Richard Burson

Prof Lin

Stats 520

15 May 2020

2019 MLB Hitters Data Analysis

The goal of every Major League Baseball team is to win a World Series. To do this it takes a complete team and not just one star player to do so. Mike Trout is an example of this, he is the consensus best player in the game and on track to be one of the best ever ,yet he still hasn’t even made the playoffs the last 5 years let alone compete for a World Series. It is fiscally impossible even for the biggest market teams to be able to pay a star player at ever position so an effective team building strategy is to try and find underrated or undervalued players by other teams across the league. Try and identify these types of players you can acquire for less money in free agency or fewer assets in a trade but will still be an effective player on your roster. The goal of this research was to find the best hitters of 2019 as well as some undervalued players that teams may be able to target for a lower price.

Data

The dataset that I used was comprised of 360 observations with 8 variables and it contained all MLB batters with at least 250 plate appearances in 2019.

* xBA – Expected Batting Average
* xSLG – Expected Slugging Percentage
* xwOBA – Expected Weighted On-Base Average
* xOBP – Expected On Base Percentage
* xISO – Expected ISO
* Average Exit Velocity – the Average Exit Velocity on balls hit by a player over a season
* Average Launch Angle - the Average Launch Angle on balls hit by a player over a season
* Barrel% - Percent of the time a player barrels up the baseball

All the “expected” variables are different than their traditional counterparts in that they are formulated using exit velocity, launch angle, and on ground balls it takes into account speed. So, each batted ball by a player is given expected values based on results that similar batted balls have had all season.

Methods

The methods used in this research were principal component analysis and cluster analysis. The original 8 variables were reduced to 2 principal components and then the players were clustered from there.

Analysis

After performing the first step of the principal component analysis the data got reduced down to 2 principal components with these weights.

![A screenshot of a cell phone

Description automatically generated]()

The Prin1 principal component appears to be an overall stat of how good a hitter is. The higher the value of Prin1 the better the hitter and the lower the value of Prin1 the worse he hitter. Prin2 is a bit more complicated because it is contrasting the launch angle, xISO, and barrel% variables with the xBA and xOBP variables. In baseball terms this principal component is comparing power hitters vs. contact hitters. Is a hitter has a high Prin2 value they are more of a power hitter whereas players with low Prin2 vlue are more of contact hitters. There is also a third option for Prin2 which is a player can have a value of close to 0 which can mean one of two things. Either the value is near 0 because the hitter is bad at both hitting for power and contact. Or the other option is that the hitter has both of these skills and can hit for contact and power.

![A screenshot of a cell phone

Description automatically generated]()

This is the plot of Prin1 vs Prin2. It makes sense to see hitters like Joey Gallo near the top since he is known as a power hitter. See players known as contact hitter near the bottom of the graph like DJ LeMahieu and Nick Markakis. Then you see plyers like Mike Trout and Nelson Cruz on the right side of the plot but have a prin2 value near zero. This is because they are elite hitters who can hit for power and contact. On the other had you have aplayer like Richie Martin who has a similar Prin2 value but is on the left side of the graph. Meaning he is a weak hitter that can’t hit for power or for contact.

![A close up of text on a white background

Description automatically generated]()

My cluster analysis yielded three clusters cluster 1 with 241 observations, cluster 2 70 observations, and cluster 3 with 49 observations. The second cluster that has 70 observations in it are the below average hitter, the first cluster with 241 observations are the hitters who are middle of the road hitters, and the third cluster with 49 are the above average hitters around the league. This is evident because all the stats across the board increase as you go from the cluster 2 to cluster 1 and then to cluster 3.

Conclusions

![A screenshot of a cell phone

Description automatically generated]()

The best 5 Prin1 values belong to Mike Trout, Nelson Cruz, Cody Bellinger, Christian Yelich, and Joey Gallo. These players are perennial all-stars and MVP candidates and the players with the lowest prin1 values are some of the worst in the league. So Prin1 seems to do a good job of predicting the positive impact a player can have for his team and who the best hitters are in the league.

Other than the outliers I plotted 4 other names in the middle of the plot of players that could possibly be under or overvalued. The three names in red Jason Castro, CJ Cron, and Justin Smoak are player I identified who could be underrated and possible players to target. Relative to other metrics their Prin1 value suggests that they could be better than originally thought. As for Gleyber Torres who is plotted with cluster 1 in the blue is someone I identified who could be overrated. He has been an all-star his first two years in the majors and is generally considered a good hitter but this analysis suggests he may be more of an average hitter.