MLB xBA Prediction: A Study Using Statcast Pitch Data

Austin Kingsley

Regis University

MSDS 696: Data Science Practicum II

Dr. John Koenig

August 21, 2022

Research Question

In an effort to predict expected batting average (xBA) for Major League Baseball (MLB) players, I intend to use the recent performance of batters and pitchers through previous xBA, strikeouts, and walks. Additionally, I will include matchup focused interpretations of Statcast pitch metric data in an attempt to better represent the players involved in each matchup.. Statcast pitch data was implemented in all ballparks with a standardized measurement system in 2017.

Data

All data is retrieved from the Baseball Savant API, which includes game files for each game. This includes entries for every single pitch thrown. The game files used in this assignment include all official MLB games between April 2, 2017 and July 27, 2022. The test data for this assignment is focused on the games from the 2022 season.

Data Cleaning

The scope was narrowed to Regular Season games in order to handle most of the missing data and to promote a more focused project. Exhibition games and Spring Training games did not have recorded Statcast pitch metrics, and most xBA data was not recorded.

Missing xBA was filled using the median of the listed outcome of the at bat and the year. Some categories, such as “sac_bunt” plays, did not have enough yearly data, resulting in the removal of the yearly grouping. Strikeouts were given an xBA value of zero.

Missing values for Statcast pitch metrics largely overlapped. Therefore, backward fill and forward fill methods were employed to fill the comparatively few missing Statcast values.

Methodology

This project can be broken down into five steps. The first step involves clustering pitches to represent the different look of pitches thrown by pitchers. This will be done using Statcast pitch metric data. The second and third step involve separate predictive modeling of batter and pitcher xBA daily outcomes based on recent performance. The fourth step involves matchup based performance modeling using the modeled clusters. These clusters will represent pitcher repertoire, and will be used to determine batter performance against pitch types they have seen. The fifth step involves combining all previous models to attempt to achieve optimal results.

Clustering

Fuzzy C-Means Clustering was employed to both hard and soft cluster pitches. The current version uses eleven clusters, which was found from the best silhouette score based on k-means clustering. This resulted in a grouping of pitches that is not fully representative. It underrepresents left handed pitchers, which is traditionally seen as an important trait in matchup outcome, and it does not distinguish between arm slots (Fig. 1). The effective speed of the pitch, which is based on the release speed and the release point distance from home plate, is clearly defined by the clusters.

Feature Engineering

All pitch metrics were replaced with a hard and soft clustered version. The last pitch of each at bat was tracked for weighted batter performance measurements against clustered pitches, a rolling average was taken of all pitches to represent pitcher repertoire, and the maximum and minimum values of the soft clusters were recorded to maintain pitch identity.

For the batter and pitcher clustering, a rolling mean of xBA, strikeouts, and walks were taken over the course of twenty and twenty-one days at increments of five and seven days respectively. For batters, the length of these increments was found to be minimally impactful, but seven days was decidedly best for pitchers. This is because it captures each start for a starting pitcher. The depth of the search was selected for the purpose of balancing applicability of the model and achieving results.

Results

There are five total models to discuss the results of. The first one models recent batter performance. The second one models recent pitcher performance. The third model is the matchup model based on clustered pitches. The fourth and fifth model combined the previous results together, one of which took in all previous predictions and one which inserted pitcher and batter predictions into the matchup model. Each model is evaluated by mean absolute error (MAE) and weighted mean absolute error (WMAE). The weight factor for WMAE is the number of plate appearances in each recorded daily matchup.

Batter Performance Modeling

The daily performance of batters experienced a value of 0.203 for MAE and a value of 0.183 for WMAE on the test matchup set. For daily batter performance, this model resulted in a MAE value of 0.127. This shows great promise for a simple random forest model. Improvements can be made by tuning the bins which plot player performance over time and model selection.

Pitcher Performance Modeling

The daily performance of pitchers experienced a value of 0.195 for MAE and a value of 0.176 for WMAE. For daily batter performance, this model resulted in a MAE value of 0.089. One reason for the better performance is that pitchers tend to have as many events or more events than batters in a given game, allowing for less error to exist in the data. However, this performance difference translated to matchup specific instances as well, providing refutation for this claim. Similar improvements can be made for this model as the batter performance model.

Matchup Modeling

The matchup specific model experienced a value of 0.201 for MAE and a value of 0.183 for WMAE. This result was not as good as I desired and the combination models suffered from the performance of this model; however, there is also the most room for improvement from this model. Improving cluster definition, introducing weights to batter performance against pitch clusters, and exploring other models are all areas of improvement.

Combination Models

The model which took in all previously discussed predictions will be called the stacked model. The model which took in the batter and pitcher models and inserted them into the matchup model will be called the combined model.

The stacked model experienced a value of 0.191 for MAE and a value of 0.174 for WMAE. The combined model experienced a value of 0.195 for MAE and a value of 0.175 for WMAE. These models performed disappointingly, but by improving the other models, these scores should increase.

Conclusion

The combination models underperformed, but there are clear areas for improvement. These areas include cluster definition, model selection, hyperparameter tuning, and performance definitions. However, the models saw increased performance with increased plate appearances in a matchup. Scores dropped nearly 0.100 points for all models at higher plate appearance matchups. This makes sense due to less error being present in their results. For three and four plate appearance matchups (~18.4% and ~0.8% of matchups), the combined model scored best at 0.134 and 0.113 MAE. At the same time, the stacked model scored best at one and two plate appearance matchups (~64.3% and 16.1% of matchups) with MAE scores of 0.211 and 0.171. This could lead to distinguishing models between reliever and starting pitchers, and provides interesting points for more analysis. Ultimately, further improvements on the model should see a smoother curve, but fewer plate appearances in a matchup should result in more error in the data. Therefore, the combined model is currently the best model. All MAE values related to this are recorded in Table 1.

Appearance per Matchup	Batter Model	Pitcher Model	Matchup Model	Stacked Model	Combined Model
1	0.2404647	0.2247453	0.2362080	0.2112090	0.2226849
2	0.1777279	0.1765643	0.1780783	0.1713727	0.1715528
3	0.1401068	0.1392430	0.1408975	0.1366073	0.1346946
4	0.1167547	0.1171960	0.1180898	0.1171464	0.1133636

Table 1. Displays the Mean Absolute Error Value across the number of plate appearances in a pitcher-batter matchup for each test model.

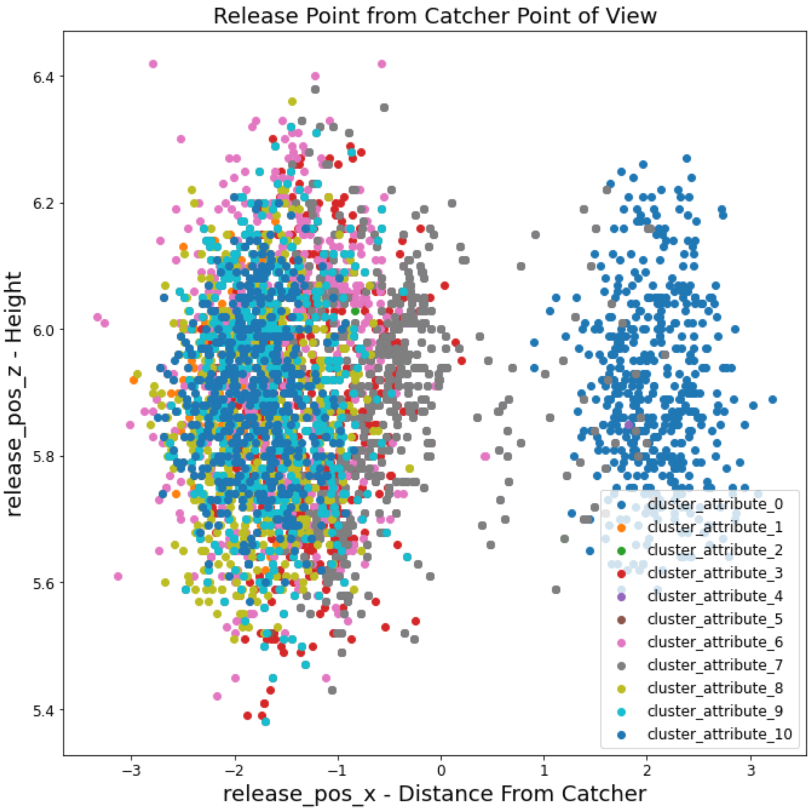

Fig. 1.1 - Involves the top 500 pitches in each cluster

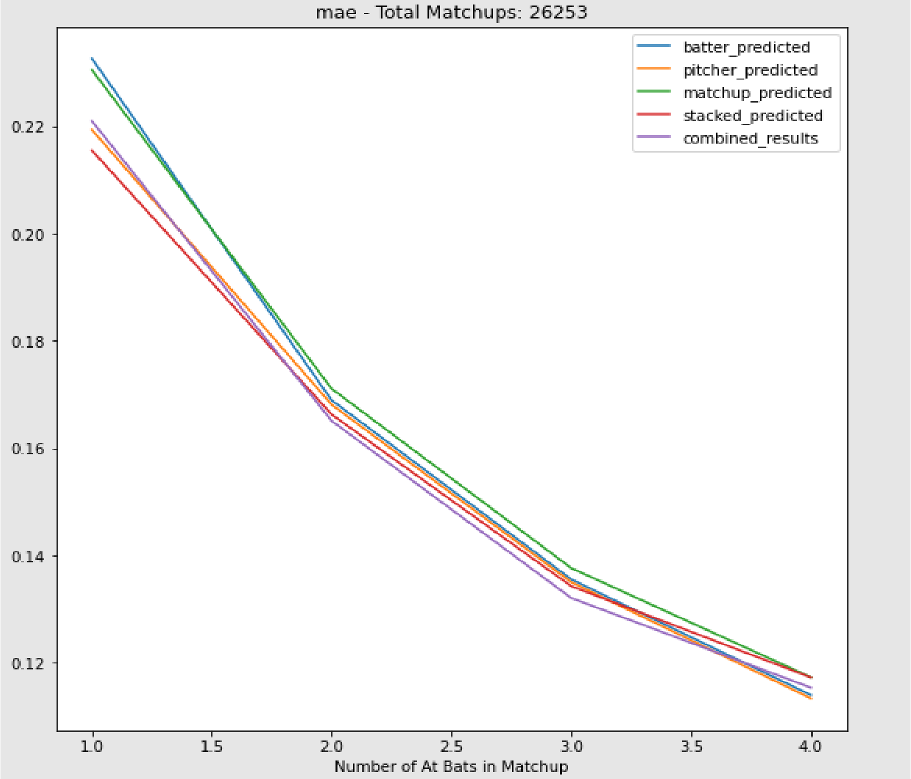

Fig 2 - Graph of MAE over the test data. Shows increased accuracy with more plate appearances per matchup in question.