Offensive Baseball Statistics

Michael Puiszis
Marquette University
MSCS Department
Data Science
Sprint 2017

Abstract: The increasing amount of attention being put on Baseball Player's statistical output has increased the need for teams to add statisticians to their coaching staff to select the best possible players. Metrics are also needed to define how teams should select the best available player. Using data from the top offensive performers from the 2010-2015, metrics were defined to select the statistically best single season performance by a player, as well as the best cumulative performance over during the defined time frame.

Author Keywords
Baseball; statistics; saber metrics;

ACM Classification Keywords
General and Reference, Social and Professional topics, Information Systems

Introduction: As baseball executives and coaches continue to utilize player data in decision-making processes when putting a team together, the need for good metrics to evaluate players becomes increasingly important. During the 2017 season, the Los Angeles Dodgers and New York Yankees total combined spending on player salary alone reaches nearly half a billion dollars. With all this money on the line, teams and owners want to ensure that they are getting the best value for their dollar.

Ever since the Oakland Athletics 2002 season in which the team spent the third least in the league on player payroll and still managed to tie for best record in the major league with the Yankees, team owners have changed how they evaluate prospective players. In 2002, the Yankees spent nearly three times as much money as the Athletics did on their respective rosters; because of the unorthodox way the Athletics' manager, Billy Beane chose his players. Relying on specific statistical metrics to evaluate potential players, Beane picked his 'skill position' players (Infielders, Catchers, and Center Fielders) by their On-Base and Slugging Percentages (The two calculations that the field of saber metrics is based upon), this caused the field of Saber-metrics to become widely used by teams to evaluate their rosters.

Technical Overview: With this in mind, I attempted to evaluate baseballs top offensive performers over a five-year period (2010-2015), and select the best amongst them. Before diving into the process and eventually conclusion of the analysis, defining the metrics used to evaluate the data is required. Players that are able to generate and score runs are a important asset to a team, and the stats associated with those skills are some of the most important ones to consider. The ability to get on base, and hit for power, are two of the other main fields to consider when grading an individual's offensive output.

Starting with an absolutely enormous data set from Baseball Reference.com [i], which contained all recorded statistics dating back to the founding of Major League Baseball (1890s). As one might expect, 125 years worth of offensive statistics makes for a rather large data set (over 80,000 entries, each containing upwards of twenty additional fields in each data entry). In order to make the data set easier to work with, and to cut down on the massive amount of data to analyze, I decided to use the most recently available data over a five-year span, 2010-2015. While this did help to reduce the overall size of the data set, a large amount of unnecessary and

unhelpful data still remained. Players that did not reach a minimum number of games played or total at-bats for a season were removed from the data set, so that only players that performed well on a daily basis were considered. Doing this helped immensely, bringing the number of entries in the data set down to just over 2000, more than one-fortieth of the size of the original set. In addition to this, players that failed to meet a minimum level of performance for certain statistical categories were also removed from consideration. Offensive standards included having a batting average greater than .200, having more than twenty RBI's (Runs Batted In), and at least thirty hits across a single season. Meeting these minimum standards are not the mark of a great player, in fact, players achieving only these benchmarks will not be playing for long. However, there exist players that shine in certain categories, but are below average in others. For example, power hitters historically have large numbers of Home Runs, Runs scored, and RBI's, but also tend to have lower batting averages. Mark McGwire for example is known as one of the best power hitters of all time, yet he only was able to have a batting average over .300 three times in his eighteen year career. [ii] So to ensure that players like Mark McGwire were not forgotten, the minimum standards were set low. Final changes to the data set included fields that would be unnecessary for this analysis, such as the number of times that an individual player was caught trying to steal a base. Once these unnecessary stats had been eliminated, additional statistics that did not exist in the data set (but could be easily calculated using the existing data) had to be calculated. Metrics such as On-Base Percentage (OBP), Slugging Percentage (SLG), and On-Base Plus Slugging (OPS) were quickly calculated by applying basic algebra to a given players season stats.

Once the data set was chopped down to a reasonable size, and contained all of the fields needed for analysis, the process of cleaning the data began. I started by separating out the various important fields that will be used to grade the numerous players, and selected the top fifty performers in each category. Doing this allowed for quick and easy visualization of the top performing players in a given category. The fields in each of these subsets included the player's name, the team they played for, and the significant statistic that the player achieved in a given year (e.g. Number of Home Runs in a particular season).

Using Seaborn, I was able to generate visualizations of the data,

which significantly helped to showcase the top performers in each category. Because of the way Seaborn visualizes the data in a strip plot[iii], players that had multiple seasons in which they led the league in a given category have multiple data points in a single row of the strip plot. Thus one can gather that a player with multiple data points in one of these visualizations was able to sustain a very high level of play, spanning a number of seasons, which is one of the most important things to consider when analyzing these stats.

After the main statistical fields had been visualized, I began looking for possible correlations between multiple pieces of data. Certain tendencies arise over the span of a season, and certain stats can be used to help predict a player's offensive performance in another category. A simple example of this is that a player with a high number of hits (e.g. more than 200 in a single season) will generally have a higher batting average (e.g. greater than .300). This is no surprise, as batting average is calculated by taking the total number of hits a players gets in a year divided by their total number of at-bats. Other observed correlations however, were significant. The number of times that a player struck out in a given year had interesting implications for other categories in a player's stat line. As the number of strikeouts increase, an individual player's batting average tended to decrease, but had the opposite effect on the player's slugging percentage. This is the opposite of what one might expect; as strikeouts negatively affect a player's slugging percentage, though some possible reasons for this occurring do exist. One explanation is that players with a higher slugging percentage generally tend to strike out more, because they are constantly swinging as hard as they can. This causes them to miss hitting a number of pitches, leading to the higher strike out numbers, but also allows the player to hit with more power. Beyond strikeouts, Homeruns were found to have positive correlations with both runs scored and runs batted in, while stolen bases were found to have a positive association with runs scored. Neither of these points was particularly surprising, however there was a significant observation found when considering on base percentage. There was a lack of correlation between on base percentage and the number of runs that a player scored in a given year. This was very surprising, as the only way one can score runs is if they get on base (or hit a lot of homeruns) a lot.

This concluded the overall analysis done to the data set; using the previously generated regressions and

visualizations, I began selecting the best players, to choose the best performer in a given year, and the player that able to maintain the highest level of play over the five year span.

Evaluation: For selecting the player that performed best over the five-year span, the visualizations that had been mentioned previously were a huge resource. Player's with multiple data entry points amongst the league leaders stood out as people to focus on when selecting a champion. The individual performer best fits this description is without a doubt Detroit Tigers First Basemen, Miguel Cabrera. Cabrera was among the top 50 performers in nearly every offensive category, almost every year between 2010 and 2015. He was among the league leaders in hits and batting average all five years. He appeared on the leaderboards four times for RBI's, three times for runs scored, and twice for home runs. During this time span, Cabrera received the lions share of season awards and accolades handed down at the end of each season, including two Most Valuable Player Awards, three Batting crowns (highest batting average in a given year), and in 2012 he became the first triple crown winner in forty-five years (To be eligible to win the triple crown, players must lead the league in Home Runs, RBI's, and Batting Average)[iv]. Other players have surpassed Cabrera's output in a single category in a given year, but none have come close to his sustained offensive contributions over the five-year time span.

It was also found that Cabrera had the single best offensive season of any other player during this time frame as well. No one else was able to achieve season totals that were as numerous or diverse across a number of categories as Cabrera. The only real problem that remained was selecting Cabrera's best season from this time period. The easy answer was to select Cabrera's Triple Crown season as the best individual season, but a closer look at the data revealed this might not be the case. Cabrera was able to follow up his Triple Crown season with one of the most impressive offensive years in baseball history. In 2013, he finished first in batting average and OPS, second in Slugging, RBI's, and Home Runs, and fourth in Runs Scored. The 2013 season saw a jump in Cabrera's numbers compared to his Triple Crown season in multiple categories; including Batting Average, Slugging, On-Base Percentage, and OPS. In the other statistical fields, such as homeruns and RBI's Cabrera was able to maintain the historic output he achieved during the 2012 season. The only reason Cabrera did not win the

Triple Crown again in 2013 was because of the 53 Home Runs that Chris Davis hit that season.

References:

---

[i] Baseball Reference.com, Retrieved May 12th. 2017: Baseball reference.com

[ii] Mark McGwire Offensive States, Baseball Reference.com, Retrieved May 12th 2017: http://www.baseball-reference.com/players/m/mcgwima01.shtml

[iii] Waskom, Michael; Strip Plot, Seaborn, Retrieved May 12th 2017: http://seaborn.pydata.org/generated/seaborn.stripplot.html

[iv] Miguel Cabrera, Wikipedia, Retrieved May 12th, 2017: https://en.wikipedia.org/wiki/Miguel_Cabrera