Hierarchical Bayesian Modeling for NFL Kicker Performance: A Production-Ready Analytics Pipeline

Author: Geoffrey Hadfield  
Date: July 2025  
Repository: <https://github.com/ghadfield32/nfl_field_goal_kicker_leaderboard>

**Abstract**

Evaluating NFL kickers poses unique challenges: sparse data, varying sample sizes, and the need for confidence estimates. In this paper, I describe a hierarchical Bayesian logistic regression model that addresses these issues through partial pooling and natural uncertainty quantification. I transform raw field-goal attempts into EPA-FG+ ratings with credible intervals, compare predictive performance against Random Forest and XGBoost baselines (AUC: 0.864 vs. 0.872), and highlight why simpler age splines outperform more complex feature sets. Finally, I outline a production-ready pipeline—integrating MLflow, FastAPI, and a React/Vite frontend on Railway—to deploy our model at scale.

**1. Introduction**

NFL kickers face wildly different workloads: veterans attempt hundreds of kicks, rookies only a handful. Raw field-goal percentage ignores attempt difficulty, and point estimates offer no measure of confidence. These gaps hinder coaches and general managers who need robust, interpretable metrics for roster decisions and in‑game strategy.

I build a hierarchical Bayesian logistic regression model that:

* Uses partial pooling to stabilize estimates across kickers with varying data.
* Quantifies uncertainty with full posterior distributions.
* Adjusts for kick distance, age, experience, and recent performance.
* Produces EPA-FG+ ratings with 90% credible intervals.

Key contributions:

1. A reproducible, production-ready Bayesian pipeline.
2. A bias–variance analysis showing why moderate feature sets beat overly complex splines.
3. Benchmarking against tree-based models.
4. An MLOps roadmap for deployment on modern infrastructure.

**2. Data and Preprocessing**

**2.1 Data Sources**

I use NFL play-by-play data from 2016–2018, focusing on regular-season field goals. My pipeline filters out:

* Non-regular-season snaps
* Attempts under 18 or over 63 yards
* Kickers with fewer than five attempts (to avoid numerical instability)

**2.2 Feature Engineering**

I create interpretable predictors that capture key drivers of success:

**Distance metrics:**

* yards (standardized)
* yards\_squared
* long\_attempt (>50 yards flag)

**Age & Experience:**

* age\_c (centered at 30, scaled)
* age\_c2 (quadratic term)
* attempt\_count (career attempts, scaled)

**Context:**

* season\_progress (0–1)
* rolling\_success (exponential decay)

**3. Exploratory Analysis**

In this section, we visualize key patterns in the cleaned dataset of 8,703 regular-season attempts (2010–2018) and summarize the main findings:

**3.1 Outcome Breakdown**

*Figure 3.1: Distribution of made (84.4%), missed (13.5%), and blocked (2.2%) kicks.*

**Takeaway:** The dataset is dominated by successful kicks, but blocked kicks—while few—warrant separate treatment in modeling or filtering.

**3.2 Distance Distributions and Success Curve**

*Figure 3.2a: Histogram of field-goal attempt distances, showing peaks around 30–50 yards.*

*Figure 3.2b: Success rate declines non-linearly with distance; bubble size ∝ number of attempts. A quadratic fit (red dashed) captures the sigmoid-like drop-off.*

**Summary of distance analysis:**

* 18–29 yards: 97.0% success (2,284 attempts)
* 30–39 yards: 90.0% success (2,552 attempts)
* 40–49 yards: 77.9% success (2,621 attempts)
* 50–59 yards: 64.5% success (1,207 attempts)
* 60+ yards: 28.2% success (39 attempts)

**Takeaway:** Distance is the dominant predictor (corr = –0.685), with a sharp drop beyond 50 yards.

**3.3 Outcome vs. Distance Distributions**

*Figure 3.3: Boxplots of attempt distance by outcome category.*

**Observation:** Made kicks center around 35 yards, while missed kicks extend to longer distances; blocked attempts span a middle range but are infrequent.

**3.4 Kicker-Level Activity and Performance**

*Figure 3.4a: Distribution of total attempts per kicker (median ~90).*

*Figure 3.4b: Distribution of raw success rates for kickers with ≥20 attempts (median ~84%).*

**Takeaway:** Broad variation in both sample size and raw accuracy underscores the need for partial pooling in a hierarchical model.

**3.5 Temporal and Demographic Trends**

*Figure 3.5: (Left) Success rate by season (2010–2018). (Center) Average attempt distance by season. (Right) Histogram of kicker ages at attempt (mean ~30 years).*

* **Seasonal success** hovers between 82–87% with no strong trend.
* **Average distance** gradually increases from ~36.5 to ~38.4 yards.
* **Age distribution** peaks around 27–33 years.

**Takeaway:** Temporal and age effects exist but are minor compared to distance and individual kicker effects.

**3.6 Feature Correlation**

*Figure 3.6: Correlation among features: distance and distance² are near-perfectly correlated (0.99), age and experience moderately correlated (0.42), others weak.*

**Takeaway:** Strong distance collinearity motivates inclusion of a squared term with care; other predictors show low multicollinearity.

**4. Model**

I specify a hierarchical Bayesian logistic model:

y\_i ~ Bernoulli(p\_i)

logit(p\_i) = α\_j[i] + β1\*yards\_i + β2\*yards2\_i + β3\*age\_c\_i + β4\*age\_c2\_i + β5\*exp\_i

α\_j ~ Normal(μ, σ)

μ ~ Normal(0,5)

σ ~ HalfNormal(1)

β\_k ~ Normal(0,1)

Partial pooling shrinks low-sample kickers toward the population mean while letting high-sample kickers express their individual skill.

**4.1 Feature Selection and Column Rationale**

For our hierarchical Bayesian model, we intentionally kept the feature set lean—picking only those variables with clear interpretability, low risk of multicollinearity, and strong domain relevance:

| **Column** | **Role** | **Reasoning** |
| --- | --- | --- |
| attempt\_yards | Distance (standardized) | Primary driver of success; we z-score to aid sampler convergence and interpretability. |
| distance\_squared | Distance² | Captures non-linear drop-off in make probability; minimal co­linearity after centering. |
| is\_long\_attempt | (>50 yds) flag | Encodes extreme attempts where logistic slope may differ. |
| age\_c | Centered age | Allows us to detect age trends around the league-average age (30 yrs). |
| age\_c2 | Age² (quadratic) | Models potential curvature in age effect. |
| exp\_100 | Scaled experience (career attempts) | Captures learning curve and veteran steadiness; standardized to zero mean/unit variance. |
| season\_progress | Fraction through season (0–1) | Reflects potential clutch or fatigue effects as season advances. |
| rolling\_success | Recent performance (exp. decay) | Adjusts for momentum or slumps; exponential decay emphasizes the last few attempts. |

**5. Feature Complexity Study**

I compare three feature sets:

1. **Baseline (3 features)**: AUC 0.752
2. **Moderate (5 features)**: AUC 0.863, condition ~230
3. **Complex splines (9 features)**: AUC 0.862, condition ~4,500 (unstable)

Complex age splines introduce multicollinearity (correlations >0.99) and convergence issues. The moderate set hits the sweet spot.

**6. Comparison to Tree Models**

| **Model** | **AUC-ROC** | **AUC-PR** | **Log Loss** | **Brier** | **ECE** | **Accuracy** | **Precision** | **Recall** | **F1** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Bayesian hierarchical** | 0.7975 | 0.9606 | 0.3151 | **0.0946** | **0.0250** | **0.8755** | 0.8750 | 1.0000 | 0.9333 |
| **Simple Logistic** | 0.8049 | 0.9637 | 0.3240 | 0.1005 | 0.0379 | 0.8692 | 0.8688 | 1.0000 | 0.9298 |
| **Ridge Logistic** | 0.7881 | 0.9562 | 0.4630 | 0.1408 | 0.1908 | 0.8692 | 0.8688 | 1.0000 | 0.9298 |
| **Random Forest** | 0.7878 | 0.9584 | 0.3410 | 0.1030 | 0.0250 | 0.8663 | 0.8663 | 1.0000 | 0.9283 |
| **XGBoost** | 0.8056 | 0.9621 | 0.3238 | 0.0990 | 0.0455 | 0.8634 | 0.8659 | 0.9966 | 0.9267 |
| **CatBoost** | 0.7423 | 0.9486 | 0.3911 | 0.1119 | 0.0631 | 0.8576 | 0.8673 | 0.9866 | 0.9231 |

While tree models slightly edge the Bayesian approach on raw AUC, only the Bayesian model provides native credible intervals and better-calibrated probabilities.

**7. EPA-FG+ Ratings**

I simulate 10,000 kicks from the empirical distance distribution, compute expected points (3 × P(make)), and subtract the league average. Using 4,000 posterior samples, I report 5th, 50th, and 95th percentiles for each kicker’s EPA-FG+.

Example: A kicker with EPA-FG+ = +0.15 adds 4.5 points over 30 attempts.

**8. Pipeline Overview**

graph TD

A[Raw CSV data] --> B[Preprocess & Features]

B --> C[EDA]

C --> D[Bayesian Model Training]

D --> E[Posterior → EPA-FG+]

E --> F[Streamlit Dashboard]

D -. Benchmark .-> G[RF / XGB]

I use Python 3.9+, PyMC 5, Pandas, Matplotlib, and Streamlit. Models and predictions are versioned via MLflow; inference is served with FastAPI and a React/Vite frontend on Railway.

**9. Diagnostics**

* **R̂ < 1.01** and ESS > 400 for all parameters.
* Posterior predictive checks show simulated make rates match observed within 95% CIs.
* Time-series cross-validation confirms generalizability from 2016–2017 → 2018.

**10. Business Impact**

**Coaches:** Real-time success probabilities with confidence—e.g., “85% ± 5% from 47 yards.”  
**GMs:** Data-driven contract evaluations and roster decisions.  
**Risk:** Rookies show wide intervals; veterans are tightly estimated.

**11. Roadmap**

Future work includes integrating weather, biomechanical data, and stadium effects; automating CI/CD with MLflow and GitHub Actions; and deploying a full React dashboard for sideline use.

**12. Conclusion**

Hierarchical Bayesian modeling turns sparse, unbalanced kicker data into actionable insights. By embracing uncertainty, we deliver calibrated, interpretable metrics that outperform point-estimate models in real-world decision-making. Our production-ready pipeline ensures these insights can scale from prototype to sideline in real time.