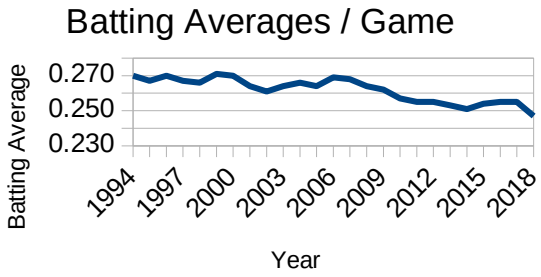
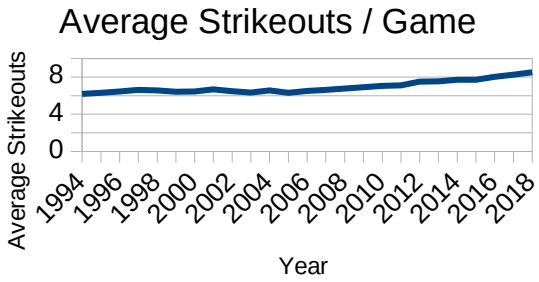
Predict Pitch Thrown by Major League Baseball (MLB) Starting Pitchers Using Machine Learning

Capstone I Project for Springboard Data Science Career Track – Mark Rojas

PROBLEM

According to Ted Williams, one of the greatest hitters of all time, "Hitting is the single most difficult thing to do in a sport". Having less than 400 milliseconds to hit a fastball traveling about 95 mph, it should be humanly impossible to hit a baseball. Even the best hitters today average less than 4 hits per 10 at-bats and since 1994, the average hits per game have decreased while the number of strikeouts per game has increased[1]. In addition, the inconsistency from the home plate umpire in how they call balls and strikes further complicates the art of hitting a baseball.





It appears that any advancements in training methods, technology, and preparation seem to have benefited pitchers more so than the hitters. While hitters can rely on historical data and past experience from facing a pitcher multiple times to guesstimate the next pitch, this does not take into account the current game-time variables that impact the pitchers decision making, and as the great Yogi Berra once stated, “You can’t think and hit at the same time”.

Because of this, to assist hitters, opposing teams have resorted to stealing signs from the catcher and observing the pitcher, like in poker, looking for telltale signs that could give away the next pitch. For example, in the 2017 Major League Baseball (MLB) World Series between the Houston Astros and Los Angeles Dodgers, the Dodgers starting pitcher for game 3 and the decisive game 7 was Yu Darvish. In both games, Darvish was quickly removed in the 2nd inning after giving up 4 and 5 runs, respectively. Following the series in which the Astros won the best of seven, it was later revealed in a Sports Illustrated article that Darvish was “tipping” his pitches, meaning he was unknowingly telegraphing what pitch he was going to throw next.

1. https://www.baseball-reference.com/leagues/MLB/bat.shtml

The problem is that not every pitcher "tips" their pitches and stealing signs is extremely frowned upon and usually leads to retaliation through hit bats-man, potentially causing injury.

CLIENT

While there are batting coaches, trainers, and managers that assist hitters throughout the season, I will say that my client would be the **hitters** themselves. Being able to provide them with real-time updates during the game including the expected pitch with say, 70% - 80% accuracy, would greatly improve their batting averages.

DATA

The primary data is available through **PITCH*f/x*** which is a Pitch tracking system, created by Sportvision and is installed in every Major League Baseball (MLB) stadium since around 2006. This system tracks the velocity, movement, release point, spin, and pitch location for every pitch thrown in baseball, allowing pitches and pitchers to be analyzed and compared at a detailed level.

http://baseball.physics.illinois.edu/pitchtracker.html

Note -- The link doesn't really take you to the data, but does provide some good information. I have already downloaded the data using R (though Python can be used also) and stored in SQL database approximately 6 GB in size.

Secondary datasets will be collected from:

1. **The Lahman Database:** contains complete batting and pitching statistics from 1871 to 2017, plus fielding statistics, standings, team stats, managerial records, post-season data, and more.

   http://www.seanlahman.com/baseball-archive/statistics/

2. **Brooks Baseball Website:** the most complete, accurate, and comprehensive data set about pitching available on the web. Utilizes the PITCH*f/x* data set and makes systematic changes that improve the quality, usefulness, and usability of that data.

   http://www.brooksbaseball.net/pfxVB/pfx.php

Additional data available, if needed:

3. **Major League Baseball Website:** official MLB website

   https://www.mlb.com/

4. **Baseball-Reference Website:** complete source for current and historical baseball players, teams, scores, and leaders.

   https://www.baseball-reference.com/

METHODOLOGY

Most MLB pitchers throw about 4 different pitch types in a game. These pitches are then varied based on grip, speed, and location. This capstone project will focus primarily on predicting discrete values (e.g., Pitch types) so I plan to approach this problem as Supervised learning classification, specifically, logistic regression. However, I expect that I will need to consider other classification techniques to appropriately assess performance.

DELIVERABLES

- Report with Results
- Collections of Resources utilized
- PowerPoint Presentation available on Slide Deck
- Code available through GitHub
- Video Presentation