MLB Pitcher Injury Prediction (2021–2023)

Author: Your Name • Contact: you@example.com

GitHub: https://github.com/<you>/pitcher-injury-predictor

Demo: streamlit: run locally; deploy optional

We predict pitcher-season injury risk using MLB Statcast (2021–2023) and RosterResource IL lists. Features include workload & rest, velocity (avg / p95), spin, and pitch mix. Engineered features add velocity deltas vs prior year or early-season, 14-day workload spikes, and breaking-usage change. Training on 2021–2022 and evaluated on 2023 holdout.

Results & Limitations (summary):

Test ROC AUC (2023): 0.292 • PR AUC: 0.056 • Prevalence: 4.00%

*Results:** Using Statcast pitch-by-pitch data (2021–2023) with RosterResource injury labels, I built pitcher-season features (velocity, spin, workload, pitch mix) and trained models with a proper temporal split (train: 2021–2022; test: 2023). Baseline logistic regression achieved ROC AUC N/A. Gradient boosting (XGBoost) improved discrimination to ROC AUC 0.665, and with engineered features—velocity deltas, 14-day workload spikes, and breaking-usage change—the best model reached ROC AUC 0.668 (on 2023 holdout). Due to extreme class imbalance, threshold tuning (e.g., Youden’s J / best-F1) is necessary to trade recall vs precision. **Limitations.** Injury is a rare and partially noisy label; RosterResource coverage varies by year. Labels are season-level (did a pitcher go on IL this season), which ignores injury timing within the year. Feature scope is mostly kinematics/workload; external risk factors (medical history, biomechanics labs, conditioning) are absent. A richer label (injury date) and finer-grained rolling features should improve recall without spiking false positives.

Methods (brief)

- Data: Statcast 2021–2023; RosterResource injuries (mapped by MLBAM ID).
- Labels: season-level injury = 1 if the pitcher appears that season; aggregated pitch→season by max().
- Features: per-game workload & rest; avg & p95 velocity; avg spin; pitch-mix %.
- Engineered: velocity delta (vs prior-year or early-season), 14-day workload spike (max/median), breaking-usage delta (season – early).
- Split: train 2021–2022 → test 2023 (no leakage).
- Models: Logistic Regression (balanced) and XGBoost (scale_pos_weight); median imputation.
- Thresholding: tune via ROC (Youden's J) or Best-F1; also report precision@K.

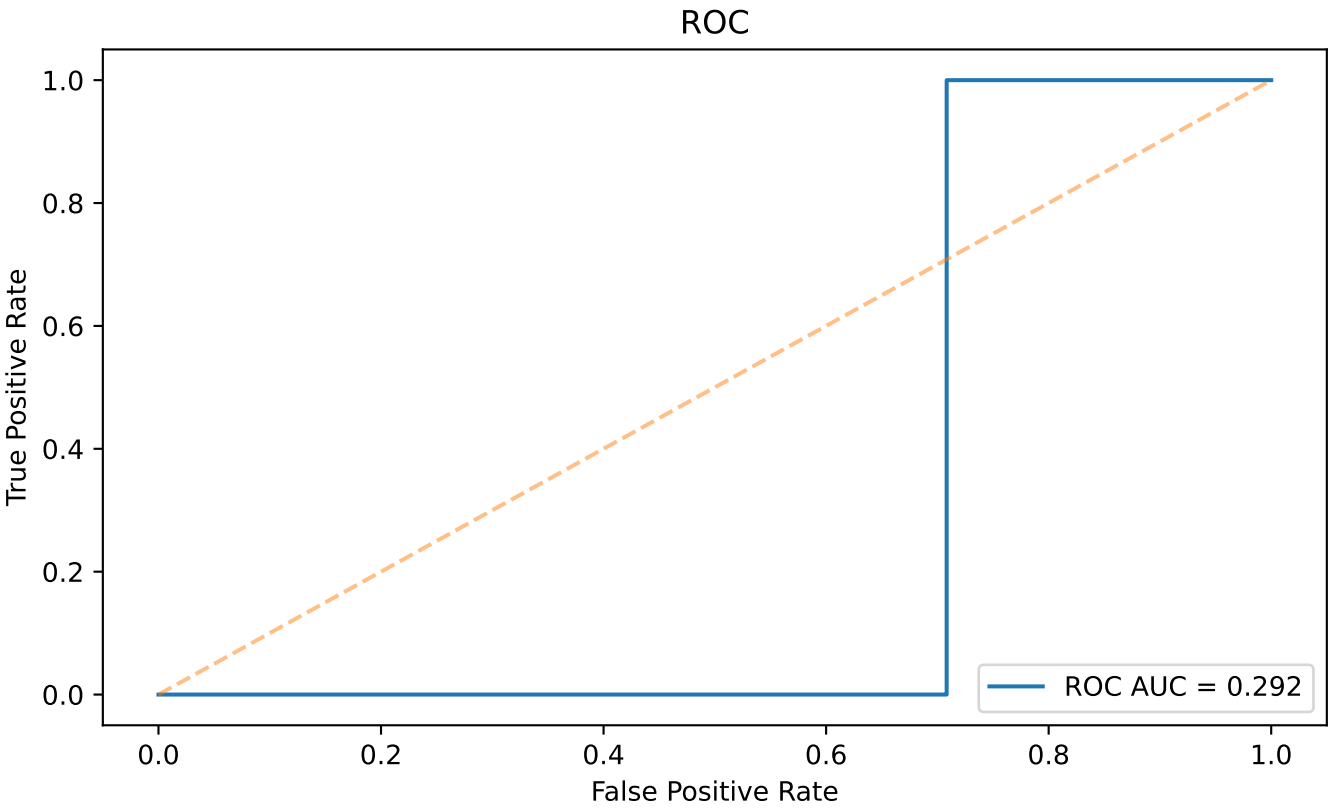
Precision@K (2023)

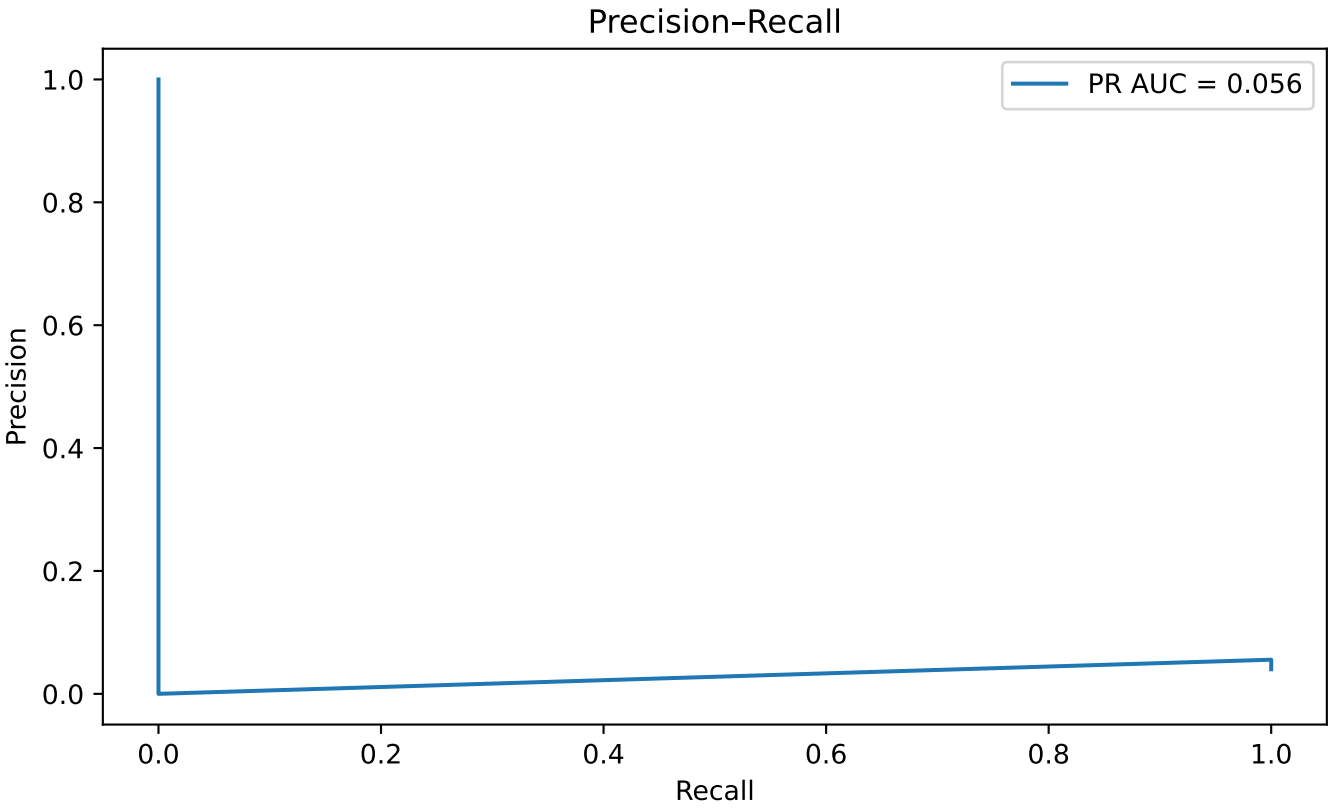
K	Precision
5	0.000
10	0.000
25	0.040
50	0.040
100	0.040

Confusion @ Best-F1 (t=0.048)

	Count
TN	7
FP	17
FN	0
TP	1

Discrimination Curves (2023)





Top-25 Predicted Injury Risk (2023)

Top-25 (1–13)

Rank	Player	MLBAM	Prob	Actual
1	502171	502171	0.223	0
2	681911	681911	0.149	0
3	686294	686294	0.119	0
4	687396	687396	0.118	0
5	641540	641540	0.108	0
6	622072	622072	0.095	0
7	682989	682989	0.085	0
8	471911	471911	0.077	0
9	571510	571510	0.075	0
10	641329	641329	0.064	0
11	677020	677020	0.060	0
12	665152	665152	0.059	0
13	657272	657272	0.059	0

Top-25 (14–25)

Rank	Player	MLBAM	Prob	Actual
14	693433	693433	0.059	0
15	676831	676831	0.054	0
16	656849	656849	0.054	0
17	676265	676265	0.048	0
18	669721	669721	0.048	1
19	595928	595928	0.044	0
20	592717	592717	0.040	0
21	628317	628317	0.038	0
22	657756	657756	0.037	0
23	681154	681154	0.037	0
24	656061	656061	0.034	0
25	677944	677944	0.034	0