Optimal Batting Approach and Season Predictions in the MLB

IST 718

Nick Lukowsky

April 22nd, 2024

Project Overview

Batters' skills in the MLB (Major League Baseball) naturally as they age over the course of their careers. With new advanced metrics being introduced every year, batters can adjust their approaches at the plate and their training to increase the value of said metrics, and therefore increase their production on the field. The goal is to provide insights into what these batters should do to stay productive year to year by focusing on these metrics. Not only will these findings be useful for the players themselves, but for coaches and managers to advise players or build rosters.

Prediction, Inference, Goals

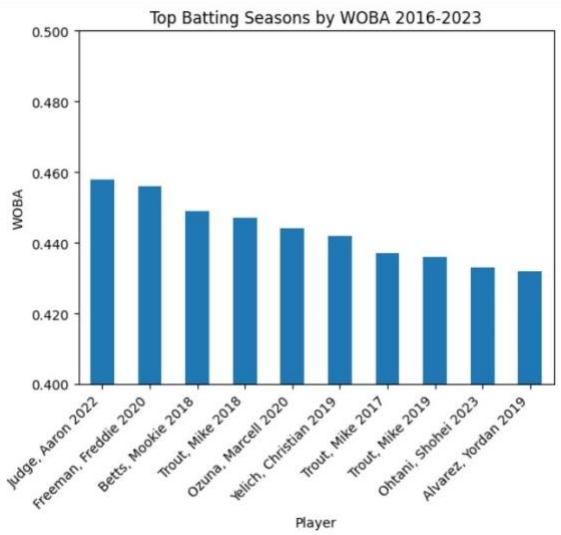
There are two main sections to this project: finding an optimal batting approach based on advanced hitting metrics, and using age as a predictor for how batters' statistics will change as they get older. For an optimal batting approach, it was hypothesized that a high exit velocity, high pull percentage, and good discipline (low swing rate for pitches outside of the zone, high swing rate for pitches inside the zone) would lead to a hitter performing very well by both the eye test and standard statistics. As for the batter projections, I generally expect a peak age to be between 27 and 30, with an incline before and a decline after. By looking at age by age differences a player's batting statistics can be predicted based on the previous three years.

Data Exploration

About the Data

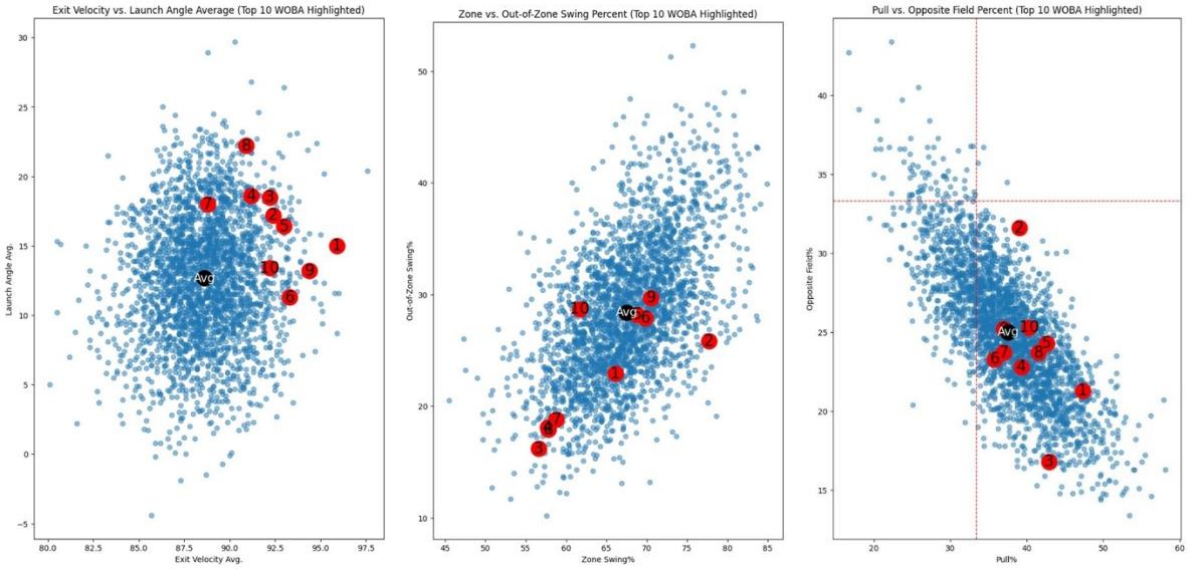
The data comes from the website Baseball Savant, which compiles traditional statistics (HR, AVG, OPS, etc.) with newer and more advanced swing tracking statistics (Exit Velocity, Launch Angle, Zone Swing%, etc.). This site is great for gathering data through the use of the Statcast Search tool. Metrics can be selected as columns with data going back to 2008 for certain stats and many filters can be used to create the subset of data that is desired. The dataset for this project contains 4264 rows and 28 columns, with each row representing a player season and each column representing different metrics used for this analysis. A minimum of 50 plate appearances per season was used, therefore the dataset included all batting seasons meeting that parameter from 2016-2023.

Visualizations



Above shows the first visualization made, which is simply the top 10 WOBA (weighted on base average) seasons in the dataset. WOBA is a more recent statistic that works as a complement to on-base percentage, as instead of simply considering whether a player reached base WOBA accounts for how that player reached base. Each way of reaching base is given a value that is the projected runs scored that each event adds.

According to the MLB Advanced Stat Glossary, “In 2014, a home run was worth 2.101 times on base, while a walk was worth 0.69 times on base. So a player who went 1-for-4 with a home run and a walk would have a wOBA of .558 -- (2.101 + 0.69 / 5 PAs)”. WOBA has become one of the best metrics for measuring hitters and their productiveness. In this top ten, it’s seen what many consider some of the best hitting seasons of the Statcast era regardless of WOBA, another good indicator of it being a strong offensive metric.



The other set of exploratory research that was done with the dataset looked at comparisons between certain advanced metrics with these top WOBA seasons highlighted. This allowed for a sense of where the best hitters stand relative to the rest of the league. The first graph on the left shows Exit velocity on the x-axis and launch angle on the y-axis. The best hitters are bunched in the top right, which means they are hitting the ball harder and higher in general than the league average. The second graph shows the swing percentage of balls inside the strike zone on the x-axis and swing percentage on balls outside the strike zone on the y-axis. Compared to the average, most of the top 10 batters do not swing at too many pitches outside the zone but their swing percentage in the zone varies. Finally, in the last graph on the right a player's pull percentage, or the percentage of time the batter hits the ball to the third of field of which they are batting from (LF side for right-handed batters and vice versa), is on the x-axis and opposite field percentage is on the y-axis. It's seen that most players pull the ball more than not, which makes sense because it is easier to generate power that way, and the top 10 hitters tend to pull the ball even a bit more than that average.

Methodology

Batting Approach

For the batting approach part of the project, a polynomial regression was used to fit each metric and WOBA. Each of the following graphs in the Results section shows where the metric is optimized for the highest WOBA. Different degrees were tested based on k-fold cross validation to find the degree with the smallest RMSE to then locate the optimal point.

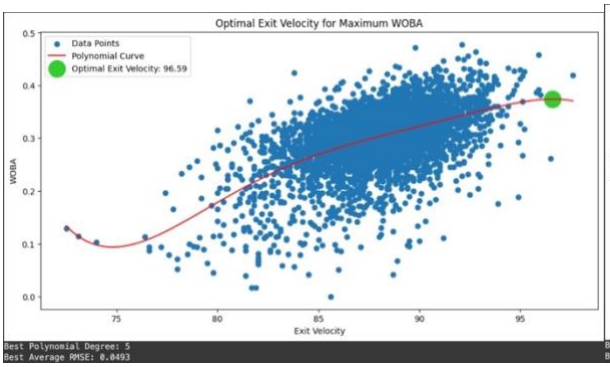
Batting Projections

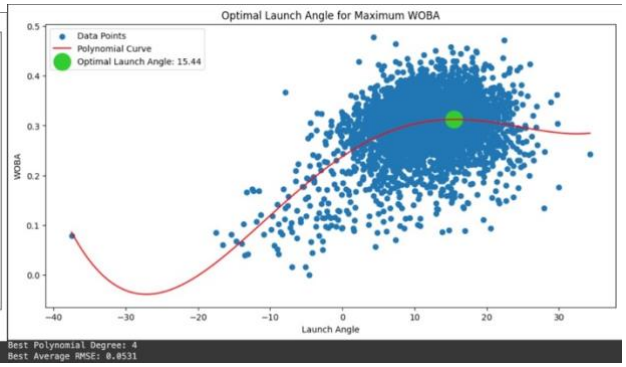
Tom Tango created the popular MLB projection system called Marcel, in which I am replicating in at a base level for my projections. His projection, "Uses three years of MLB data, with the most recent data weighted heavier. It regresses toward the mean. And it has an age factor". Then a function is created that when a player's name is inputted, the function will output the projections for the following season.

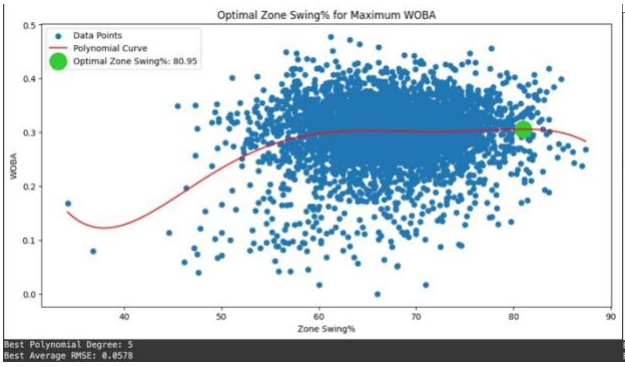
Results

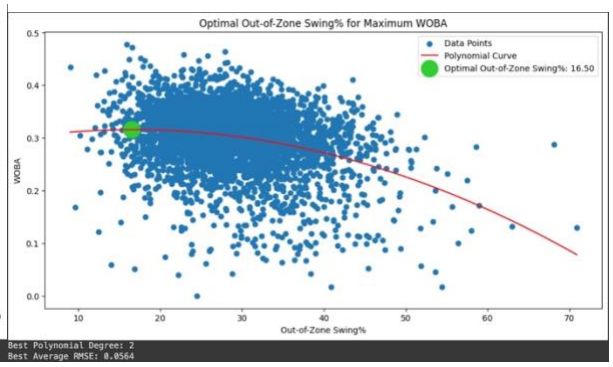
Batting Approach

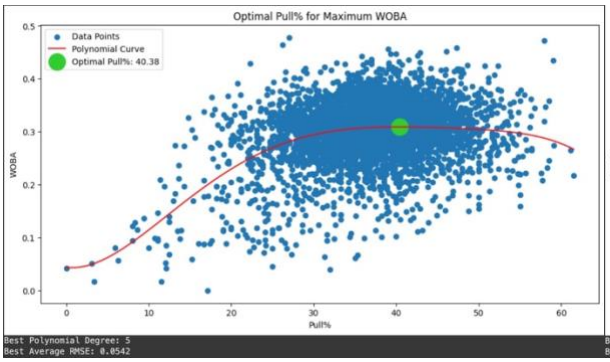
Below are all the graphs that go through each of the highlighted advanced statistics and their optimal points compared to WOBA.

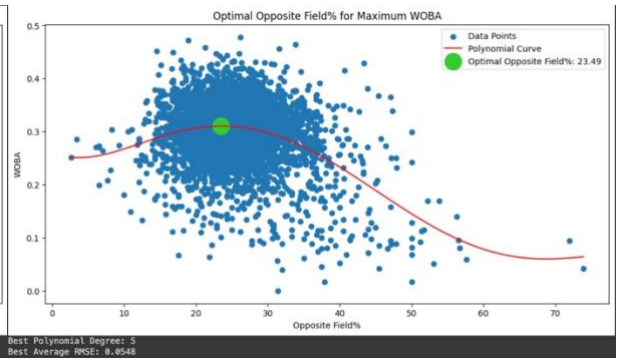




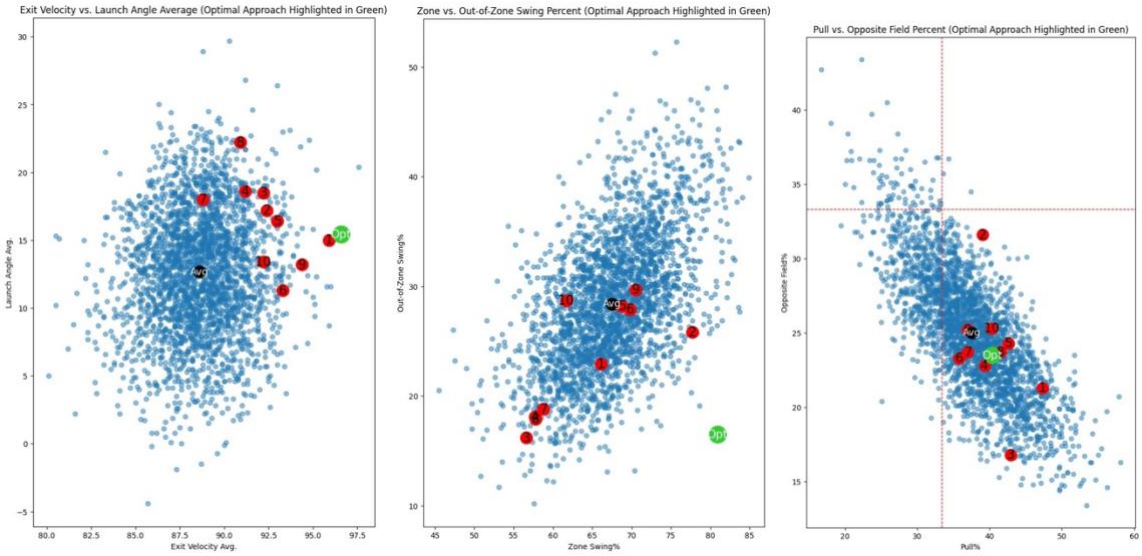








From top to bottom left to right, the graphs show the desired advanced statistics compared to WOBA are exit velocity, launch angle, swing percentage in the zone, swing percentage outside the zone, pull percentage, and opposite field percentage. The optimal point for all 6 of these graphs is highlighted in green so it can be clearly seen and compared with the rest of the players in the dataset.

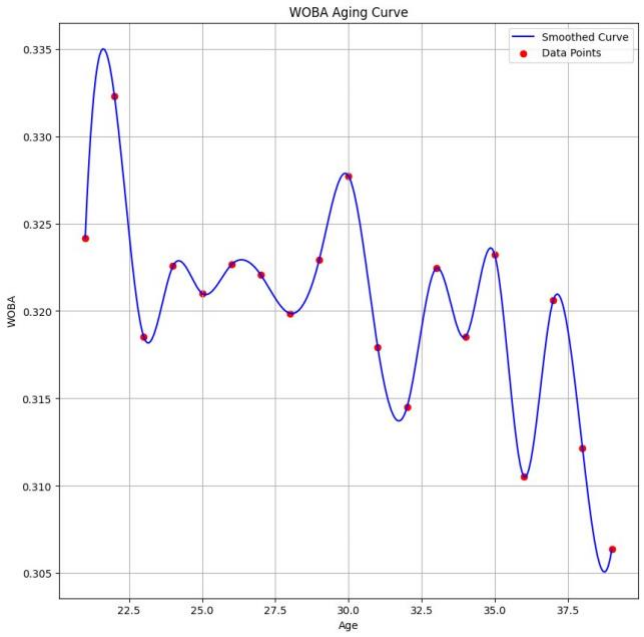


Above are the three graphs made during data exploration, but with the optimal points for each metric highlighted in green this time. A few interesting notes, on the first graph the optimal point is right next to Aaron Judge's 2022 season (the number 1 point in red), who had the highest WOBA season of the dataset, suggesting that Judge was spot on with his exit velocity and launch angle approach that season according to the regression. For the middle graph it's seen that ideally you want to swing at pitches in the zone and not swing on pitches outside of the zone. However, to be that disciplined at the plate is almost impossible. Finally, the third graph shows the optimal pull% vs opposite field% is close to the league average, suggesting that there are benefits to each side of the batting approach in those metrics. The optimal statistics for best WOBA are as followed:

- Exit Velocity: 96.59 mph
- Launch Angle: 15.44 degrees
- Zone Swing%: 80.95%
- Out-of-Zone Swing%: 16.50%
- Pull%: 40.38%
- Opposite Field%: 23.49%

Batting Projections

The second phase of the project investigated age and how it affects WOBA, and then how that information could be used to project a player's statistics from looking at the past three years.



	player_age	avg_woba_aging_curve	percent_change_woba
0	21	0.324167	NaN
1	22	0.332300	2.508997
2	23	0.318528	-4.144448
3	24	0.322566	1.267599
4	25	0.321004	-0.484090
5	26	0.322650	0.512606
6	27	0.322075	-0.178099
7	28	0.319864	-0.686439
8	29	0.322932	0.959150
9	30	0.327730	1.485678
10	31	0.317921	-2.992828
11	32	0.314497	-1.077152
12	33	0.322447	2.527905
13	34	0.318529	-1.215213
14	35	0.323211	1.469817
15	36	0.310529	-3.923484
16	37	0.320625	3.251089
17	38	0.312143	-2.645503
18	39	0.306364	-1.851467

Above is the curve that shows the average WOBA by age. From there, that curve was taken and the percent change to each age is found, appearing in the last column of the right image above. This percent change value is a part of how the players' statistics for the upcoming season are determined.

To determine the players projected statistics, a weighted mean was created based on recency of the season played, taking from the Marcel system. For my weights, and after some testing, I settled on a 60%, 30%, 10% weighting system for progressive distant seasons from that player's career. To explain further, each statistic was calculated by taking the 3rd most recent season stat value and multiplying it by .1, taking the 2nd most recent season stat value and multiplying by .3, taking the most recent season stat value and multiplying it by .6, and then finding the sum of those three values. Then, the projection was created by taking that weighted mean and multiplying it by the percent change for that age progression from the aging curve.

name	year	age	pa	hit	home_run	avg	ops	woba	ev	la	z_swing%	oz_swing%	pull%	opp_field%
Ohtani, Shohei	2021	26	639	138	46	0.257	0.964	0.393	93.6	16.6	69.7	27.3	46.6	22.9
Ohtani, Shohei	2022	27	666	160	34	0.273	0.875	0.370	92.9	12.1	72.1	28.4	36.0	27.8
Ohtani, Shohei	2023	28	599	151	44	0.304	1.066	0.433	94.4	13.2	70.5	29.7	37.0	25.2

name	year	age	pa	hit	home_run	avg	ops	woba	ev	la	z_swing%	oz_swing%	pull%	opp_field%
Ohtani, Shohei	2024	29	645	159	45	0.293	1.01	0.414	94.0	12.8	70.4	29.0	37.5	25.7

Projected (Scaled) To Real-Life 2024 Statistics

name	year	age	pa	hit	home_run	avg	ops	woba	ev	la	z_swing%	oz_swing%	pull%	opp_field%
Ohtani, Shohei	2024	29	110	27	8	0.293	1.01	0.414	94.0	12.8	70.4	29.0	37.5	25.7

name	year	age	pa	hit	home_run	avg	ops	woba	ev	la	z_swing%	oz_swing%	pull%	opp_field%
Ohtani, Shohei	2024	29	110	35	5	0.368	1.094	0.467	95	13.8	71.4	27.1	46.8	23.4

The two images from the previous page are the projections for Shohei Ohtani for the 2024 season based on the previous three seasons statistics. Next, above are Ohtani's projected stats through today, April 22nd, based on the 110 plate appearances he has taken this year so far, as seen in his actual stats through today. This year, he has not hit for as much power compared to his projections but is hitting for a higher average and still playing very well, with a higher WOBA than projected.

Problems and Future Considerations

There are several challenges and areas that could be worked on or considered for this project's future. First, for an ideal batting approach, everything needs to be taken into consideration, and these numbers do not tell the whole story. For example, off speed pitches do not get hit as hard because of simple physics as the harder the ball is thrown, the more likely it will be hit harder off the bat. Also, if there is a runner on first, the batter is more likely to try and hit the ball in the air to avoid a double play, these are just a few scenarios that would affect a batter's approach. Next, there is always some sort of luck associated with the game of baseball. You can hit the ball super hard at an optimal launch angle, but the ball is caught because a player is in the right spot at the right time. On the other hand, take weak contact to a soft spot in the opponent's positioning, and you end up on base.

For the aging curve, different positions in the field have different lifespans. Catchers tend to wear down a lot faster because their position takes a higher physical toll on a player, whereas first basemen can play for a longer time for the opposite reason.

Lastly, the COVID year and different juiced ball occurrences could have had a big impact on the dataset due to sample size concerns or changes in overall league hitting or pitching strength that were not accounted for.

For future considerations, it would be beneficial to predict rookie players or players still in the farm system, with more advanced minor league statistics as the predictions function needed 3 years of MLB data to make the predictions. With more time, applying the same process to pitchers could give insight to an entire other position group and could lead to interesting results as well. Overall, this project allowed me to find results related to hitting approach that could be used at all levels of an organization, for a player, coach or front office. Furthermore, I was able to apply the foundation of a widely known projection system along with an aging curve to make batter season predictions.

Works Cited

Baseball Savant: Trending MLB Players, Statcast and Visualizations baseballsavant.com. “Baseball Savant: Trending MLB Players, Statcast and Visualizations.” *Baseballsavant.com*, 2019, baseballsavant.mlb.com/.

Ansari, Aamir Ahmad. “Polynomial Regression with K-Fold Cross-Validation.” Medium, 2 Nov. 2021, aamir07.medium.com/polynomial-regression-with-k-fold-cross-validation-bc5275137546

“Checking in on the Aging Curve.” FanGraphs Baseball, 4 Oct. 2021, blogs.fangraphs.com/checking-in-on-the-aging-curve/.

“Marcel the Monkey Forecasting System (Marcel) | Glossary.” MLB.com, www.mlb.com/glossary/projection-systems/marcel-the-monkey-forecasting-system

“Weighted On-base Average (wOBA)| Glossary.” MLB.com, https://www.mlb.com/glossary/advanced-stats/weighted-on-base-average