Caleb Lee

Runner Advance Prediction (Problem 1)

Objective: Predict runner_advance as a calibrated probability in 0,1 for sacrifice-play opportunities and submit probabilities for runner_advance_predictions.csv. Primary evaluation is log loss; secondary is the soundness of method.

Data and leakage control.

- **Train:** 15,533 labeled opportunities. Target: runner_advance (1 = runner advanced, 0 = held or thrown out).
- **Tracking:** 62,574 outfielder tracking rows at sub-play resolution. To prevent plays with many tracking samples from dominating and to avoid target leakage, I aggregated tracking to one row per play_id (numeric mean, min, max), then broadcast those per-play features to each runner opportunity on that play.
- Dropped any outcome-like columns beyond the target per the glossary.

Features.

- **Context:** inning, score differential, balls, strikes, outs.
- **Situation:** runner_base, runner_on_3rd, two_outs, high_leverage, plus targeted interactions (for example runner_on_3rd × two_outs).
- **Ball off bat:** exit speed, launch angle, direction, hang time, distance, and simple bins (ground ball, line drive, fly ball, popup; weak/medium/hard/very hard; shallow/medium/deep).
- **Positioning (from tracking):** outfield depth and distance-to-home, plus aggregated pos_x/pos_y summaries (mean/min/max).
  Total features used in modeling: 47, including eight aggregated tracking features.

Validation and baselines.

- **CV protocol:** 5-fold GroupKFold on game_id so entire games stay within a fold and never appear in both train and validation.
- **Baselines:**
  - Constant prevalence: 0.6931 log loss (p = 0.499).
  - Out-of-fold contextual baseline that only uses runner_base, outs, and hit_trajectory: 0.5907 ± 0.0050.

Models and selection:
I compared three families on identical grouped CV with neg_log_loss:

- **RandomForest CV Log Loss: 0.3683 (+/- 0.0126)**
- **HistGradientBoosting CV Log Loss: 0.3269 (+/- 0.0220)**
- **GradientBoosting CV Log Loss:**

HGB is a strong fit for structured tabular data with mixed types and non-linear interactions, and it includes built-in regularization.

Calibration:
I evaluated isotonic and sigmoid post-hoc calibration on grouped CV. Both increased log loss slightly (isotonic 0.4041, sigmoid 0.4045), so I retained the uncalibrated HGB. Predicted probabilities show healthy dispersion (mean 0.521, SD 0.363).

Result:
Final model: **0.3269 log loss**, which is an absolute gain of **0.305** over the constant baseline (about **44%** reduction) and about **34–35%** better than the contextual baseline (**0.5907** → **0.3882**). Most influential signal groups align with baseball intuition: runner_base, ball-flight physics (distance, angle, exit speed), and outfield depth/positioning.

Reproducibility and deliverables.

- Code: notebook and script implement the full pipeline, including tracking aggregation, feature build, GroupKFold CV, model training, and export.

- Predictions: runner_advance_predictions.csv (probabilities for test_data.csv).

Next steps with more time and data.

1. **Richer context:** park effects, weather, batter spray tendencies, and fielder arm/relay timing.

2. **Entity modeling:** partial pooling or embeddings for runners, hitters, and fielders to stabilize sparse groups.

3. **Calibration at scale:** fold-wise stacked calibration or constrained reliability optimization; report reliability curves and Brier score.

4. **Temporal robustness:** rolling backtests by month/season to monitor drift and stability.

5. **Explainability:** global and local effect estimates for coaches and analysts (for example SHAP grouped by feature families).

-