DSML – Fall 2025

Final Project Report: Predicting NFL Field Goal Success

Quinn Peters

Date: November 12th, 2025

I understand and have adhered to all the tenets of the Duke Community Standard in completing every part of this assignment. I understand that a violation of any part of the Standard on any part of this assignment can result in failure of this assignment, failure of this course, and/or suspension from Duke University.

1 Problem Description and Executive Summary

A single NFL field goal can swing win probability by 30 percentage points or more, yet coaches must decide in real time whether to kick, go for it, or punt. Our central question is:

> **Given measurable environmental, contextual, and kicker-specific features, what is the probability that an NFL field goal attempt will be successful?**

We build and compare several probabilistic models for field goal success using regular-season and playoff data from 2013–2024. Rather than simply classifying makes vs. misses, our goal is to produce *calibrated probabilities* that can plug directly into decision tools (expected points, win probability, and 4th-down calculators). A motivating example from our presentation is Cam Little's 68-yard attempt in a dome with neutral pressure: our final model assigns a make probability of roughly 15%, illustrating both the usefulness and limitations of analytics in extreme situations.

We evaluate Logistic Regression, Bayesian Logistic Regression, Generalized Additive Models (GAM), Bagging, LightGBM, and Bayesian Additive Regression Trees (BART), along with ensembles. Models are trained on 2013–2023 data and evaluated on a held-out 2024 season to mimic genuine forecasting. Our primary metric is the Brier score; we also consider ROC–AUC, PR–AUC for misses, and Expected Calibration Error (ECE).

The best weighted ensemble attains a Brier score of approximately 0.110 on 2024 data, with good AUC and low ECE. For typical NFL kicks (30–55 yards), predicted probabilities are very well calibrated and track empirical make rates closely across distance and probability bins. A BART-based causal analysis confirms that kick distance is by far the dominant driver of success, while weather, pressure, and stadium effects matter but only at the margins (usually 1–3 percentage point shifts). The model struggles most in the long tail: very long kicks, unusual weather, and rare high-pressure scenarios with little historical data.

Our main conclusion is that **field goal success can be predicted reliably for normal NFL situations using calibrated ML models**, but extreme edge cases remain inherently uncertain and should still rely heavily on coaching judgment.

1.1 Why Predict Probabilities Instead of Make/Miss?

A natural starting point is to ask why we bother building a probability model at all, instead of a simple classifier that predicts make vs. miss. In our data, 86.3% of field goal attempts are successful. A trivial baseline model that always predicts "make" would therefore achieve about 86% accuracy without using any features at all. From a coaching perspective, however, such a model is useless: it cannot distinguish between a 25-yard chip shot and a 55-yard attempt into the wind.

NFL coaches think in terms of *risk*, not just binary outcomes. On a fourth down near the edge of field-goal range, the decision is not "will the kicker make it?" but rather "is a 68% kick, plus the downside of a miss, better than the expected value of going for it or punting?" A model that returns calibrated

probabilities is directly compatible with this decision calculus. For example, a change from 68% to 82% make probability is a meaningful shift in expected points and win probability, even though both cases would look identical to a classifier that only outputs "likely make."

This motivates our emphasis on Brier score, calibration error, and probability reliability instead of accuracy alone. In particular, we want:

- probabilities that reflect the empirical frequency of makes in similar situations (calibration), and
- sufficiently smooth behavior that small, realistic changes in distance or weather do not cause discontinuous jumps in the prediction.

The rest of our modeling and evaluation pipeline—including per-distance isotonic calibration and an ensemble of diverse models—is built around this goal of delivering decision-grade probabilities rather than just labels.

2 Data

2.1 Source and Scale

We construct our dataset from the `nflfastR` play-by-play logs for the 2013–2024 NFL seasons. We keep all standard field goal attempts and remove obvious data errors and blocked kicks (where responsibility is ambiguous). After cleaning, the dataset contains roughly 12,449 field goal attempts by 114 unique kickers with an overall make rate of 86.3%.

Each observation corresponds to a single kick with associated game context, environment, and kicker history at the time of the attempt. The temporal ordering of plays allows us to compute historical kicker statistics without leakage from future kicks.

2.2 Feature Set

We group our features into four conceptual categories:

Environmental: temperature, wind speed, binary indicators for rain and snow, altitude, and a roof indicator (dome vs. open). For dome stadiums we impute typical indoor values (approximately 68°F and low wind), while for outdoor stadiums we use weather summaries provided by `nflfastR`.

Contextual: quarter, season type (regular vs. playoffs), score differential, time remaining, and a buzzer-beater indicator for kicks in the final seconds. We also incorporate the Vegas win probability at the time of the snap as a summary of game flow.

Kicker history: rolling career field goal percentage and total attempts computed *up to* each kick, as well as simple indicators of experience. These statistics are re-computed season-by-season to avoid peeking into the future.

Physical/positional: kick distance (line of scrimmage + 17 yards), turf vs. grass surface, and home vs. away.

Exploratory analysis shows that distance is the dominant empirical signal: make rates are around 98% under 30 yards, roughly 80% for 40–49 yards, and about 68% for 50+ yards. Truly high-pressure situations (buzzer-beaters, playoff kicks) are quite rare, and snow games are less than 1% of the sample, which foreshadows the challenges of modeling extreme scenarios.

2.3 What a "Normal" NFL Field Goal Looks Like

To ground the modeling problem, it is helpful to characterize what a typical NFL field goal attempt looks like in our dataset. Using play-by-play logs from 2013–2024, we find:

- **Average distance:** 38.4 yards.

- **Distance distribution:** most kicks fall between 30 and 49 yards. Roughly 59% of attempts are in the 30–49 yard range, with only 17.3% coming from 50+ yards.
- **Weather:** games are usually played in mild conditions, with an average temperature of about 61.5°F and wind around 5.5 mph.
- **Stadium and surface:** about 71.8% of kicks are outdoors and 55.9% are on turf.

Distance buckets illustrate the steep performance gradient we are trying to model. Aggregating attempts into four bins yields:

- 0–29 yards: 24.2% of attempts, 98.4% make rate,
- 30–39 yards: 28.9% of attempts, 93.6% make rate,
- 40–49 yards: 29.6% of attempts, 80.1% make rate,
- 50+ yards: 17.3% of attempts, 67.8% make rate.

These descriptive statistics reinforce three key points that informed our modeling choices:

1. **Distance dominates the empirical pattern.** The drop from short kicks to 50+ yard attempts is large enough that any realistic model must treat distance as the primary driver of risk.
2. **Truly extreme conditions are rare.** Very cold temperatures, strong winds, snow, and record-length kicks appear in only a small fraction of plays, which helps explain why models struggle most in these tails.
3. **Most training examples are "boring" kicks.** The majority of data comes from moderate distances in reasonable weather. This is exactly where a well-calibrated model can deliver the most value to coaches making routine fourth-down decisions.

Throughout the rest of the paper, we use these baseline frequencies to sanity-check our fitted probabilities. For example, we are skeptical of any model that claims a 90% make rate on 50+ yard attempts, because such predictions are inconsistent with the empirical 67.8% average even before conditioning on additional features.

3 Methods

Our modeling pipeline is designed around three priorities: (1) realistic data cleaning that respects football domain knowledge, (2) a diverse set of predictive models, and (3) probability calibration with a focus on decision quality.

3.1 Preprocessing and Experimental Design

We first drop blocked kicks and obvious data errors, standardize surface and stadium labels to consistent categories, and parse weather strings to extract temperature, wind, and precipitation flags. For domes we set temperature and wind to fixed indoor-like values; for outdoor games we leave the recorded values, which occasionally produces very cold or very hot outliers. Kicker career statistics are computed cumulatively to prevent leakage.

To evaluate generalization, we use a *time-based split*: 2013–2023 for training and tuning, and the full 2024 season as a held-out test set. Within the training block we use cross-validation for hyperparameter tuning. This setup mimics how a real team would deploy a model before a new season.

Because we care about calibrated probabilities, not just ranking, we define the following evaluation metrics:

- **Brier score** (primary): mean squared error between predicted probabilities and outcomes.
- **ROC–AUC** and **PR–AUC (miss class)**: discrimination, especially for the relatively rare misses.
- **ECE@10**: expected calibration error across 10 probability bins.

These metrics capture both accuracy and the reliability of probabilities that will feed into downstream decision tools.

3.2 Model Families

We implement a suite of models with different strengths:

- **Logistic Regression:** a simple, interpretable baseline with regularization.
- **Bayesian Logistic Regression:** similar structure with explicit priors to control coefficient sizes.
- **GAM (Generalized Additive Model):** allows smooth nonlinear effects of distance, temperature, and wind while preserving interpretability.
- **Bagging:** ensembles of decision trees to reduce variance and capture interactions.
- **LightGBM:** gradient-boosted decision trees tuned for tabular data with complex feature interactions.
- **BART:** Bayesian Additive Regression Trees, which provide flexible nonlinear fits and a natural framework for causal analysis via counterfactual predictions.

Hyperparameters are chosen by minimizing the Brier score on validation folds. After training, we apply **isotonic calibration within distance bands** (e.g. 0–29, 30–39, 40–49, 50+ yards) to enforce realistic probability shapes and prevent overconfident predictions, especially at long distances where data are sparse.

Finally, we construct both equal-weight and optimized weighted ensembles of the best-performing models. Ensemble weights are chosen by grid search to minimize Brier score on validation data while avoiding extreme weights that would collapse back to a single model.

3.3 Causal Extension with BART

To better understand *why* probabilities change, we use BART in a causal mode: for each kick we generate counterfactual predictions under modified values of selected variables (e.g., moving from outdoor to dome, or from regular season to playoffs) while holding other features fixed. Averaging these differences across the sample yields approximate Average Treatment Effects (ATEs) for distance, weather, roof, pressure, and kicker skill. Rather than presenting a long table, we summarize key magnitudes in the Results and Conclusions.

4 Results

4.1 Predictive Performance

Across individual models, GAM, LightGBM, and BART form the top tier. On the 2024 hold-out season, typical performance is:

- **GAM:** Brier ≈ 0.111, AUC ≈ 0.76, low ECE.
- **LightGBM:** Brier ≈ 0.111, AUC in the mid 0.75 range.
- **BART:** best single model Brier, just under 0.110.

A simple equal-weight ensemble of GAM and LightGBM improves the Brier score slightly, and an optimized weighted ensemble of Bagging, LightGBM, GAM, and Logistic Regression achieves our best overall performance with Brier ≈ 0.1098, AUC around 0.76, and ECE on the order of 1–2%. The improvement in Brier score over the best single model is small but consistent across validation splits.

In practical terms, these Brier scores correspond to well-calibrated probabilities: when we group predicted probabilities into bins (e.g. 0.7–0.8, 0.8–0.9), empirical make rates closely match the predictions, especially in the high-density 30–55 yard range that coaches care about most.

4.2 Calibration and Distance Effects

Our distance-binned isotonic calibration substantially improves reliability. Before calibration, tree-based models tend to be slightly overconfident on long kicks and underconfident on chip shots. After calibration, reliability plots by distance look nearly monotone, and the empirical miss rate in each probability bin aligns closely with the predicted miss rate.

From a football perspective, the most important structural finding is that **each additional yard of distance reduces the make probability by roughly one percentage point** near common NFL distances. This effect dominates the influence of all other features. For example, moving from a 40-yard to a 50-yard attempt typically reduces make probability by more than 10 percentage points, a much larger shift than typical changes in temperature or wind.

The Cam Little 68-yard example illustrates the model's behavior in the tail. Even in ideal dome conditions and without extreme pressure, the model gives only about a 15% chance of success, and our residual analysis confirms that such extreme kicks are highly variable and poorly supported by historical data. The model captures the basic difficulty but cannot guarantee accuracy on one-off record-distance attempts.

4.3 Causal Insights

The BART-based causal analysis supports several intuitive but important claims:

- **Distance is the dominant causal driver.** Increasing distance by one yard reduces make probability by roughly 1.1 percentage points on average, far larger than any other per-unit effect.
- **Pressure matters, but modestly.** Moving from regular season to playoffs or from low- to high-pressure (late game) situations produces only small causal decreases in success probability (on the order of 1–3 percentage points).
- **Stadium conditions help at the margins.** Switching from outdoors to a dome increases make probability by around 2 percentage points due to the removal of wind and weather.
- **Weather effects are noisy and driven by selection.** Snow and heavy rain show counterintuitive positive ATEs in some specifications, a sign that only strong kickers attempt those kicks and that sample sizes are tiny. This reinforces that extreme-weather predictions are inherently unreliable.

Kicker career field goal percentage and attempt counts have relatively small causal effects once distance and context are controlled for, suggesting that much of the "kicker skill" signal is already encoded through the types of kicks each player is asked to attempt.

4.4 Decision-Level Case Studies: Three High-Stakes Kicks

To complement our aggregate metrics, we evaluated the model on three real, high-stakes kicks drawn from recent NFL and college games. These examples were used in our in-class interactive demo and illustrate both the strengths and weaknesses of our system.

Kick 1: 52-yard playoff game winner (Evan McPherson). This attempt is a long but not absurdly long kick: 52 yards, on turf, at roughly 35°F with low wind. It is a playoff game-winning situation, so the pressure is extremely high, but other environmental factors are relatively normal. Our ensemble assigned this kick a make probability of 81.54%. This value is materially higher than the unconditional 50+ yard make rate (67.8%), reflecting both the kicker's strong track record (87.5% season FG%) and the absence of adverse weather. Qualitatively, this probability aligns with how many observers would describe the kick: difficult but very much in the kicker's range.

Kick 2: 43-yard playoff kick in heavy wind (Cody Parkey). The second case is a 43-yard attempt—a distance where league-wide make rates are usually around 80–90%—but under much worse conditions. The game is in the playoffs, the temperature is just under 40°F, and the wind is very strong. The kicker is experienced (121 career attempts) with a slightly below-average make rate (84.3%). Our model nevertheless produced a high predicted probability of 88.81%. In hindsight, this case highlights one of our main failure modes: the model tends to underweight rare combinations of high wind, cold, and pressure relative to the more common "normal" 40–49 yard attempts that dominate the training data. This helps explain why, as we note in the conclusion, the system performed poorly on two of our three extreme case studies.

Kick 3: 68-yard record-attempt in a dome (Cam Little). The final case is a 68-yard attempt indoors, at altitude, by a young kicker with an 86% make rate on 43 prior attempts. If successful, the kick would have set a record. The dome conditions remove rain, snow, and wind, but the distance is far outside the support of normal NFL behavior. Our ensemble assigned this kick a make probability of 15.42%. On one hand, this is directionally reasonable: it is much lower than typical 50+ yard attempts, reflecting the extreme distance. On the other hand, the model is effectively extrapolating here, because there are very few historical examples anywhere near 68 yards. This case reinforces our broader point that the model is not an oracle in the far tail and that coaches should treat such outputs as rough signals rather than precise probabilities.

Across these three scenarios, we see the same pattern emphasized by our global metrics:

- In "normal" but high-stakes situations (Kick 1), the model produces probabilities that are consistent with both historical rates and football intuition.
- In rare, structurally different situations (Kicks 2 and 3), small misspecifications in how we handle wind, pressure, and extreme distance translate into visibly miscalibrated predictions.

These case studies therefore provide qualitative evidence for our main conclusion: the calibrated ensemble is highly useful for everyday coaching decisions but should be applied with caution in the extreme tail where data are sparse.

5 Current Conclusions and Future Work

5.1 Conclusions

Our main conclusions are:

- **Reliable probabilities for typical kicks:** Using a calibrated ensemble of GAM, tree-based models, and logistic regression, we can predict field goal success with good Brier scores and excellent calibration on standard NFL kicks between roughly 30 and 55 yards.
- **Distance dominates decision-making:** Causal and predictive analyses agree that distance is the primary driver of success probability; other factors like weather, pressure, and stadium type matter but typically only move the needle by a few percentage points.
- **Extreme situations remain uncertain:** Very long kicks, unusual weather, and rare high-pressure moments have little historical data, leading to unstable estimates. In these situations, analytics should inform but not dictate coaching decisions.

Overall, we answer the project question in the affirmative: **for the majority of real-world NFL situations, field goal success can be predicted well enough to support data-driven 4th-down and end-of-game strategy.**

5.2 Limitations

Key limitations include the lack of tracking data (snap and hold quality, ball trajectory), the coarse nature of publicly available weather summaries, and the scarcity of extreme scenarios in the historical record. Our reliance on play-by-play data also means we cannot directly observe kicker health or psychological state.

5.3 Future Work

Several directions could substantially improve this system:

- Incorporating player tracking data and ball-flight metrics to distinguish clean kicks from mishits and blocked attempts.
- Extending the historical window backward and across leagues to better populate the extreme tail of long-distance attempts.
- Building hierarchical kicker-level models that shrink inexperienced kickers toward league averages while allowing established veterans to deviate.
- Integrating our probabilities directly into a decision model that jointly optimizes whether to kick, punt, or go for it, enabling richer "what-if" analysis for coaches.

6 References

References

[1] A. Baldwin, R. Yurko, and S. Ventura, *The nflfastR Project: Reproducible NFL Analytics with R*, `https://github.com/nflverse/nflfastR`, 2021.

[2] M. Lopez, G. Matthews, and B. Baumer, *How Often Does the Best Team Win? A Unified Approach to Understanding Randomness in North American Sport*, *Annals of Applied Statistics*, 12(4):2483–2516, 2018.

[3] R. Yurko, S. Ventura, and M. Horowitz, *Going for it: Fourth Down Behavior in the NFL*, *Journal of Quantitative Analysis in Sports*, 15(4):289–308, 2019.

[4] H. A. Chipman, E. I. George, and R. E. McCulloch, *Bayesian Additive Regression Trees*, *Annals of Applied Statistics*, 4(1):266–298, 2010.