Cincinnati Reds Hackathon 2025

NC State IAA
Andrew Buelna, Landon Docherty, Brett Laderman, Daniel Ryan, Jacob Segmiller

February 16th, 2025

Table of Contents

Cincinnati Reds Hackathon 2025

Overview

This report outlines our approach to forecasting 2024 playing time using historical pitch-by-pitch data from 2021 to 2023. By developing a roster-agnostic model, we estimated each player's expected workload independent of team context. Our methodology combines statistical feature selection, domain expertise, weighted historical performance projections, and machine learning models to identify key drivers of playing time. Our models for hitters, starting pitchers, and relief pitchers achieved Root Mean Squared Errors (RMSE) of 16.11, 15.07, and 7.88, respectively, demonstrating strong predictive accuracy. Our results will provide the Cincinnati Reds front office with data-driven insights to enhance roster planning and decision-making.

Methodology & Analysis

This section outlines the data, variables, and model selection used to predict plate appearances and batters faced.

Data Used

Our team received Baseball Savant pitch-by-pitch data for regular-season MLB games from 2021 to 2023. Using this data, we aggregated player statistics by summing plate appearances (PA) for batters and batters faced (BF) for pitchers based on events that marked the end of a plate appearance. Additionally, we incorporated biographical information, such as age, from the Lahman People dataset to enhance our analysis. This approach enabled us to develop a roster-agnostic model to predict each player's total playing time in 2024, independent of roster competition, focusing solely on their expected individual contributions.

We split the dataset into three subsets: hitters, starting pitchers, and relief pitchers. We trained the hitter and pitcher models independently, as their statistics and target variables differed. We categorized pitching statistics into starting and relieving roles based on the designated indicator in the dataset. Each subset was further split into training and test datasets, with 2021 and 2022 data used to train the model, and 2023 being used as a test set. After tuning the model, we refit the model with all data to predict 2024.

Variable Selection

To predict plate appearances and batters, we predicted individual statistics. Using correlation matrices for hitters, starting pitchers, and relief pitchers, we identified statistics that were strongly correlated with plate appearances and batters faced while minimizing multicollinearity among the remaining variables in the model. We factored in some domain knowledge to select the statistics to keep when multicollinearity was an issue among equally strong predictors.

For hitter statistics, we selected the following ten variables, ordered by feature importance:

- Total Bases
- Sacrifice Flies
- On-Base Plus Slugging (OPS)
- Runs Created Per Plate Appearance
- Extra Base Hit %
- Weighted Spot in the Batting Order (Average)
- Walk %
- Strikeout %
- Contact %
- Age

For starting and relief pitcher statistics, we selected the following eight variables, ordered by feature importance:

- Hits Allowed
- Strikeouts
- Walks and Hits per Inning Pitched (WHIP)
- Walks
- Strikeout-to-Walk Ratio
- Hit by Pitch
- Earned Runs Allowed
- Home Runs Allowed

Model Selection

After selecting the variables for each model, we created two machine learning models for all subsets: a random forest model and an Extreme Gradient Boosting model (XGBoost). We chose machine learning models to capture the non-linearities within the data, such as the age variable. For instance, plate appearances typically increase during a player's 20s but tend to decrease as they enter their 30s. We evaluated model performance using the metric RMSE. Across all data subsets, we found that the XGBoost model outperformed the random forest model.

2024 Predictions

We predicted the 2024 statistics using data from the previous three seasons. To determine the optimal weighting system, we tested different year-by-year weights in 0.1 intervals by using 2021 and 2022 statistics to predict 2023 statistics. We selected the weights that minimized the RMSE of the predicted 2023 statistics compared to the actual 2023 statistics on a variable-by-variable basis. Finally, we adjusted the two-year weights into three-year weights to incorporate players with three years of data.

If a player had only 2023 data, their 2023 statistics were carried forward to 2024, as reflected in the one-year weights. Similarly, for players with only 2022 and 2023 data, the two-year weights were applied. Players who did not have a plate appearance or did not face a batter in 2023 were excluded from our predictions due to uncertainty about their status for the 2024 season.

Table 1 displays the one-year, two-year, and three-year weights for each batting statistic used in our prediction models. We ensured that each player's respective weights summed to one, representing a full season based on past performance.

Table 1: Weights Used for Predicting 2024 Batting Statistics

Variables	One-Year Weights	Two-Year Weights	Three-Year Weights
Total Bases, Expected Spot in Hitting Order	{2023}: 1.0	{2023}: 0.60 {2022}: 0.40	{2023}: 0.50 {2022}: 0.30 {2021}: 0.20
Strikeout %, Contact %, Runs Created per PA	{2023}: 1.0	{2023}: 0.70 {2022}: 0.30	{2023}: 0.60 {2022}: 0.25 {2021}: 0.15
Sacrifice Flies, OPS, Walk %, XBH %	{2023}: 1.0	{2023}: 0.80 {2022}: 0.20	{2023}: 0.70 {2022}: 0.20 {2021}: 0.10

As noted in Table 1, we found that using a weight equal to or greater than 0.50 for 2023 was the most effective across all of the statistics. The same methodology used to predict batting statistics was applied to predicting pitching statistics. We found that all pitching statistics in our models were most accurate when using the same weighting scale across all variables, as shown in Table 2.

Table 2: Weights Used for Predicting 2024 Batting Statistics

Variables	One-Year Weights	Two-Year Weights	Three-Year Weights
WHIP, Hits, Home Runs, Earned Runs, Strikeouts, Walks	{2023}: 1.0	{2023}: 0.60 {2022}: 0.40	{2023}: 0.50 {2022}: 0.30 {2021}: 0.20

The weights noted in Table 1 and Table 2 provided us with a 2024 projection for each player in the dataset, which we used to make predictions once we built and selected our final models.

Results

This section presents the results of our XGBoost models, including the variables that were most important for predicting playing time and the accuracy of the models.

Variable Importance

Fitting three different models to predict playing time for hitters, starting pitchers, and relief pitchers allowed us to assess which of the selected statistics were most important in projecting playing time for each group. Figure 1 displays the importance of the different variables used as inputs in the XGBoost model for hitters.

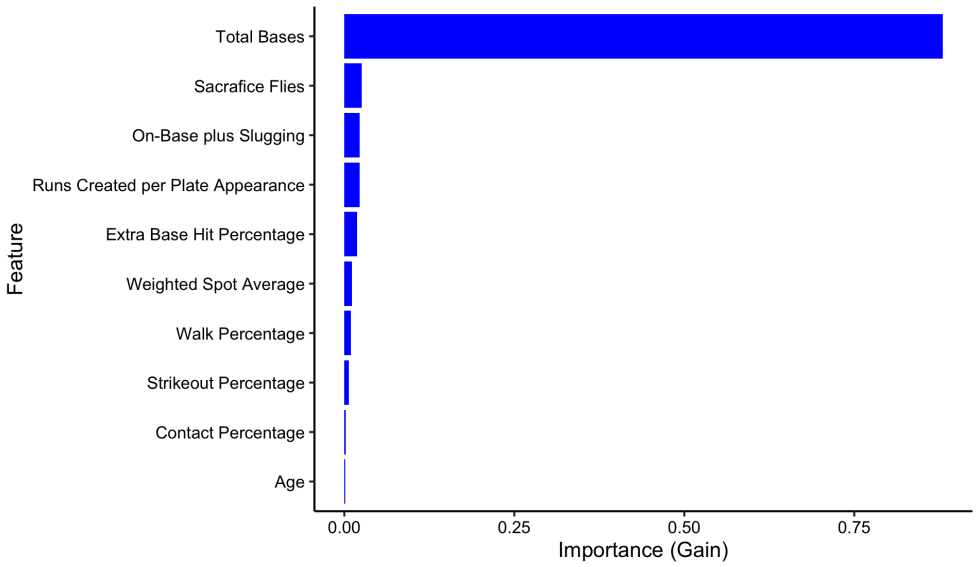

Figure 1: Variable Importance for Hitters Model

Figure 1 shows that total bases is the most important variable for predicting plate appearances, with other factors like OPS also playing a significant role. Since accumulating total bases requires frequent plate appearances, its importance is expected. However, the inclusion of non-counting statistics highlights the XGBoost model's ability to account for both a player's workload and overall performance.

For the pitcher models, the same variables were used to predict batters faced, but since different models were fit, the variable importance results differ slightly. Figure 2 displays the variable importance for the starting pitchers model.

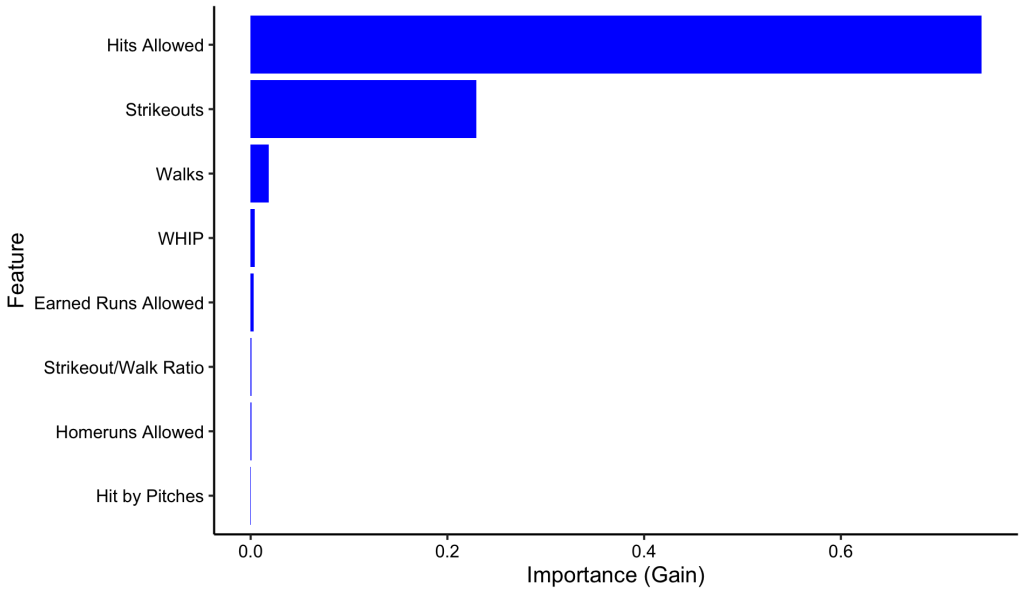

Figure 2: Variable Importance for Starting Pitchers Model

Figure 2 shows that similar to hitters, the most important variable for starting pitchers is the counting statistic hits allowed. However, unlike the hitters model, non-counting statistics play a smaller role, though they still contribute to the overall predictions.

Figure 3 shows the variable importance for the model predicting relief pitcher playing time.

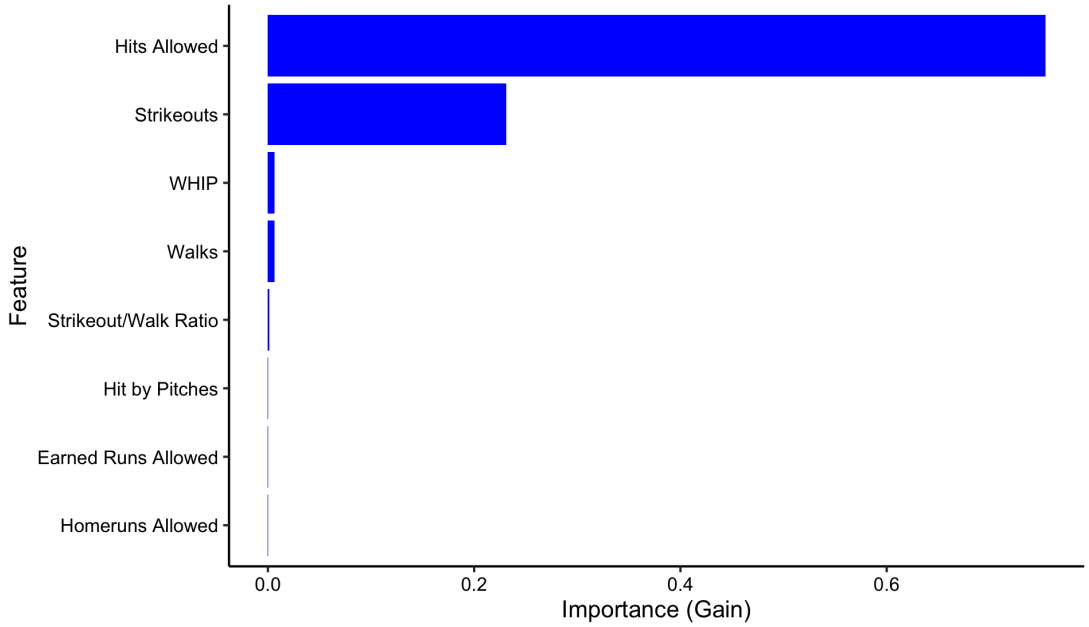

Figure 3: Variable Importance for Relief Pitchers Model

As shown in Figure 3, the most influential variable for relief pitchers was hits allowed, while the most important non-counting statistic was WHIP.

Accuracy Measures

For each of the models, 2021 and 2022 stats were used to train the model with 2023 acting as a testing set. Table 3 displays the RMSE for the three models to predict playing time.

Table 3: Accuracy of All Models for 2023

Model (Response Variable)	RMSE
Hitters (Plate Appearances)	16.11
Starting Pitchers (Batters Faced)	15.07
Relief Pitchers (Batters Faced)	7.88

According to Table 3, all three models were fairly accurate in their predictions for 2023 playing time. For hitters, the model was off by about 16.11 plate appearances, on average. For pitchers, projections differed by about 15.07 batters faced for starters and 7.88 for relievers, on average.

Recommendations

Using XGBoost models, we projected plate appearances and batters faced for players in the 2024 season. Projections were adjusted to account for players expected to make both relief and starting appearances, as well as those anticipated to have plate appearances. While the models demonstrated strong accuracy for 2023, we recognize several limitations when applying them to future projections. Since we cannot determine whether fluctuations in playing time result from injury, poor performance, or a late-season call-up, some projections may be inaccurate. Additionally, the absence of roster context may lead to misestimations for players whose roles significantly changed in 2024. Furthermore, new players entering the league in 2024 will have no predicted values, as they were not included in the dataset. While this is not an exhaustive list of limitations, these challenges highlight that this predictive approach is best suited for scenarios where only historical and projected statistics are available.

Conclusion

Our predictive models effectively estimated 2023 playing time, capturing key variables influencing player usage. The models for hitters, starting pitchers, and relief pitchers achieved RMSE of 16.11, 15.07, and 7.88, respectively, highlighting their strong predictive accuracy. While their performance for 2024 cannot yet be evaluated, these models offer valuable insights for the Cincinnati Reds' roster planning. However, factors such as injuries, roster changes, and the absence of data for new players may impact projections. Despite these limitations, our XGBoost approach provides a robust foundation for data-driven decision-making.

Appendix

All code is available on Github.

Thank you to Dr. Sarah Egan Warren, Dr. Aric LaBarr, and Dr. Susan Simmons for their advice throughout the process.