An iterative model for quantifying batter “approach”

2017 TruMedia Baseball Hackathon Submission
S. Brian Huey
sbhuey@gmail.com
github.com/brianhuey/trumedia
February, 20, 2017

Introduction

A good hitting approach can be defined as the batter’s ability to maximize his team’s run scoring expectation via the decisions he makes within his plate appearance. The decision to swing or to take a pitch should be informed by the count, number of outs and the runners on base prior to the pitch. It should also be informed by his assessment of the probability of getting a hit if he chooses to swing, and his assessment of the likelihood of a called strike if he chooses to take.

The weighing of a batter’s options is most plainly illustrated in three ball, no strike counts, where batters are less likely to swing at a probable strike; the intuition being that the cost of strike is relatively minimal, but the benefit of a ball is great. Viewed through the framework of run expectancy, we would expect the run expectancy from all of the outcomes that could occur from a swing (hit, out, swing and miss, foul) to be lower than the run expectancy associated with all of the outcomes that could occur from taking the pitch (walk, 3-1 count). The 3-0 count example is intuitive because it’s commonly practiced by many batters. It is less intuitive whether or not it is optimal to swing at a 2-1 pitch with 70% probability of being called a strike.

Previous studies have examined hitting approach by looking at rate stats such as strikeout rate and walk rate. Russell Carlton[1] expanded upon this by developing a methodology to measure a hitter’s strike zone judgement, contact rate and swing rate. Carlton uses the outcome after the pitch as the method of quantifying a successful or unsuccessful approach. A successful outcome from a swing is contact; an unsuccessful outcome is a swing and miss. A successful outcome from taking a pitch is a called ball, a called strike is considered unsuccessful. At the time of his study, PitchFX data was unavailable, so it was not possible to know if contact from a swing was on a pitch that would have otherwise been called a ball, or if a pitch that was taken for a strike was located on the edge of the strike zone.

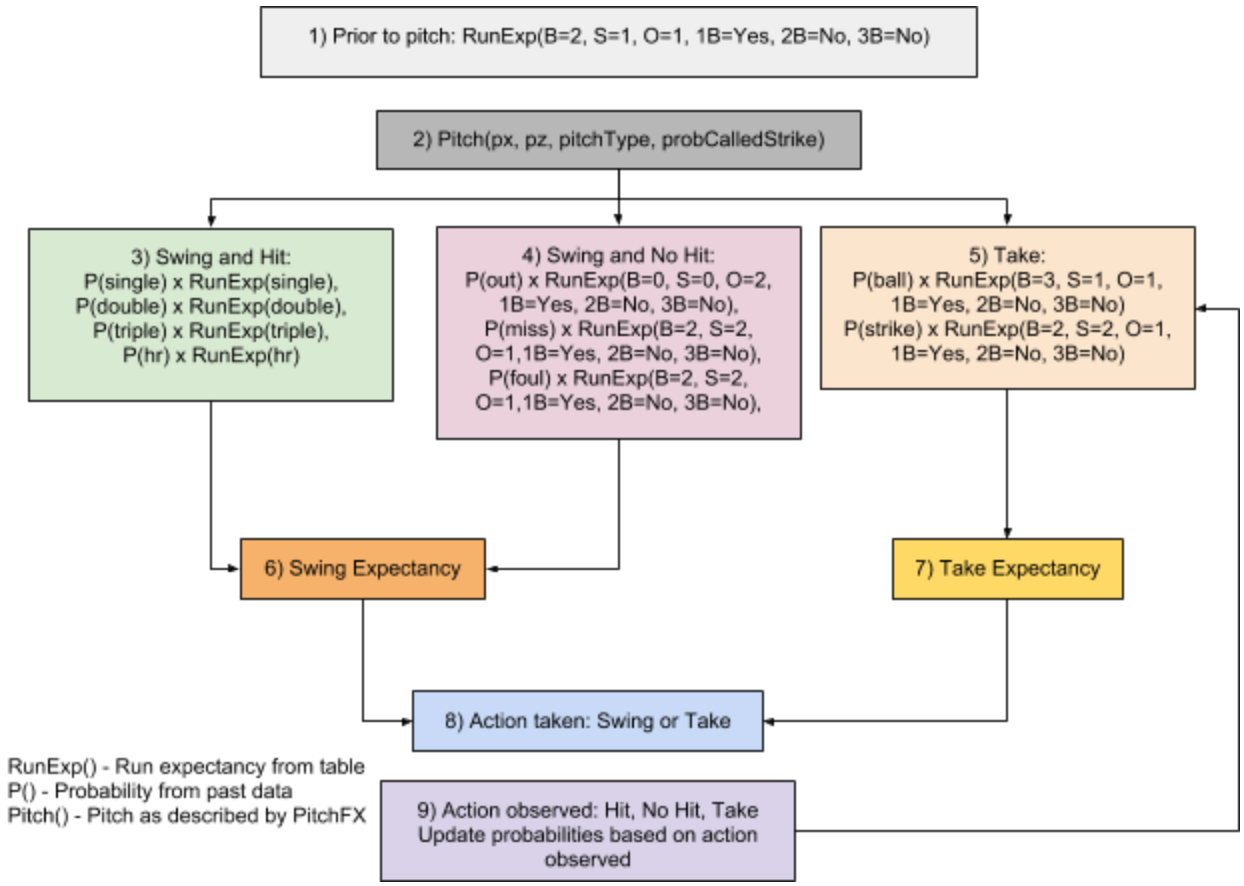
Building upon this framework, I seek to analyze pitch-by-pitch data to quantify successful and unsuccessful outcomes via run expectancy. I explore whether batters can demonstrate an ability or inability to leverage contextual factors (outs, balls, strikes, runners), information about the pitch (strike probability, location, type) and their past hitting history to reliably increase run expectancy within their plate appearance.

---

[1] Carlton, Russell “Is Walk the Opposite of Strikeout?” By The Numbers: The Newsletter of the SABR Statistical Analysis Committee, Vol 17, No 1, Feb 2007

Like Carlton's model, the batter is faced with two choices: swing or take, but instead of outcomes, the measure of success is whether the choice increased the team's run expectancy over the alternative. The model incorporates PitchFX and TruMedia's called strike probability metric, which allows the model to assess run expectancies weighted by probabilities conditional on the details of the pitch.

Methodology



The diagram above depicts the data modeling process for an example base state where the count is two balls, one strike, one out with a runner on first base.

1) The run expectancy is calculated prior to the pitch, which is referred to as the prior run expectancy.
2) The pitcher throws a pitch with the properties location, type and probability of being called a strike. From here the batter can either choose to swing or take the pitch.
3) If the batter chooses to swing, then the run expectancy from hitting the ball is the sum of the run expectancies for all possible outcomes, weighted by their probability. The four outcomes are single, double, triple, home run.

4) The batter may also swing and not get a hit, so the non-hit run expectancies, weighted by their probabilities, must also be added to the swing expectancy total. The possible outcomes are an in play out, swing and miss or foul ball.
5) If the batter chooses to take the pitch then the two possible outcomes are a called ball or called strike, with the run expectancies for both being weighted by the probability of the pitch being called a strike or a ball.
6) Swing expectancy is the sum of 3) and 4)
7) Take expectancy is equal to 5)
8) The batter is observed either swinging or taking. When the run expectancy from swinging is greater than the run expectancy from taking the pitch (i.e. 6 greater than 7 in the diagram), we would want to see the batter swing and vice versa.
9) Based on the action observed in 8), the outcome is used to update the probabilities for that batter when it comes time to generate 3), 4) and 5) next time.

Run Expectancy

Run expectancy tables are used to estimate the average number of runs a team would be expected to score given a certain state in the game. For example FanGraph's RE24[2] measures the change in run expectancy from the beginning of a player's plate appearance to the end of it. Others[3] have measured run expectancy conditional on the ball/strike count. These tables are generated by looking at past game data (for this analysis the 2015 season data set was used), grouping by base state and averaging the number of runs scored corresponding to each base state.

Because this analysis is concerned with run expectancies within a plate appearance, I generate two types of run expectancy tables that are more detailed than 24 base state table: expected runs conditional on outs, balls, strikes and base runners and expected runs conditional on hit type and base runners. The first is used to generate run expectancies prior to the pitch and after the pitch, and the second is used to generate run expectancies for the range of possibilities resulting from the batter hitting the ball.

Heat Maps

In order to generate probabilities associated with all possible outcomes resulting from swinging at a pitch, batter-specific "heat maps" are used to record the batter's past performance based on specified pitch criteria. After each swing, the heat map is updated to reflect the outcome, thereby updating the probability estimate.

Each collection of heat maps produces a probability of all possible swinging outcomes based on pitcher handedness, pitch type and location within a 5 x 5 grid centered on home plate and the

[2]http://www.fangraphs.com/library/misc/re24/

[3] http://www.nssl.noaa.gov/users/brooks/public_html/feda/datasets/expectedruns.html

vertical midpoint of the strike zone. Since there are seven possible outcomes (single, double, triple, home run, swing and miss, foul, in play out), two options for pitcher handedness and up to fifteen unique pitch categories, each heat map could contain up to 210 unique 5 x 5 matrices, not including matrices used to manage the total number of swings and pitches seen. The advantage of the heat map is that it's batter specific and provides a running total, so that changes in the batter's approach and pitchers' approaches to the batter are captured over time.

Model

The model was run sequentially for all games and plate appearances in the 2016 season, with the exception of pitches that resulted in bunt attempts, hit by pitches and other rare or uncategorizable outcomes. Because at the start of the 2016 season there would be no observed probabilities for the batters, the 2015 season data set is used as the initial condition.

For each batter, $i...k$, and each pitch, $j...n$, the run expectancy prior to the pitch ($RunExp_{Prior,i,j}$), run expectancy of a swing ($RunExp_{Swing,i,j}$) and the run expectancy of a take ($RunExp_{Take,i,j}$) was calculated.

In order to evaluate the batter's performance, $RunExp_{Swing,i,j}$ and $RunExp_{Take,i,j}$ must be measured in terms of their change from $RunExp_{Prior,i,j}$. The following calculations are made for each pitch:

$$\Delta Swing_{i,j} = RunExp_{Swing,i,j} - RunExp_{Prior,i,j}$$

$$\Delta Take_{i,j} = RunExp_{Take,i,j} - RunExp_{Prior,i,j}$$

Since $\Delta Swing_{i,j}$ and $\Delta Take_{i,j}$ measure the change in run expectancy regardless of whether the batter swings or takes, it is also necessary to measure both in terms of what the batter actually did. For instance, if $\Delta Swing_{i,j} > \Delta Take_{i,j}$, then swinging is considered optimal; if the batter is observed taking the pitch then the model assigns the batter $\Delta Take_{i,j}$, which is notated as $\Delta Take_{i,j}|Swing$.

$\Delta Swing_{i,j}|Swing$ - Run expectancy from swinging, swing observed (optimal)

$\Delta Swing_{i,j}|Take$ - Run expectancy from swinging, take observed (suboptimal)

$\Delta Take_{i,j}|Take$ - Run expectancy from taking, take observed (optimal)

$\Delta Take_{i,j}|Swing$ - Run expectancy from taking, swing observed (suboptimal)

$\Delta Gain_i = \sum_{j=0}^{n} \Delta Swing_{i,j}|Swing + \Delta Take_{i,j}|Take$ - Total gain from optimal decisions

$$\Delta Loss_i = \sum_{j=0}^{n} \Delta Swing_{i,j} | Take + \Delta Take_{i,j} | Swing$$ - Total loss from suboptimal decisions

$$\Delta NetTotal_i = \Delta Gain_i - \Delta Loss_i$$ - Net total from all decisions

Results

Seasonal Net Total

The leaderboard below shows the top 20 and bottom 20 batters (min 500 plate appearances) in net total run expectancy added ($\Delta NetTotal_i$) for the 2016 season. The top 20 includes a number of players who reside in the top 20 in WPA or offensive WAR, however there are a couple players that rank higher here. Similarly the bottom 20 has a couple players that are not included in leaderboards measuring offense.

Batter	NetTotal
Josh Donaldson	176.55
Mike Trout	172.37
David Ortiz	166.78
Ben Zobrist	161.83
Joey Votto	154.74
Carlos Santana	150.04
Bryce Harper	149.38
Paul Goldschmidt	148.28
Brandon Belt	140.91
Jason Kipnis	139.81
Kris Bryant	138.76
Edwin Encarnacion	135.34
Corey Seager	132.8
Nolan Arenado	132.11
Mookie Betts	130.97
Anthony Rizzo	130.79
Freddie Freeman	129.04
Jose Bautista	128.2
Matt Carpenter	124.05
Dexter Fowler	120.22

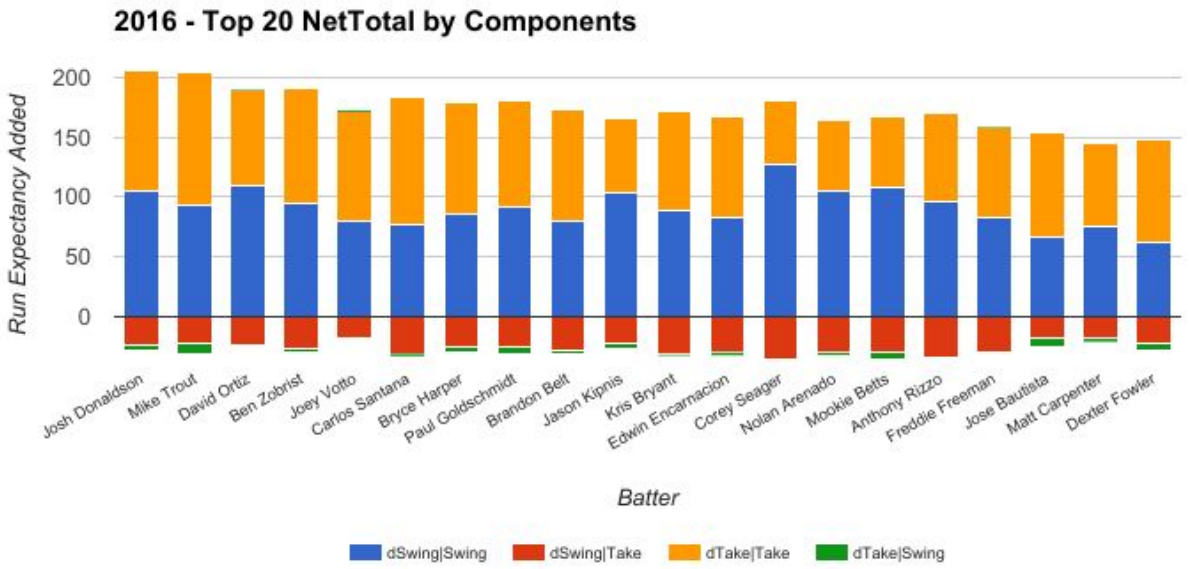
Batter	NetTotal
Adeiny Hechavarria	26.19
Alexei Ramirez	31.55
Freddy Galvis	32.28
Brandon Phillips	33.69
Didi Gregorius	35.53
Alcides Escobar	41.72
Wilson Ramos	44.11
Adonis Garcia	45.5
Chase Headley	47.36
Javier Baez	50.56
Jose Iglesias	50.87
Melvin Upton	51.07
Rajai Davis	51.22
Elvis Andrus	52.06
Starling Marte	52.15
J.T. Realmuto	52.18
Josh Harrison	53.45
Salvador Perez	55.52
Danny Espinosa	56.22
Cheslor Cuthbert	57.4

Seasonal Component Totals

In order to better understand the $\Delta NetTotal_i$ metric, it is useful to break down its component parts graphically. Below is a plot of the top ten in the leaderboard showing $\sum_{j=0}^{n} \Delta Swing_{i,j}|Swing$, $\sum_{j=0}^{n} \Delta Swing_{i,j}|Take$, $\sum_{j=0}^{n} \Delta Take_{i,j}|Take$, $\sum_{j=0}^{n} \Delta Take_{i,j}|Swing$ as individual components.

Looking at the components reveals that the majority of value comes from making optimal decisions, either swinging when it's better to do so, or taking when it's better to do so. Within that, some batters get to their total by generating a high amount of their net total by maximizing the impact of their optimal decisions (Josh Donaldson, Mike Trout), while others get there by minimizing the negative impact of their sub optimal decisions. In the cases of David Ortiz and Joey Votto, they even generate a small amount of positive run expectancy change from the suboptimal decision, meaning that when they took a pitch even though it would have been better to swing at it, they still improved their run expectancy from prior to the pitch.

For example, if a hitter has excellent contact skills, a bases loaded, no out situation has a high run expectancy associated with swinging, even if the ball has a low probability of being strike, but hitters like Ortiz and Votto also have excellent strike zone judgement and prefer to take the ball, have the count be in their favor and wait for a better pitch to hit.



Normalized Rates

Normalizing the net total and its components by pitches provides a better indication of a batter's approach since it makes comparable batters who didn't see the same number of pitches or

didn't play in the same number of games. Since the component rates are normalized by a different number of pitches, i.e. different subsets of n, the sum of the component rates will not equal the net total rate.

Looking at the individual components reveal some characteristics of batters. For example hitter aggression can be viewed in terms of the balance between $\Delta Swing_{i,j}|Swing$ and $\Delta Swing_{i,j}|Take$, that is the benefit of swinging when it's optimal minus the disbenefit of swinging when you shouldn't.

For example the table below shows that Jose Iglesias' run expectancy rate from his optimal swings, is comparable to Jason Kipnis. However Iglesias loses much more in run expectancy from swinging when it is not optimal to do so. There are also indications that Kipnis' approach is much more advanced when looking at how he takes pitches. Compared to Iglesias, Kipnis is able to use better strike zone judgement to gain a larger run expectancy advantage when taking pitches while not hurting himself by swinging at bad pitches.

I believe this has a lot to do with their different approaches. If Iglesias were to try and reduce the penalty of his suboptimal swings, he would want to do so while still maintaining aggression when it is optimal. This would not necessarily come from swinging less, but from swinging less at certain pitches and/or swinging less in game situations where there is a higher penalty for being wrong.

	Optimal	Sub-Optimal	Optimal	Sub-Optimal
Batter	$\frac{\sum_{j=0}^{n} \Delta Swing_i|Swing}{100\ pitches}$	$\frac{\sum_{j=0}^{n} \Delta Swing_i|Take}{100\ pitches}$	$\frac{\sum_{j=0}^{n} \Delta Take_i|Take}{100\ pitches}$	$\frac{\sum_{j=0}^{n} \Delta Take_i|Swing}{100\ pitches}$
Jason Kipnis	11.9954	-5.8908	8.744	-0.4811
Jose Iglesias	11.9487	-8.0785	5.6353	-1.0039

Conclusions / Further Work

This iterative model quantifying hitter approach produces a data set that can be used to evaluate how batters make decisions from pitch to pitch and how those decisions accumulate over the course of a season.

The hypothesis underlying the model is not that batters have run expectancy matrices or cumulative batting average in mind when they come to the plate, rather that the framework provided may potentially expose underlying skills that would not be observable by looking at outcomes after plate appearances or swing rate stats.

Further work is required to establish a link between what is being measured in the data set and actual skill. For example I create a relatively simple model of the information that hitters receive

from pitchers and also assume that all of that information (pitch location, pitch type) is perfectly observable to the hitter even though a primary goal of pitching is to deceive the hitter.

Similarly, the model used to generate probabilities of hit outcomes relies heavily on historical data and can take too long to reflect a change in approach or skill level. For example, Wilson Ramos was greatly penalized by the model for swinging when it said he should be taking because his hit probabilities were initially based on the 2015 season when he hit .229/.258/.358 over 504 PA. The model was not able to update Ramos' 2016 probabilities fast enough to understand it shouldn't assume he was hitting like he was in 2015 anymore.

Finally, further work is necessary in validating how the model's measures correlate with actual outcomes. For example, does swinging when it's not optimal produce outcomes that are also not optimal? As a predictive measure it would be expected to be correlated, but also subject to a high degree of variance.