Do Better Batters Face Higher or Lower Quality Pitches?

Member: Jean An, Elliot Lee, Ziang Zhou
Mentor: Alex Caravan, Driveline Baseball

Abstract

Even before the concept of "Moneyball" became prevalent, statistical analysis has always been at the core of baseball. With that head-start, baseball has maintained its position as the leader of all sports in the development and the usage of analytics. Currently, in Major League Baseball, "using analytics" is no longer a competitive advantage, but a bare minimum requirement for teams to stay competitive; the true edge lies within the depth of the analysis and the strength of the data.

The goal of this project is to develop models that **quantify the quality of individual MLB pitches** (commonly known as "**Stuff+**"), and to examine if the implemented results can be used to successfully answer the question **whether there is an observable relationship between how good a batter is and the difficulty of the pitches they face**. Our hypothesis is that better batters perhaps are not facing better pitches, but are simply facing better pitchers.

There are currently multiple versions of Stuff+ models developed by different sources—including one from Driveline Baseball—that evaluate pitch quality against intrinsic run values developed through the calculation of simple linear weights. We present a state-of-the-art model to evaluate the quality of pitches against change in **xwOBA** instead. Rather than a measurement of batted ball results, xwOBA—developed by Statcast—measures the expected value of batted balls based on exit velocity, launch angle, and sprint speed. The contributions of this model are focused on evaluation based on quantifiable attributes of the pitch delivered, rather than the outcome of the event or the game state.

Introduction

Pitching Data: The dataset consists of roughly 972K pitches from the 2020 & 2021 seasons, with 80% going into the training set and 20% going into the testing set. The trained models are then applied to roughly 707K pitches from the 2022 MLB season.

Data Source: Models were trained and tested on publicly available pitch-by-pitch data from the 2020 and 2021 MLB seasons that were downloaded from Baseball Savant, the official website of Statcast.

Pitching Model:

- **Model:** XGBoost
- **Features:** Pitch Velocity, Vertical Break, Horizontal Break, Extension.
- **Pitch Types:** FF, SI, FC, SL, CU, CH.

Target Variable: Delta xwOBA = End xwOBA - Start xwOBA

- **Start xwOBA:** Average xwOBAcon of initial ball-strike count
- **End xwOBA:**
  - **Batted balls:** Actual xwOBAcon
  - **Strikeout, Walks, HBP:** wOBA
  - **Non-events:** Average xwOBAcon of new ball-strike count

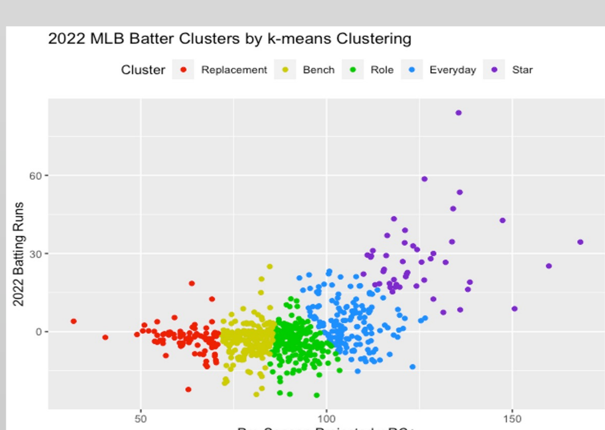
Count	xwOBAcon
0-2	.328
1-2	.344
2-2	.357
0-1	.357
1-1	.376
3-2	.383
0-0	.384
2-1	.389
1-0	.394
2-0	.421
3-1	.423
3-0	.463

Batting Data: The dataset consists of seasonal data of the 687 non-pitcher batters who had at least one plate appearance in the 2022 MLB season, combined with the 2022 Pre-Season Projections from Steamer, a renowned projection system that develops forecasts for MLB players.

Data Source: Batting statistics were downloaded from FanGraphs. Steamer pre-season projections were provided by Driveline Baseball.

Batting Model:

- **Model:** k-Means Clustering
- **Features:** Batting Runs, Projected wRC+
- **Clusters:** 5
  - Star players (n = 47)
  - Everyday players (n = 153)
  - Role players (n = 211)
  - Bench players (n = 197)
  - Replacement-level players (n = 79)



Methodology

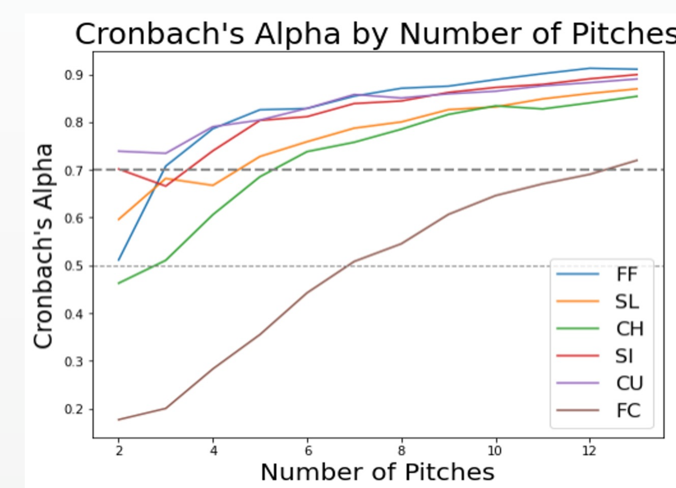
As introduced in the previous section, our Stuff+ model **relies on XGBoost to predict our own target metric "Delta xwOBA"** using several input features. The Pitch Type classifications are provided within the data set from Baseball Savant, but we grouped Knuckle Curves (KC) together with Curveballs (CU) and Splitters (FS) together with Changeups (CH), resulting in six unique pitch type groups.

Each pitch type is trained and implemented separately. The input features of the models are **Pitch Velocity (MPH), Horizontal Break (inches)**, **Vertical Break (inches)**, and **Extension (feet)**, which is the distance between the front of the pitching rubber and where the pitch is actually released from the pitcher's hand. The following are the training and testing errors of each pitch type group:

Pitch Type	Training RMSE	Training COD (R^2)	Testing RMSE	Testing COD (R^2)
FF	0.215	0.006	0.215	0.001
SI	0.204	0.012	0.206	0.0004
FC	0.209	0.020	0.205	0.001
SL	0.206	0.009	0.211	0.001
CU	0.192	0.018	0.194	0.002
CH	0.207	0.014	0.206	0.0004

RMSE is the metric used to measure the **descriptiveness** of the predictions, and **R^2** is used to determine the **predictiveness** of the predictions. Since our target variable Delta xwOBA is very noisy by nature (majority of the pitches are not PA-ending events), the low R^2 observed is actually reasonable. After all, the metric is neither linear nor normally distributed.

In addition to being descriptive and predictive, it is also important to understand the **reliability** of the produced Stuff+ metric. For this, we used **Cronbach's Alpha**, which is a coefficient that measures the "internal consistency" of a metric. The results show that five of our six models were able to produce predictions that are consistent immediately. However, the model for cutters lags behind:



The fact that our cutter Stuff+ takes more than double the number of pitches to stabilize compared to the other pitch types is interesting, given that the errors were pretty inline with the others. It is worth noting that the cutter model also produced the highest training R^2. It is possible that it may be more difficult for pitchers to replicate their cutters in terms of pitch characteristics, although that is unlikely given that other Stuff+ models have not generated similar results.

Lastly, our batter classification is produced by a simple two-dimensional **k-Means clustering model with five clusters**. The input variables are the batters' actual **Batting Runs** and **Projected wRC+**. The produced clusters are labeled in the order of their skill level as: **star** players, **everyday** players, **role** players, **bench** players, and **replacement-level** players.

Results

To make our produced metric easily interpretable, we scaled and centered them using Z-scores. In the remainder of this piece, all Stuff+ numbers, unless otherwise stated, will be represented in the same scale: **100 is average**, 200 is one standard deviation better than the average, -100 is two standard deviations worse than the average…etc.

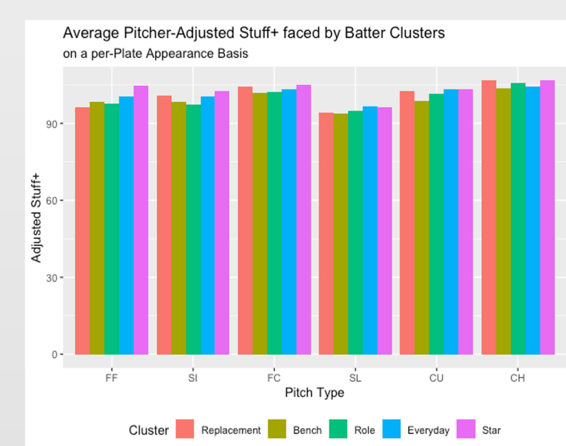
With Stuff+ applied to every pitch from the 2022 MLB season and all the batters classified into five separate categories, we can now examine if there is any observable relationship between the quality of the pitch and the general ability of the batter. If we simply look at the average Stuff+ across all pitch types that batters in each category faced, there is a clear relationship: the better the batter, the tougher the pitches he sees:

Cluster	Stuff+
Star	103.0
Everyday	101.0
Role	98.7
Bench	98.1
Replacement	97.1

However, we suspect that there may be sampling bias in this case. It is possible that better players see better pitches on average because they face better pitchers more often. If opposing teams wants to have success against these elite hitters, they should use their best pitchers to face them.

To adjust for pitcher strength, we simply **centered the Stuff+ for each pitch by the average Stuff+ of that pitcher** for that given pitch type. For instance, if a pitcher threw a slider that graded as 120 Stuff+, but his personal average slider Stuff+ is 140, that pitch now has an adjusted Stuff+ of 80 instead of 120, because the figure is 20 below his personal average. A 100 adjusted Stuff+ in this case means the pitcher threw a pitch that is in line with his personal average, and not necessarily in line with the average of the entire population. Interestingly, even after adjusting for pitcher strength, we see no difference in Stuff+ at the top end. **Star players still face the toughest pitches.**

Cluster	Adjusted Stuff+
Star	103.0
Everyday	101.0
Role	98.8
Bench	98.4
Replacement	98.9



The results seem to suggest that while better batters indeed face better pitches, it is not because of what we hypothesized, where it is mostly due to the strength of the pitchers they face, at least not on a per-pitch basis. However, in theory, better pitchers could be throwing fewer pitches because they are relatively better at putting away the batters like they are supposed to, without having to waste many pitches. Throwing more pitches does not mean a pitcher is deployed more often.

To test for this, we simply looked at each pitcher's personal Stuff+ average and grouped them together **on a per-PA basis** rather than **on a per-pitch basis** and did not distinguish by pitch type. Essentially, this will tell us the average strength of pitchers used against batters from the different clusters. A relationship, albeit small, is observed in this case: **the better the batter, the better the pitcher they face**. However, the differences among clusters is smaller, particularly at the top end, where there is barely any discrepancy observed between the star players and the everyday players.

In summary, while there is an observed difference, it is clearly not large enough to offset the previously observed relationship between batter cluster and the pitcher-adjusted Stuff+. It is likely that both cases are true: **star players face better pitchers, but also face better pitches** from all the pitchers.

Cluster	Pitcher Stuff+
Star	100.0
Everyday	99.6
Role	99.1
Bench	98.6
Replacement	97.3

Conclusions and Future Work

This project has successfully demonstrated that batter MLB batters—specifically the very elite star hitters—**do face better quality pitches**. At the same time, they also **face better pitchers in general**. Through the pitch quality model that we built and implemented, we found that MLB pitchers bring out their best "stuff" and elevate their game when they are tasked to battle against the best of the best in the batter's box. However, since there were no clear differences observed within specific pitch types, it is likely that **pitcher choose to use their best pitches more**, rather than truly throw their pitcher better when facing the best hitters.

It is worth noting that the concept of this pitch quality model is similar to most others, where we aim to quantify the raw "stuff" of any given pitch. Additional factors such as pitch location and pitch sequencing are not taken into account, even though these are clearly variables that have impact on the outcome of pitches. Future studies could aim to incorporate pitch location into consideration to find out if pitchers pitch more carefully when they face the best hitters.

Some additional adjustments to this study can also be made. Platoon splits and times-through-order penalties are some examples of factors that this study did not include given various constraints. It would be interesting to see how pitchers perform facing batters of different handedness, but the effect of the potential change in pitch type usage would have to first be adjusted for.

Glossary

- **wOBA:** Weighted On-Base Average, a metric developed by Tom Tango that uses Linear Weights to assign values to different outcome events in baseball and quantify the production a player produces with his bat. Adjusted to match the OBP scale; not adjusted for ballpark or league.
- **xwOBA(con)**: Expected Weighted On-Base Average (on Contact), a metric developed by Statcast that uses Exit Velocity, Launch Angle, and in certain cases Sprint Speed to predict the wOBA of a given batted ball. Uses a k-NN model plus a GAM model to incorporate Sprint Speed.
- **Pitch Types:**
  - **FF:** Four-Seam Fastball
  - **SL:** Slider
  - **SI:** Sinker (aka. Two-Seam Fastball)
  - **CU:** Curveball
  - **FC:** Cutter (aka. Cut Fastball)
  - **CH:** Changeup
- **wRC+:** Weighted Runs Created Plus, a metric developed by FanGraphs that aims to quantify the production a player produces with his bat. Adjusted for ballpark and league average, which is set at 100. A 110 wRC+ means a player's production is 10% better than the average; a 80 wRC+ means a player's production is 20% worse than the average…etc.
- **HBP:** Hit By Pitch, when a batter is struck by the pitch thrown by the pitcher, he is awarded first base.
- **PA:** Plate Appearance, representing each turn a batter comes to bat at the plate.

Acknowledgements and References

We would like to thank our mentor, Alex Caravan of Driveline Baseball, for his guidance and support throughout this project. We would also like to thank Mark Ho, New York University, for his feedback and constructive criticisms.

Asel, John. "Rethinking the True Run Value of a Pitch With a Pitch Model." Driveline Baseball, 24 Sept. 2021, www.drivelinebaseball.com/2021/09/rethinking-the-true-run-value-of-a-pitch-with-a-pitch-model

Langin, Chris. "Pitch Design: What Is Stuff+? Quantifying Pitches With Pitch Models." Driveline Baseball, 9 May 2022, www.drivelinebaseball.com/2021/12/what-is-stuff-quantifying-pitches-with-pitch-models.

Sharpe, Sam. "An Introduction to Expected Weighted On-Base Average (xwOBA)." MLB Technology Blog, 20 Sept. 2019, https://technology.mlblogs.com/an-introduction-to-expected-weighted-on-base-average-xwoba-29d6070ba52b