
CINCINNATI REDS BASEBALL ANALYTICS
2025 STUDENT HACKATHON
Analytics
REDS

Overview

In this competition, we were tasked with estimating players' plate appearances and batters faced using variables we would have to generate exclusively from the provided Savant pitch-by-pitch dataset.

Intuition

Brainstormed key statistical factors influencing playing time

Identified and generated relevant statistical features

Focused on metrics that best predict future playing time

Variables Generated: Generic

- Player Age/Years in MLB at Season Start
- Player Height/Weight
- Plate Appearances + Batters Faced (Appearances)
- # of Stretches of 10+ Missed Days
- Avg # of Appearances over Previous Two Seasons
- Trend in Number of Appearances over Previous Two Seasons
  - 1 if Appearances Increased
  - 0 if Appearances Decreased
- # of Games Played over Previous Two Seasons
- Dummy binary variables for whether player was batter or pitcher

Variables Generated: Batters

- On-Base Percentage, Slugging Percentage, OPS as Batter
- Batter Weighted Averages (using lauch_speed_angle frequency) and Subsequent Percentiles in the Following Categories:
    - Launch Angle
    - Launch Speed
    - Hit Distance
    - Estimated Batting Average
    - Estimated wOBA

Weighted Avg Statistic	10th Percentile	25th Percentile	50th Percentile	75th Percentile	90th Percentile
w_avg_lnch_spd_wo_risp	71.54026316	83.73333333	87.44003268	89.5525	91.4
w_avg_lnch_spd_w_risp	70.19298246	81.24791667	85.95719178	88.93320413	90.92031579
w_avg_lnch_ang_wo_risp	-14.79487179	3.625	10.78391357	15.41666667	19.46766917
w_avg_lnch_ang_w_risp	-13.18	3.517018779	10.66779279	15.52205882	20.87788462
w_avg_est_ba_wo_risp	0.1821272727	0.2676545455	0.3082368421	0.339373913	0.3679941124
w_avg_est_ba_w_risp	0.16204	0.2563683099	0.3022310268	0.3388612099	0.3701766667
w_avg_est_woba_wo_risp	0.1772571429	0.2751928783	0.3360204082	0.3857759336	0.4341632653
w_avg_est_woba_w_risp	0.1534375	0.2585395833	0.3248315589	0.378743617	0.4277346535
w_avg_hit_distance_wo_risp	51.96	126.875	156.2341137	174.2752757	190.7403706
w_avg_hit_distance_w_risp	41.78	119.4866586	151.6806616	173.5164378	192.6824561

Variables Generated: Pitchers

- On-Base Percentage, Slugging Percentage, OPS Allowed as Pitcher
- Average Pitch Spin Rate
- Maximum Pitch Spin Rate
- Average Pitch Velocity
- Maximum Pitch Velocity
- Average Effective Pitch Velocity
- Maximum Effective Pitch Velocity
- Change in Average Pitch Spin Rate from Previous Two Seasons

Additional Variables Considered

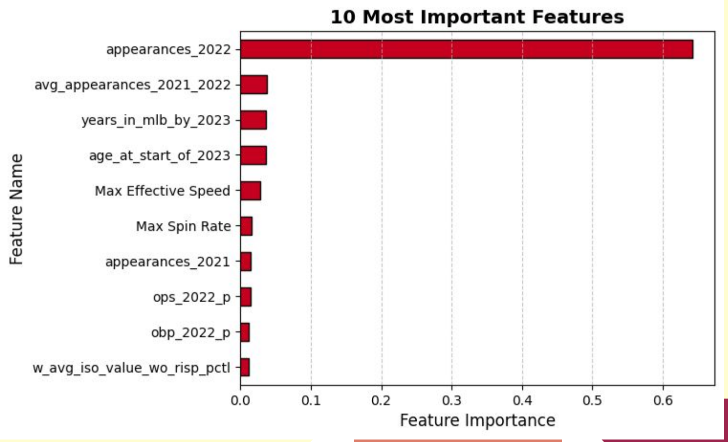
Some additional variables were thought up and considered but ultimately not generated:

- Starter rotation spot/batting order slot
- Leverage use for relievers/Team leverage instances from the last 2 years
- Ability to hit against the shift using:
    - Outfield alignment
    - Infield alignment
- Number of times a pitcher gets through the order
- Variance in pitcher arm angle between pitch types
- Catcher pop time
- Player sprint speed
- Player fielding percentage
- # of outs generated per batter faced
- wRC+

Modeling Attempts

Models were trained using 5-fold cross-validation to reduce overfitting. We used recursive feature elimination to select independent variables, choosing the model with the lowest average RMSE across validation iterations.

- Linear Regression (from scikit-learn)
- Lasso Regression (from scikit-learn)
- Random Forest 1 (from scikit-learn)
- Random Forest 2 (from tensorflow)
- Gradient Boosting (from scikit-learn)
- Histogram Gradient Boosting (from scikit-learn)
- Extreme Gradient Boosting (from xgboost)
- AdaBoosting (from scikit-learn)

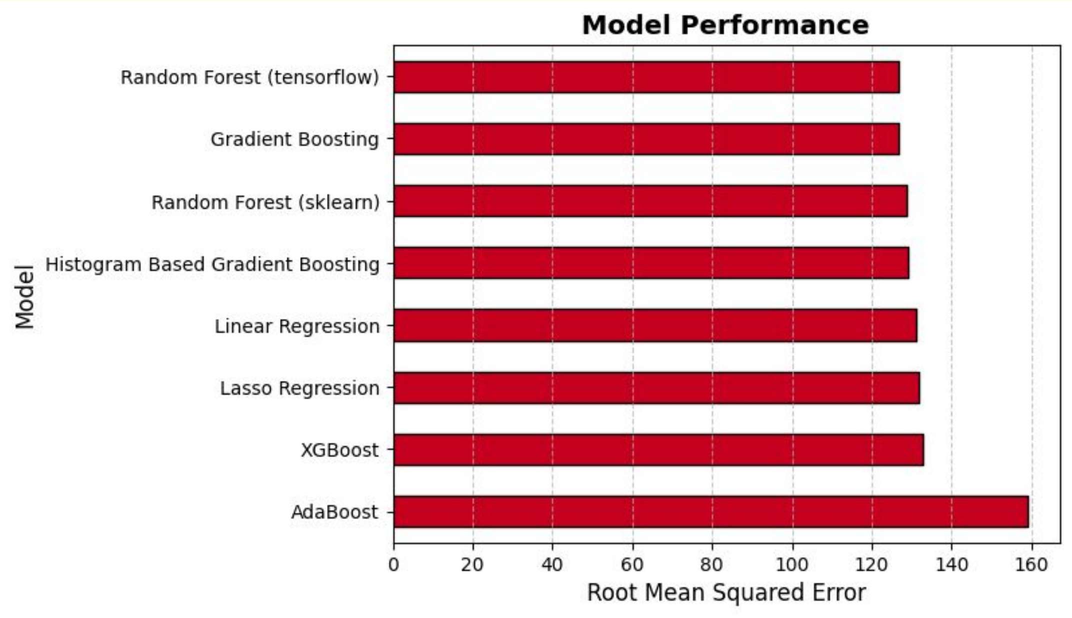


Modeling Results

Best model:

TensorFlow Random Forest

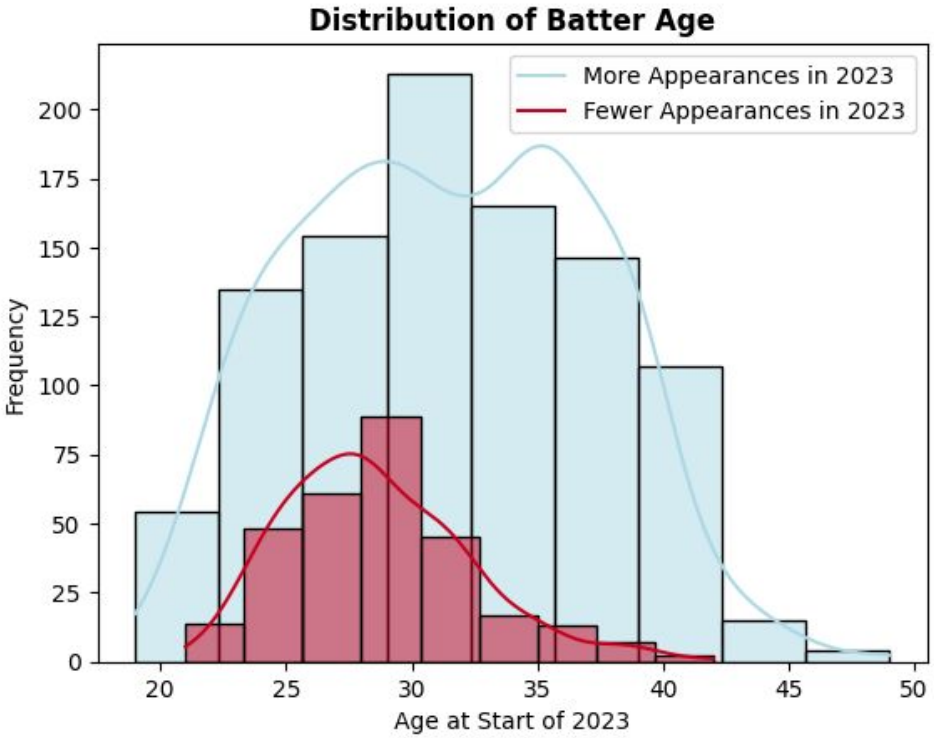
- Feature selection achieved with recursive feature elimination using gradient boosting
- Achieved 126.76 RMSE across five cross-validations

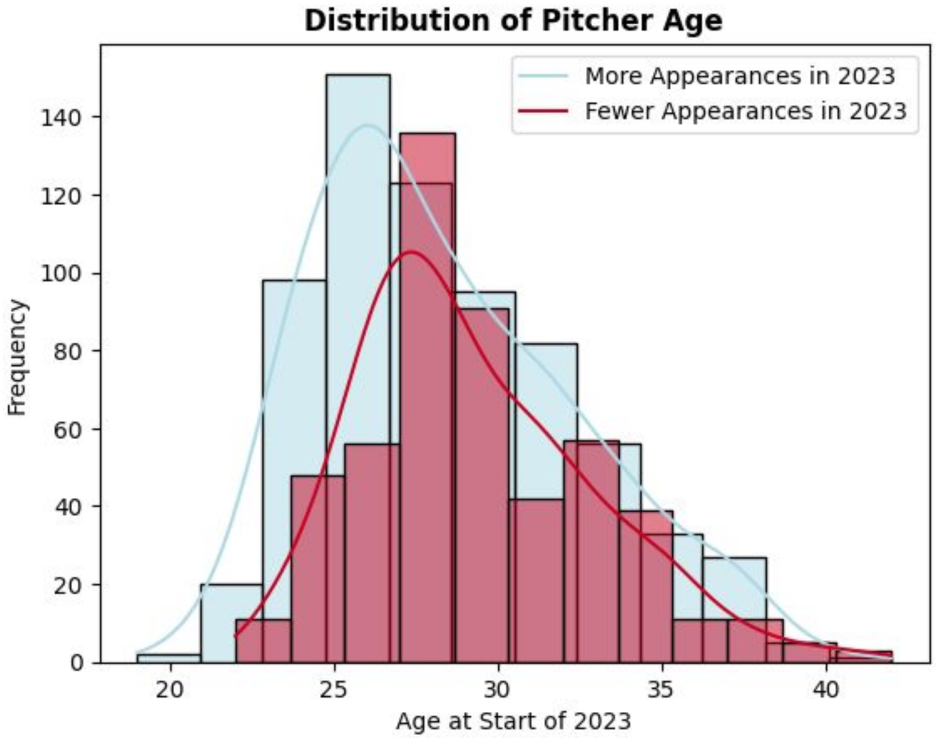


Optimization

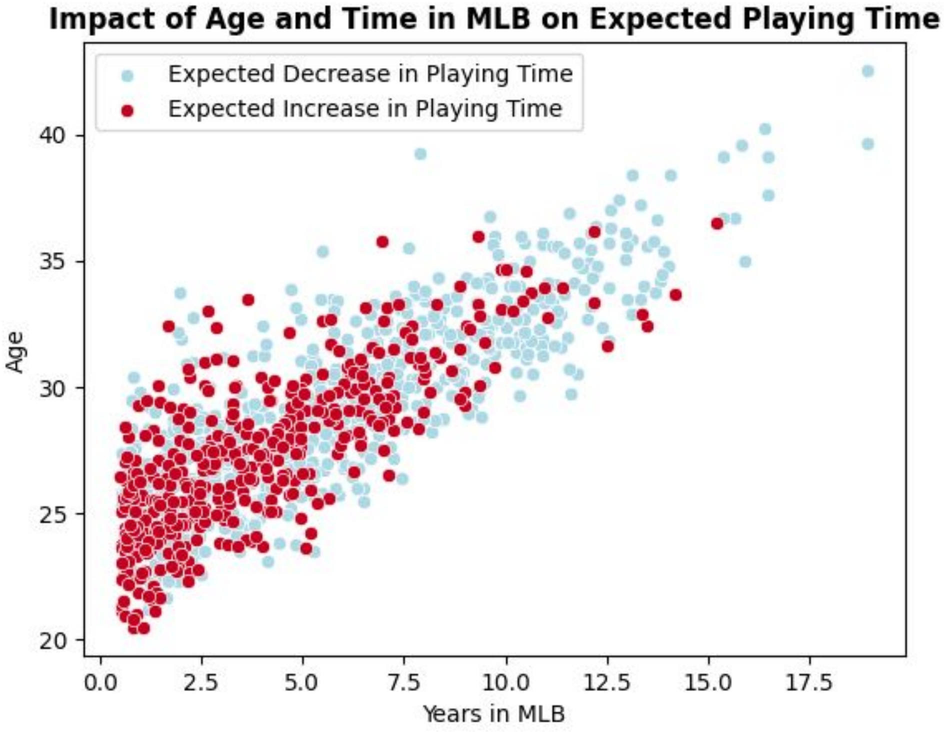
- Sorted through possible outliers and discrepancies within our data.
- Tweaked model parameters to further reduce error in predictions
- Scaled data for models that can be sensitive to feature scale
- Attempted Principal Component Analysis for feature reduction

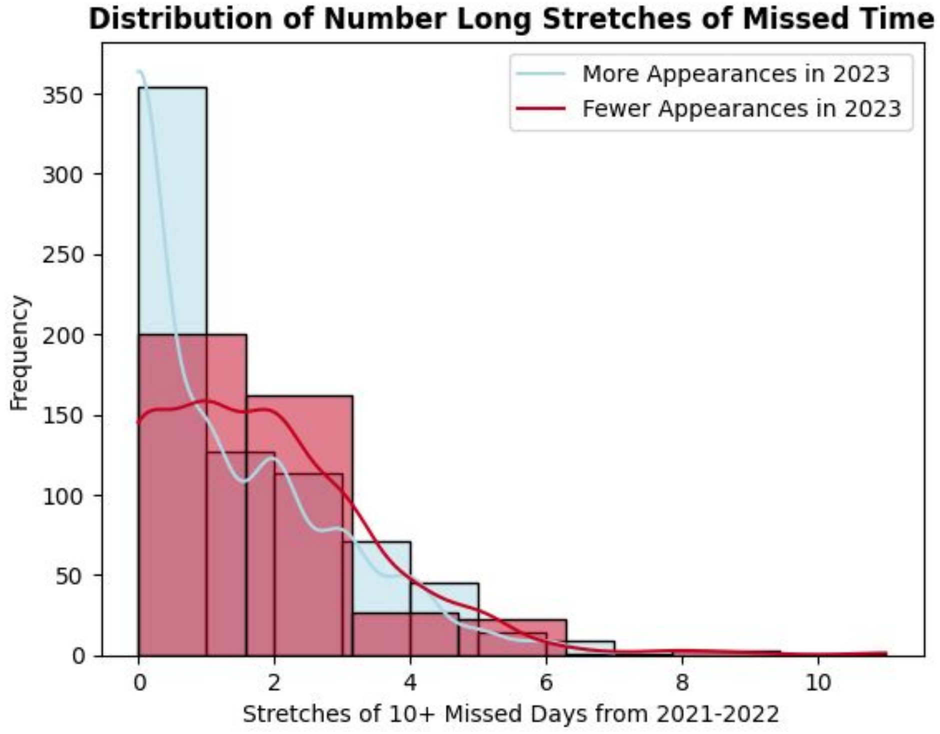
Playing Time Analysis





Playing Time Analysis (cont.)

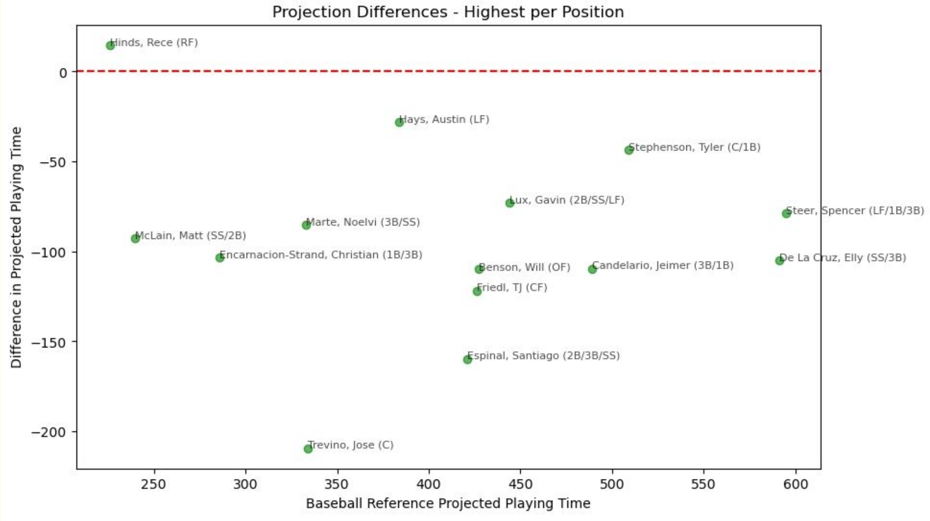




Application to 2025 Reds Playing Time - Batters

Top 10 Batters by Projected Playing Time

- Spencer Steer - 516
- Elly De La Cruz - 485
- Tyler Stephenson - 465
- Jeimer Candelario - 379
- Gavin Lux - 371
- Austin Hays - 355
- Will Benson - 317
- TJ Friedl - 304
- Jake Fraley - 285
- Santiago Espinal - 261

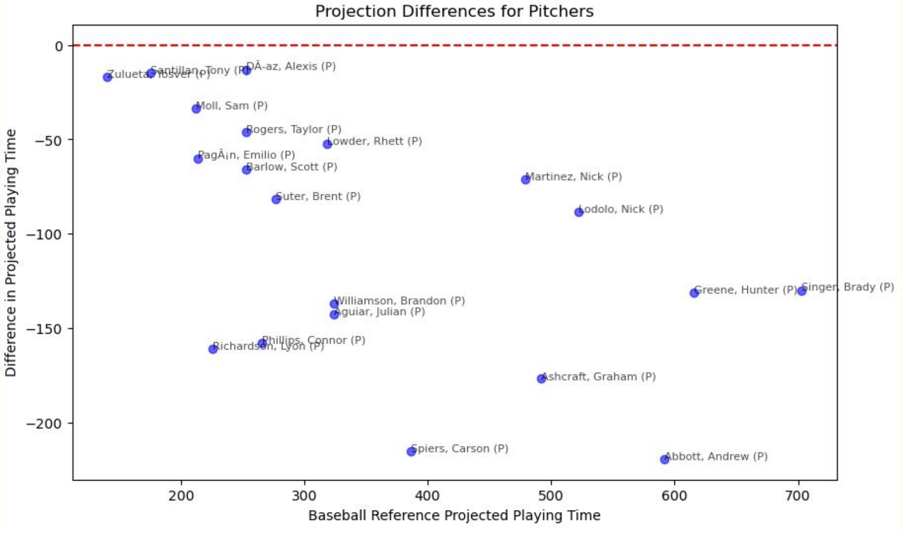



T. Friedl
A. Hays
W. Benson
E. De La Cruz
M. McLain*
G. Lux
S. Steer
T. Stephenson
J. Candelario

Application to 2025 Reds Playing Time - Pitchers

Top 10 Pitchers by Projected Playing Time

- ❑ Brady Singer - 572
- ❑ Hunter Greene - 484
- ❑ Nick Lodolo - 433
- ❑ Nick Martinez - 407
- ❑ Andrew Abbott - 372
- ❑ Graham Ashcraft - 315
- ❑ Rhett Lowder - 265
- ❑ Alexis Díaz - 253
- ❑ Taylor Rogers - 206
- ❑ Brent Suter - 195



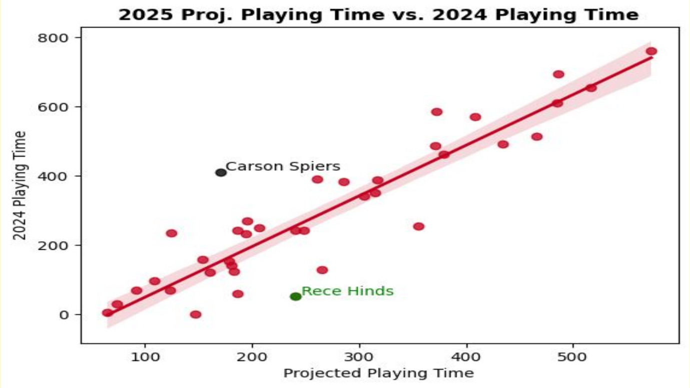
Limitations

The inclusion of a number of additional factors would likely significantly improve the prediction despite the purpose of the exercise in predicting playing time purely on performance. These factors include:

- Contract data
- Roster positional talent context
- Top prospect standing
- Team performance on the year
- IL transaction data

Application Possibilities

- Apply modeling to minor leagues to assess prospects, guiding decisions on playing time, call-ups, and demotions
- Identify undervalued bench players or prospects worth acquiring
- Determine optimal spring training invitation candidates/40 man roster members
- Analyze player performance in frequently played ballparks to maximize impact in key environments



Let's Hear From You