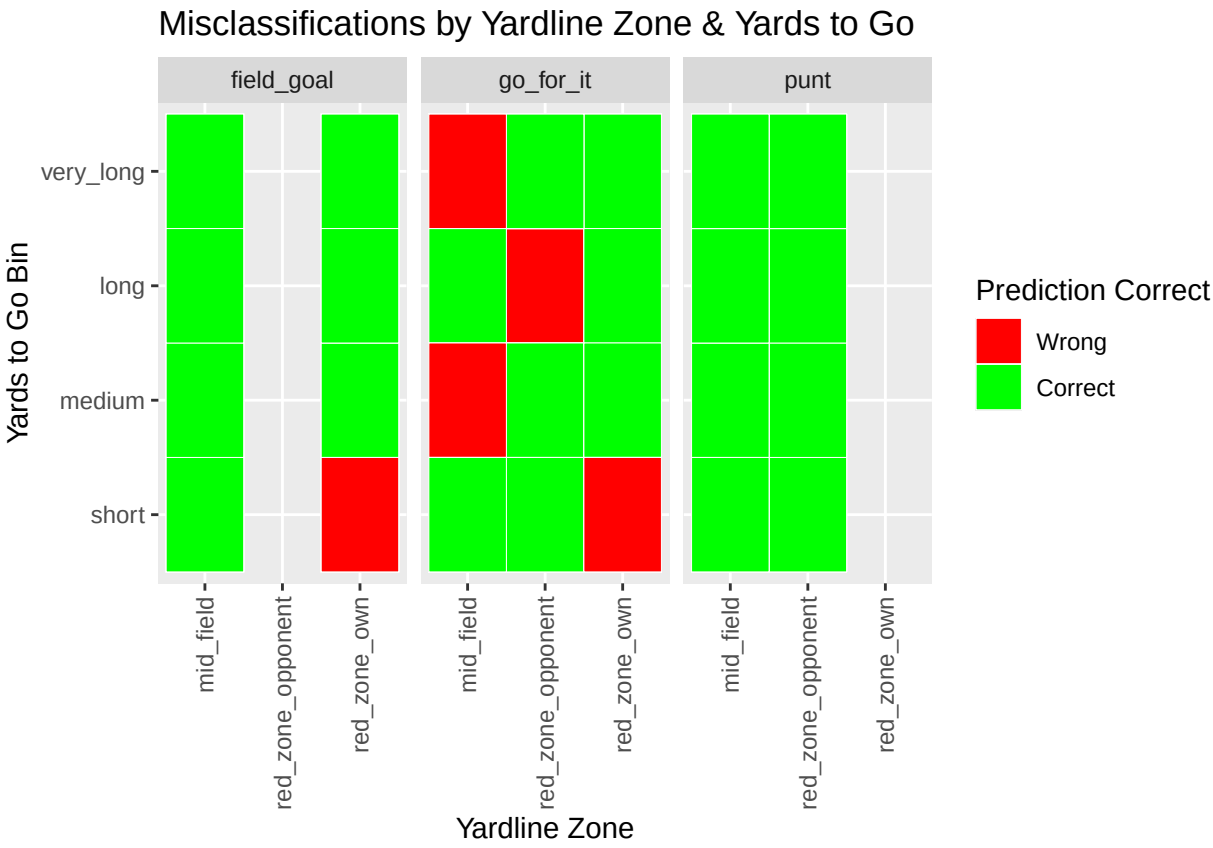
- After performing 5 fold cross validation to tune of LASSO function it was able to output the 20 most predictive features in our data set of 374 features

Feature 1	Feature 2	Feature 3
yardline_100	ydstogo	log_ydstogo
own_half	season	week
roof_open	roof_outdoors	surface_dessograss
surface_matrixturf	score_diff	qtr
game_seconds_remaining	half_seconds_remaining	roof_dome
surface_a_turf	surface_astroplay	surface_sportturf
surface_astroturf	surface_fieldturf	

2.3 Overall Accuraccy Results

	field_goal	go_for_it	punt
field_goal	4349	581	15
go_for_it	248	2419	131
punt	214	916	10319

2.4 Accuracy Results cont.



2.5 Testing Models to find best performing one

model	accuracy
top3	0.8197165
top6	0.8377449
top9	0.8436328
all_lasso	0.8878699

2.6 Model Work: Pass vs Run (Within Go-For-It)

While the primary 4th-down model determines whether a team should punt, attempt a field goal, or go for it, that recommendation alone does not specify the type of play a team should run if "go for it" is selected. To extend the decision framework, we developed a second-stage model focused specifically on predicting whether an offense should run or pass the ball in true go-for-it situations.

We first isolated all plays in the dataset where the offense actually attempted to convert on 4th down with either a designed run or pass. We then constructed a new binary outcome variable, "go_type", with levels run and pass. Importantly, we engineered this feature set to be consistent with the predictors used in the 4th-down decision model to ensure that both stages of the advisor could take the same type of input.

Predictors included contextual variables such as yardline, yards to go, score differential, quarter, game-clock states, and stadium conditions (roof and turf surface). We also applied the same transformation used earlier: including the log-scale transformation of yards-to-go and the indicator for whether the offense was on its own half of the field, which made the model robust to skewed distances and field position patterns.

We trained a random forest classifier on all run/pass go-for-it plays between 1999 and 2019, reserving the 2021–2025 seasons for testing (excluding 2020). This approach mirrors the train/test scheme of the main model, ensuring that both stages of the advisor evaluate modern NFL behavior using patterns learned from historical play-calling.

.metric	.estimator	.estimate
accuracy	binary	0.8243105

The resulting model achieved strong predictive performance, correctly classifying 82.4% of unseen go-for-it plays as runs or passes. This accuracy indicates that situational variables alone (without personnel or formation data) still provide considerable predictive signal about how teams choose to attack on 4th down.

To integrate both models into a single decision engine, we created an advisor that:

1. Uses the main 4th-down classifier to choose among punt / field goal / go for it.
2. If "go for it" is recommended, passes the same game situation to the run/pass model.
3. Returns a complete recommendation such as "go for it – run" or "go for it – pass" along with associated probabilities.

This two-stage approach allows the advisor to not only mimic coaching tendencies but also provide interpretable, actionable guidance for specific 4th-down situations.

2.7 Results on Current NFL season

After combining the two models, we evaluated the advisor on real NFL plays from the 2025 season. Our goal was to determine whether the system's recommendations aligned with both league-wide tendencies and actual in-game decisions.

Across the five sample situations we tested, the advisor made recommendations consistent with historical behavior and provided clear probabilistic reasoning for each decision. For example:

2.7.1 Commanders v Packers Week 2

For this play, it is week 2 of the season in the 4th quarter with 12:00 min on the clock. The Commanders have the ball down 11 points and are on the opponents 34 yard with 13 yards to go until the 1st down. For this situation, our model predicts that roughly 68% of coach's would kick a field goal here, 16% would go for it, and 17% would punt. Here the Commanders elect to kick the field goal which our model agrees with, despite the team being down by 11 and the goal posts being 52 yards out.

predicted_decision	prob_field_goal	prob_go_for_it	prob_punt	recommended_play
field_goal	0.676	0.156	0.168	field_goal

To view the actual play (with valid prime account, play_id 2682) start stream at (1:46:40): - https://www.amazon.com/gp/video/detail/B0F8KWZG8C/ref=atv_dp_amz_c_n9jOf6_6_13?jic=8%7CEgNhbGw%3D

2.7.2 49ers v Rams Week 5

For this play, it is week 5 of the season in overtime with 3:22 min on the clock. The Rams have the ball down 3 points and are on the opponents 1 yard with 1 yards to go until the 1st down. Our models says 47%

of coach's would kick a field goal here, 53% would go for it, and less than 1% would punt. The Rams here elect to go for it and run the ball rather than kicking the FG to tie the game. (*they go on to lose 26-23*)

predicted_decision	prob_field_goal	prob_go_for_it	prob_punt
go_for_it	0.47	0.528	0.002

predicted_go_type	prob_run	prob_pass
run	0.748	0.252

To view the actual play (with valid prime account, play_id 4950) start stream at (3:24:10): - https://www.amazon.com/gp/video/detail/B0DWTQMTMP/ref=atv_dp_amz_c_n9jOf6_6_10?jic=8%7CEgNhbGw%3D

2.7.3 Lions v Cowboys Week 14

For this play, it is week 14 of the season in the 2nd quarter with 0:55 sec on the clock. The Cowboys have the ball down 11 points and are on the opponents 37 yard with 4 yards to go until the 1st down. The advisor again recommended going for it and identified pass as the preferred play type (93% probability), reflecting the league-wide aggressiveness of teams trailing by double digits in plus territory. Dallas here elects to go for the field goal.

predicted_decision	prob_field_goal	prob_go_for_it	prob_punt
go_for_it	0.156	0.602	0.242

predicted_go_type	prob_run	prob_pass
pass	0.071	0.929

To view the actual play (with valid prime account, play_id 2163) start stream at (1:25:50): - https://www.amazon.com/gp/video/detail/B0F9GZ984M/ref=atv_dp_amz_c_n9jOf6_6_1?jic=8%7CEgNhbGw%3D

2.7.4 Broncos v Raiders Week 10

For this play, it is week 10 of the season in the 2nd quarter with 9:22 min on the clock. The Raiders have the ball up 7 points and are on the opponents 31 yard with 2 yards to go until the 1st down. Our models says 61% of coach's would kick a field goal here, 34% would go for it, and only 5% would punt. The Raiders

here elect to go for it when they were up for 7, which results in a TD (**that was overturned by a offensive PI, w/ TD off the board they elect to punt it*)

predicted_decision	prob_field_goal	prob_go_for_it	prob_punt	recommended_play
field_goal	0.614	0.336	0.05	field_goal

To view the actual play (with valid prime account, play_id) 1316 start stream at (51:15): - https://www.amazon.com/gp/video/detail/B0F83DDLSD/ref=atv_dp_amz_c_n9jOf6_6_5?jic=8%7CEgNhbGw%3D

2.7.5 Ravens v Dolphins

For this play, it is week 9 of the season in the 1st quarter with 8:50 sec on the clock. The Ravens have the ball down 3 points and are on the opponents 2 yard line with 2 yards to go until the 1st down. Our model says 73% of coach's would kick a field goal here, 27% would go for it, and less than 1% would punt. The Ravens here elect to go for it when they were down by 3, which results in a TD

predicted_decision	prob_field_goal	prob_go_for_it	prob_punt	recommended_play
field_goal	0.732	0.268	0	field_goal

To view the actual play (with valid prime account, play_id 455) start stream at (17:50): - https://www.amazon.com/gp/video/detail/B0F195LTV8/ref=atv_dp_amz_c_n9jOf6_6_6?jic=8%7CEgNhbGw%3D

3 Final Words

Across the league, fourth-down decision-making follows clear and consistent strategic patterns. Coaches are highly conservative in their own territory, strongly favoring punts, while aggressiveness increases steadily as teams approach midfield and the opponent's side of the field. Short yardage situations significantly raise the likelihood of going for it, particularly when teams are trailing or when the game context reduces the relative cost of failure.

However, the analysis also highlights that the most interesting decisions occur in a narrow set of "gray-area" situations, typically near midfield or just outside conventional field goal range. In these cases, no single choice overwhelmingly dominates while coaching behavior becomes far less predictable. This variability suggests that factors such as individual risk tolerance, game flow, and organizational philosophy play a larger role than pure situational metrics alone.

Overall, the findings indicate that while much of fourth-down strategy aligns with broadly shared league norms, meaningful deviations occur precisely where strategic trade-offs are most complex.

4 Sources

Baldwin, Ben; Carl, Sebastian; & Yurko, Ronald (2021). nflfastR: Efficient Scraping and Aggregation of NFL Play-by-Play Data. Journal of Open Source Software, 6(62), 3262. DOI: 10.21105/joss.03262.

Amazon Prime Video (2022). NFL game broadcast clips used for examples. Retrieved from primevideo.com.