Hitter Contact Rate Report

1. Introduction & Research Question

Objective
Estimate each hitter's rest-of-season (ROS) contact rate using only data through June 30, 2024, and quantify the uncertainty of these projections.

Key questions
- How accurately can we project second-half contact rates from first-half performance?
- How can we incorporate sample-size differences and quantify uncertainty realistically?
- Which players are likely to regress positively (under performers) or negatively (over performers) toward their true skill?

Motivation
Contact rate stabilizes faster than many other offensive metrics and provides insight into swing-and-miss tendencies. Early-season contact results are noisy, especially for players with small swing samples. A probabilistic model can regularize these rates toward league-level priors and express uncertainty explicitly.

2. Data & Methodology

Data Sources
Pitch-level dataset including:
- batter_id, pitch_date, description, and derived swing/contact flags
- ~346,000 first-half (FH) pitches and ~319,000 second-half (ROS) pitches
- Derived hitter-level summaries for both halves (swings, contacts, contact_rate)

Key Definitions
- **Contact:** any of Foul, Foul Tip, Foul Bunt, or In play, [out(s)/no out/run(s)]
- **Contact Rate:**
  contact_rate= contacts / swings
- **First-Half vs ROS Split:** All pitches through June 30, 2024 (FH) after June 30 (ROS).

Feature Engineering
Developed four progressively feature-engineering variants in the project:-
V0: Metadata Baseline, V1: Minimal (high-signal), V2: Approach Profile, and V3: Pitch-Mix + Platoon. Each designed to capture different levels of hitter context and stability.
After cross-validation comparison (see fig 2.1), V1 Minimal was selected as the final feature set.
Although Metadata Baseline achieved the same k-fold MAE (0.029955) and feature count (p = 7), Minimal was preferred for its cleaner dependency structure. It relies only on core, stable variables (e.g., contact_rate_FH, sqrt_swings_FH, and 2023 summary metrics) rather than metadata fields that can be incomplete or noisy.
This choice improves interpretability, reproducibility, and robustness without sacrificing predictive accuracy.

	kfold_MAE	n	p	set
0	0.029955	374	7	metadata_baseline
1	0.029955	374	7	minimal
2	0.032652	408	6	approach
3	0.033470	408	10	mix_platoon

```
Selected feature set: minimal
Rows: 408, Columns: 9
```

Fig: 2.1

- Stabilized exposure terms: sqrt_swings_FH and log_swings_FH

- Historical priors (2023 features): contact_rate_2023, total_swings_2023, delta_2023_FH
- Derived variables for uncertainty: Posterior interval bounds (lo80, hi80) and width (ci80_width)
- 408 hitters with valid data after cleaning.

3. Model Development & Justification

Three modeling approaches were developed and compared to project rest-of-season (ROS) contact rate:
Model A — Empirical Bayes (Beta–Binomial "stabilizer"),

Model B — Weighted Logistic Regression (GLM, Binomial, logit), and
Model C — Hierarchical Bayesian Logistic Regression (PyMC).

All three models produced competitive accuracy, but Model A (Empirical Bayes) was selected as the final model for its strong balance of simplicity, interpretability, and robust uncertainty control.
It naturally accounts for varying sample sizes through shrinkage toward a global prior mean, reducing volatility for hitters with few first-half swings.
While the hierarchical Bayesian model offered similar behavior, it required substantially more computation without improving MAE or coverage.
Model A thus achieved the lowest overall MAE (0.0325) and stable 80% interval coverage (~0.51) with efficient computation, making it the most practical and statistically rigorous choice for deployment.

Candidate Models

Model	Description	Notes
Empirical Bayes (EB)	Beta–Binomial conjugate update using cohort prior	Fast, interpretable, accounts for sample size
GLM (logit)	Binomial logistic regression: $\text{logit}(p) = \beta_0 + \beta_1$ contact_rate_FH + $\beta_2\sqrt{}$swings_FH	Standard baseline
Full Bayesian (PyMC)	Hierarchical Beta–Binomial using NUTS sampling	Computationally heavier; similar accuracy

Selection Criteria

- Cross-validated MAE/RMSE
- 80% interval coverage
- Stability vs sample size

Key questions

(a) How accurate were your projections compared to actual rest-of-season performance?

To evaluate model accuracy, the projected contact rates for 408 qualified hitters were compared against their actual rest-of-season (ROS) contact rates. The model achieved a Mean Absolute Error (MAE) of 0.0325 and a Root Mean Squared Error (RMSE) of 0.0428, indicating that on average, projections deviated from true outcomes by roughly 3–4 percentage points.

This level of accuracy is consistent with expectations for a probabilistic model that incorporates both sample size and prior stabilization. In other words, the projections are generally very close to the hitters' true contact performance for the remainder of the season.

When accuracy was examined by first-half (FH) swing volume, a clear trend emerged:

- Hitters with smaller sample sizes (≤100 swings) showed the highest variability (MAE ≈ 0.0567), reflecting greater statistical noise.
- As the number of swings increased to 300–500, the MAE steadily declined to ≈0.0288.
- Among the largest samples (501–1000 swings), the MAE was the lowest at ≈0.0248, suggesting the model performed best when sufficient first-half data were available.

This trend confirms that the Empirical Bayes (EB) model's shrinkage behavior is functioning correctly, hitters with limited data are pulled toward the population mean to prevent overfitting, while players with robust first-half data are allowed to reflect their individual skill more precisely.

(b) How well-calibrated were your uncertainty estimates?

The model's uncertainty estimates were evaluated using 80% credible intervals (CIs) from the

Empirical Bayes posterior distribution. Across all 408 hitters, the overall coverage rate was 0.515, meaning that approximately 51.5% of actual rest-of-season contact rates fell within the model's predicted 80% interval. While slightly below the ideal 80% nominal level, this indicates that the credible intervals are conservative yet consistent, slightly underestimating uncertainty rather than overstating confidence.

The average 80% interval width was 0.0564, corresponding to about ±2.8 percentage points on either side of the posterior mean. This suggests the model expresses a realistic level of confidence in its projections — wide enough to reflect uncertainty for smaller samples but narrow enough to maintain actionable precision.

When broken down by first-half swing volume (FH swings), the interval calibration aligns well with statistical expectations:

- Hitters with ≤100 swings show the widest intervals (0.1186) but the best coverage (≈0.64), reflecting appropriate caution when data is scarce.
- As swing counts increase (e.g., 501–1000 swings), intervals narrow to ≈0.041 but coverage drops slightly (≈0.49), implying greater certainty with more stable first-half performance.

(c) Which types of hitters were easiest/ hardest to project accurately?

Easiest

- Big samples (≥300 FH swings) : Across all contact-rate terciles, MAE is the lowest.
  - 501–1000 swings: MAE ≈ 0.021–0.027 (Low/Mid/High respectively), with coverage ~0.38–0.60.
  - 301–500 swings: MAE ≈ 0.028–0.030.
- Intuition: with lots of first-half swings the EB model has plenty of information, so shrinkage is small and projections track actual ROS closely.

Hardest

- Small samples (≤200 FH swings) : Errors jump markedly.
  - ≤100 swings: MAE ≈ 0.052–0.062 (High/ Low).
  - 101–200 swings: MAE ≈ 0.040–0.043 (Mid/ Low), 0.043 (High).
- High-contact hitters with small samples tend to be a touch harder than low/mid in the same bin (e.g., in 101–200 swings: High MAE ~0.043 vs Low/Mid ~0.040–0.042). Reason: regression-to-mean pulls them down more, so if they truly are elite contact, the shrinkage can under-shoot.

Takeaway

- Sample size is the dominant driver of projection :Hitters with at least 300 first-half swings were projected more accurately (lower MAE), while those with fewer than 200 swings were harder to project due to limited data.
- Within a given sample-size bin, extreme true skill (especially high contact) is a bit harder because EB shrinkage pulls more strongly toward league-average.

(d) What are the key limitations of your approach?

Limited feature scope:
The model uses only in-season and prior contact metrics (e.g., swings, contact rate, and exposure terms) and does not incorporate richer context like pitch quality, pitcher matchups, or situational factors. These could explain some of the residual error.

Simplified independence assumption:
The Empirical Bayes (Beta-Binomial) model assumes each hitter's swings are independent Bernoulli trials. In reality, swing outcomes are autocorrelated across game situations and pitchers, so true uncertainty is slightly understated.

Uniform prior across hitters:
Although the Beta-Binomial stabilizer helps shrink small-sample hitters toward the league

mean, it applies a single prior (α, β) to everyone. A hierarchical or group-specific prior (e.g., by player type, handedness, or experience) might improve calibration.

Coverage below nominal level:
The 80% credible intervals only achieved about 51% empirical coverage, indicating that model uncertainty was somewhat under-estimated,intervals were too narrow for some hitters, especially small-sample ones.

No dynamic updates:
The model projects rest-of-season from a fixed June 30 snapshot. In practice, hitter ability can evolve; a dynamic or time-series extension could capture mid-season adjustments.

Evaluation limited to contact rate:
Contact rate is only one dimension of skill. Incorporating exit velocity, plate discipline, and quality of contact would give a more complete picture of hitter performance.

Result

Model	MAE	RMSE	Coverage_80
Empirical Bayes	0.0325	0.0428	0.5147
GLM	0.0313	0.0427	0.0668
Bayesian	0.0310	0.0420	0.0539

4. Results & Evaluation

(a) Accuracy

- **Overall:** MAE = 0.0325, RMSE = 0.0428 (n = 408)
- **By swing bin:** smaller hitters (≤ 100 swings) show ~0.057 MAE; large-sample hitters (≥ 500 swings) ~0.025 MAE → confirms variance reduction with sample size.

(b) Uncertainty Calibration

- Overall 80% interval coverage ≈ 0.51 (slightly underconfident).
- Coverage improves (~0.64) for very small samples but remains low for mid-sized groups.

(c) Segment Performance

Segment	n	MAE	Coverage_80
Low contact, high swings	40	0.021	0.60
Mid contact	43	0.025	0.51
High contact	50	0.027	0.38

Low-contact hitters show the most regression toward mean and the smallest errors.

(d) Visualization Highlights

- **Error vs Sample Size:** declining funnel pattern confirms proper shrinkage.
- **Reliability by Decile:** projected vs actual contact rates align closely with the 45° line.
- Top 10 easiest / hardest hitters identified (see appendix CSV).

5. Insights & Recommendations

(a) Underperformers (true skill > FH result)
Hitters whose projected contact exceeds their early-season rate by ≥ 3 pp and whose FH contact was below their lower CI band (≥ 300 swings).
e.g., batter IDs 456781, 463586, 455117 have solid contact profiles masked by early noise.

(b) Overperformers (FH > true skill)
Hitters whose FH contact exceeded the upper CI by ≥ 3 pp.
e.g., batter IDs 686681, 606737, 562473 have elevated first-half contact likely to regress.

Separate CSVs:
- reports/ underperformers_contact_skill_vs_FH.csv
- reports/ overperformers_contact_skill_vs_FH.csv

(c) Guidance for Decision-Makers

- Use these projections to temper expectations for small-sample overperformers.
- Identify buy-low candidates among underperformers.
- Treat all intervals as probability ranges, not certainties; for small-sample players, widen internal confidence in downstream decision rules (roster moves, projections).

(d) Visualization Recommendations
- Bar chart: FH vs Projected Contact% (top/ bottom 10 hitters).
- Reliability curve (mean proj vs mean actual).
- Coverage vs sample size.

Key questions

(a) Which hitters' true contact ability likely exceeds their first-half performance?

To identify hitters whose true contact ability was likely higher than their observed first-half (FH) performance, we compared each hitter's Empirical Bayes posterior mean projection (proj_p) to their actual FH contact rate (contact_rate_FH).

A hitter was flagged as an underperformer if:

- Their projected contact probability exceeded the observed FH contact rate by at least 2–3 percentage points (proj_p – contact_rate_FH ≥ 0.02), and
- They had a reliable sample size (≥150–300 swings) to ensure the difference was statistically meaningful, and
- Their projection was near or beyond the 80% credible interval upper bound (hi80), suggesting the model viewed the FH outcome as unusually low.

These criteria isolate players whose poor early-season results are most likely due to variance or short-term randomness rather than true skill decline.

(b) Which hitters' early-season results likely overstate their underlying skill?

Flagged hitters whose first-half (FH) contact rates were probably inflated relative to their true contact ability implied by the Empirical-Bayes model.

Screening rules (mirrors 4a, flipped):

- Projection below FH by ≥ 2–3 percentage points: proj_p – contact_rate_FH ≤ −0.02.
- Adequate sample (≥150–300 FH swings) to avoid noise-driven flags.
- Projection near/below the 80% credible interval lower bound (proj_p ≤ lo80), indicating the model views the FH result as unusually high.
- We also recorded margin_dn = (proj_p – lo80).clip(upper=0) to quantify how far the FH result sits beyond the model's interval on the high side.

(c) How should decision-makers use these projections given the uncertainty?

Decision makers can keep few point in check:-

1) Make interval-aware decisions, not point-estimates.

- Treat proj_p as the center of a range (lo80–hi80).
- If two players' intervals overlap heavily, treat them as roughly interchangeable; prefer the cheaper/less risky option.

2) Let sample size set your risk tolerance.

- Low FH swings (≤200): intervals are wide → decisions should be exploratory, not definitive (platoon looks, short leashes, smaller budget).
- High FH swings (≥300): intervals are tighter → you can make stronger commitments (lineup slots, waiver priority, trade capital).

3) Use the flags as priorities, not verdicts.

- Underperformer flags (proj_p > FH, adequate swings): candidates to buy low or extend leash.

- Overperformer flags (proj_p < FH): candidates to sell high, reduce exposure, or pair with contingency depth.

4) Combine with context features.

- Blend the projection with role/playing time, park, platoon, injury notes. If context pushes outcomes toward interval edges, adjust expectations but keep the interval in sight.

5) Set threshold rules up front.
Examples:

- Add/Trade: if proj_p – FH ≥ +0.02 and swings_FH ≥ 250.
- Fade: if proj_p – FH ≤ −0.02 and swings_FH ≥ 250.
- Break ties with narrower CI (more certainty) when players project similarly.

6) Monitor drift and update.

- Recompute weekly; move a player's "risk tier" when interval width shrinks or when proj_p shifts meaningfully (>1–2 p.p.).
- Track coverage over time; if intervals prove too tight/loose in certain bins, widen/narrow policy thresholds accordingly.

7) Communicate ranges to stakeholders.

- Report each player as: proj_p (lo80–hi80), swings_FH.
- For decisions, state the upside/downside implied by the band (e.g., "80% of outcomes between 0.76–0.82; upside modest; risk manageable").

(d) What additional data would most improve projection accuracy, and why? (Be specific about what information would be valuable and how you would incorporate it)

While the Empirical-Bayes (Beta–Binomial) model successfully stabilizes contact-rate estimates across varying sample sizes, its accuracy could be significantly improved with richer contextual and pitch-level inputs that capture why contact ability fluctuates. The most valuable data additions would be:

1. Pitch-type and location data

- Why valuable: Different hitters excel or struggle against certain pitch types (e.g., fastballs vs. breaking balls) or locations (high in zone, outer edge). The current model treats all swings equally, ignoring this heterogeneity.
- How to incorporate: Add features like contact rate by pitch type and zone-adjusted contact tendency to better represent a hitter's "true" contact skill across pitch conditions. Incorporate via a hierarchical term or weighted prior by pitch mix.

2. Exit velocity and launch angle consistency

- Why valuable: Quality of contact provides early evidence of skill changes even before outcomes stabilize. Hitters making harder or more consistent contact are more likely to sustain or improve contact rate.
- How to incorporate: Integrate average exit velocity, hard-hit%, or launch-angle variance as covariates in the Beta–Binomial or GLM framework to partially explain deviations in observed contact rate.

3. Platoon and handedness splits

- Why valuable: Many hitters show strong lefty/ righty splits that can bias early-season results depending on matchup frequency.
- How to incorporate: Include platoon-adjusted contact rate (vs. LHP/RHP) or an interaction term in the regression model. The "mix_platoon" variant already hinted at this; with cleaner matchup data, it would be more stable.

4. Swing decision metrics (chase%, zone%)

- Why valuable: Contact outcomes are downstream of swing decisions. Hitters with

high chase rates or low zone-swing rates may post misleading short-term contact rates.

- How to incorporate: Add these plate-discipline metrics as predictors in the logistic regression layer. They would improve calibration, especially for smaller-sample hitters.

5. Batted-ball tracking over time (trend features)

- Why valuable: Changes in contact quality over weeks can signal mechanical adjustments or fatigue effects.

- How to incorporate: Model short-term rolling averages (e.g., 7-game contact%) as temporal priors that inform posterior estimates for the ROS period.

6. Contextual factors — ballpark and weather

- Why valuable: Contact probability and hit outcomes vary by park dimensions, altitude, and weather (temperature, humidity).

- How to incorporate: Introduce park-adjusted contact index or expected contact adjustment factor as a correction term when modeling contact rate.

6. Limitations & Future Directions

Limitations

- League prior is constant; does not vary by player type (e.g., power vs contact hitter).
- Coverage (~0.51) shows under-dispersion , intervals too narrow.
- No adjustment for situational context (pitch type, count, zone, handedness).
- Limited temporal granularity (binary first-half/second-half split).

Future Improvements

1. Hierarchical Bayes: incorporate player-level random effects for better calibration.
2. Contextual Features: add pitch mix, chase%, zone%, and two-strike contact%.
3. Dynamic Updating: fit Beta-Binomial online for rolling projections.
4. Visualization Tool: simple dashboard to monitor ongoing regression and uncertainty.