Out “Hitting” the Odds

By Jack Arbuckle, Lauren Beaudreau, and Ben Scratz

Motivation

This project looks to analyze the probability of an MLB player getting a hit during a game. This idea stems from common hitter prop bets that have become very popular in online sports betting. Each day the odds of a person getting hit are released which usually fall between -400 and +200. Although most starters have a favored chance of getting a hit, sometimes this bet is parlayed together with a couple of players to create + odds. People look to capitalize on something that looks very enticing. All they need is one hit in about 5 to 6 at bats for the whole game. But as with most betting, the success rate usually falls with the sportsbook. Some of this betting failure comes from the nature of the sport. A better would think that a person hitting 0.300 would get a hit 3/10 at bats, so getting 1 hit per day seems very favorable. But unfortunately, baseball does not work like that. Baseball is a sport of average. Players hit 0.300 over the season, not 0.300 daily. Instead of going 1.5/5 each day, they might go ⅗ the first day and then 0/5 the next day. These drastic changes have to do with hitting streaks, pitching matchups, game situations etc. All these different factors make the hitter prop bets more of a tossup then they appear. Without reliable information it is hard to predict when these bunch hits or some hits or no hits will occur. Even the most productive hitters can go 0/5 any given day. The goal for this project is to create a model that takes hitter and pitcher metrics and situations into account to better predict the probability of that player getting hit for that certain day. The probability generated by the model for one or multiple players combined can then be used to make an educated bet that is more in the favor of the bettor than the sportsbook. The goal is to gain meaning behind the odds in order to bet smart, not bet on blind faith.

There is some previous research on this topic, especially since hitting prop bets and games like “Beat the Streak” [1]. Some of the data is a very basic interpretation of using batting average and other “old school” odds to build not predictions, but analysis. For these more basic insights there are many websites that have information on hitting probability and different line graphs for their average over time. [2] While this is a good analysis for other things like long term success and predictions, it fails to provide real usable value for betting on daily odds. On the other hand, there is some more advanced research on predicting hits for MLB players.

Numerous github pages had basic to advanced machine learning models that looked to analyze player hits on a daily and weekly basis. An example of one of the more advanced methods was a group [3] that used batter metrics, pitching metrics, stadium data, and weather data in order to try to predict the chances of a player getting a hit. They were inspired by the "Beat the Streak" game and wanted to see how long they could predict different players to get a hit on consecutive days. They also attempted with individual players on a given day but had mixed results due to limited machine learning models. At heart their end goal was to predict the odds of consecutively getting hits correct while this project's focus is on daily prop betting, not consecutive consistency.

Problem Framing

This project focuses on building an xgboost model to predict whether a player will get a hit on a certain day. It will use various metrics to use as explanatory response variables for the response variable of hit or no hit. The first metric used was the expected batting average metric. Expected Batting Average is a statcast metric that measures the probability that a batted ball will become a hit. Every batted ball is assigned an xBA on comparable hit balls in terms of exit velocity, launch angle, and hit location. In addition to xBA metric, the percentage a player hits balls with an exit velocity of 90 mph or above on certain types of pitches was chosen as an explanatory variable, as this is a better indication of solid batting and overall batting performance over time. By looking at exit velocities of 90 mph or above, the model can examine the data on a pitch by pitch level rather than just overall plate appearances. The last main metric of interest is Batter Z-O Swings, with Z-Swings being pitches inside of the strike zone that the batter swings at while O-Swings are pitches outside of the strike zone that the batter swings at. The Z-O Swing metric is calculated by subtracting the O Swings from the Z swings, therefore a higher value is better. The Z-O Swing Percentage is a better indicator of plate discipline. The predictive model also includes pitcher data such as xBA against, and types of pitches thrown. These two results will then be combined and put into a predictive model.

To build the predictive model that can be used for betting, every MLB player will be listed and a percentage of hits based on each pitch type that exists. For there to be a percentage listed, the batter must have 10 swings per pitch type and 25 overall batting appearances. This is to eliminate the players with a small sample size that may skew the data. Since the project goal is

to be able to predict how many hits an MLB player will get during a given game, it is necessary to include all three of these metrics in the predictive model. In simple terms, the model will predict how many hits batter X will get when facing pitcher Y who tends to throw certain types of pitches.

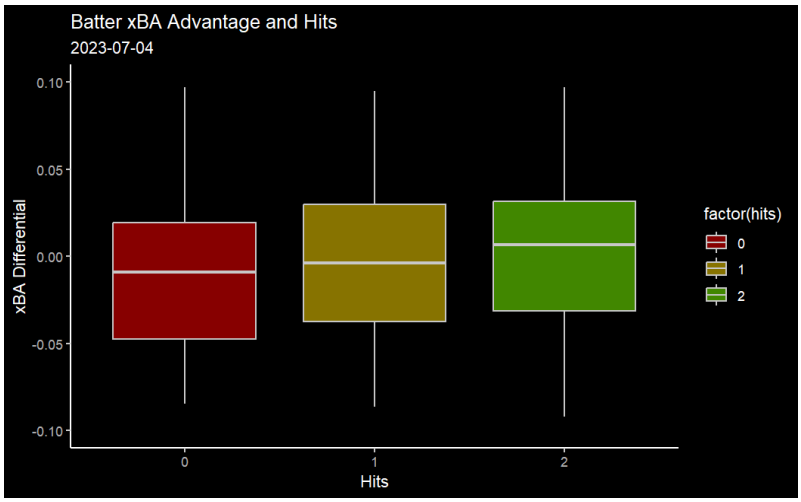
Data Overview

The dataset that will be used to build our model is the 2023 Statcast data. This dataset shows every pitch from each game throughout the season. Each row represents an individual pitch with 99 variables attached to that pitch. Pitch type, game date, pitcher ID, batter ID, and the event that occurred from each pitch are some of the variables of most interest to us. Statcast_2023 has 729,774 observations.

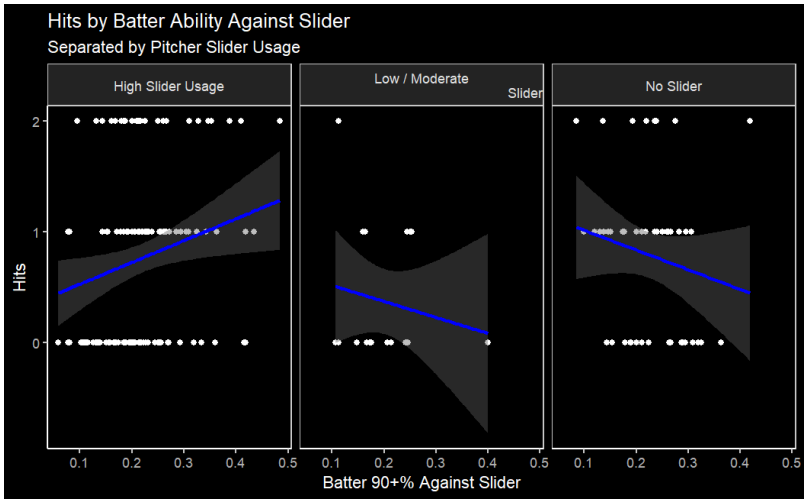
From the original dataset, statistics will be calculated, forming new tables grouped by batters, pitchers, and dates. First, the batter_stats table displays the statistics described in detail above, namely xBA, 90+%, and ZO-swings, for a given date. Pitcher_stats displays xBA, 90+%, and pitch usage by date. Finally, bullpen_data displays xBA for all relievers by team.

In creating a complete data set for testing, all statistics to date will be joined to the corresponding game results: did the hitter get a hit on the particular date, either against the starting pitcher or bullpen, whose statistics are described in the row.

Data Visualization



This graph shows the distributions of xBA differential (Batter xBA - Pitcher xBA) for matchups where the hitter records 0,1, and 2 hits in the game. It shows that the differentials were clearly higher in games where the batters recorded more hits.



This graph shows the relationship between hits and the batter's ability against a certain pitch (slider). It is separated by the frequency with which the pitcher throws a slider. It shows that a hitter's ability against the slider is only relevant when the pitcher throws a high frequency of sliders. This demonstrates that it is important to include pitch-by-pitch usage as well as abilities.

Contribution

Jack 33% Motivation and Code

Ben 33% Data Overview and Code

Lauren 33% Problem Framing and Code

Bibliography

[1] **https://www.mlb.com/apps/beat-the-streak/official-rules**

[2] **https://blogs.fangraphs.com/the-odds-of-hitting-for-the-cycle/**

[3] **https://eglouberman.github.io/MLB-hit-predictor/.**