**Field Goal Success Probability Model**

**Purpose**:

Estimate the intrinsic difficulty of a field goal attempt using only objective, kicker-agnostic inputs. The model answers the question, “How hard is this kick inherently?” rather than “How likely is this specific kicker to make it?” or “Was it a good coaching decision?” By focusing solely on physical factors like kick distance, wind, temperature, and surface, it avoids bias from game context or individual performance history. The result is a clean, interpretable baseline that supports pregame prep, scouting, and in-game decision-making.

**Data & Feature Engineering**  
We used field goal play-by-play data from nflfastR. After filtering for valid attempts, we parsed messy weather strings into structured variables and filled missing values using default indoor conditions or available outdoor data. To simplify modeling and improve interpretability, continuous weather variables like temperature and wind were binned into logical categories based on inflection points in accuracy rates and football logic. For example, temperatures below 50°F were labeled “Cold” due to a noticeable drop in make rate. We excluded game context features such as leverage, quarter, and playoff status to avoid strategic bias and keep the model focused on intrinsic kick difficulty. Finally, we added interaction terms like Wind × Distance and Cold × Wind to help the model capture realistic effects (i.e it makes sense that wind impacts longer kicks more than short ones).

**Modeling Approach**

* Final Model: Logistic Regression (SMOTE, interaction terms, threshold = 0.5)
* Key Inputs: Kick distance, weather (binned temperature and wind), field surface, altitude, precipitation, humidity, interaction terms like Wind × Distance and Cold × Wind
* Target: Binary (Make / Miss), ~84% base rate

We tested multiple classification models, including logistic regression and XGBoost. To address class imbalance, we used SMOTE with logistic regression and class weighting for XGBoost. Logistic regression offered interpretability and allowed us to understand the impact of individual features, while XGBoost captured non-linear patterns but struggled with overpredicting makes. Including interaction terms added signal without hurting calibration. We evaluated multiple thresholds and selected 0.5, which provided the best tradeoff between sensitivity and specificity.

To improve probability accuracy, we applied Platt scaling for post-hoc calibration, aligning predicted probabilities with real-world outcomes — especially on difficult kicks.

Final model coefficients aligned with football logic:

Longer kicks, precipitation, wind, and cold-wind combinations all decreased make probability. Night games and cold weather showed slight positive effects, potentially reflecting preparation or sample bias. We retained some statistically insignificant variables based on domain knowledge, potential future value, and contribution to overall model stability.

**Performance Summary**

**Include table**

We used a mix of metrics to evaluate performance: AUC to measure separation between makes and misses, Brier Score and Log Loss to assess calibration, and Sensitivity, Specificity, and Balanced Accuracy to evaluate performance across both classes given the 84% make rate.

While XGBoost showed stronger raw calibration, it failed to identify tough kicks — predicting nearly every attempt as a make. Logistic regression with SMOTE struck the best balance: strong AUC, much better class balance, and clear interpretability.

We applied Platt scaling to recalibrate predicted probabilities because the original model tended to underestimate the chance of success on difficult kicks. Calibration improved the Brier Score from 0.2041 to 0.1148 and aligned predictions more closely with actual outcomes.

**Model Scope and Use**

This model provides a reliable, kicker-agnostic estimate of field goal difficulty based solely on physical conditions. It’s best used as a baseline for scouting, in-game decisions, or pregame planning when you need a clean read on how tough a kick truly is. It intentionally excludes kicker history, game context, and execution errors like bad snaps or pressure, which are important for post-mortems but can bias real-time assessments. Those factors can be layered on later, but the core model is built to be modular, interpretable, and focused on the kick itself — not the decision or the outcome.