Pitch Prediction: Statistical Learning

Christian Stratton, Andrew Hoegh, Jennifer Green
Montana State University
Department of Mathematical Sciences

January 28, 2020

Abstract

Keywords:

1 Introduction

Analytics, or more specifically sabermetrics, have long played a role in strategical choices in Major League Baseball (MLB). Most famously chronicled in the book Moneyball Lewis (2004), sabermetrics were used to construct a roster of players primarily using On Base Percentage (OBP) rather than traditional statistics such as batting average. While baseball organizations tend to keep the most useful sabermetrics techniques proprietary, there are books (Baumer and Zimbalist, 2014; Law, 2017) and entire conferences denoted to applying statistical techniques to the game of baseball. Koseler and Stephan (2017), and the reference therein, provide an overview of many statistical and machine learning based approaches for analyzing baseball data.

From a hitter's perspective, sabermetrics tools to determine the type of pitch that will be thrown would be a major advantage. While we are not aware of publicly available, and perhaps more importantly, legal methods for predicting pitch sequencing; in recent years,

teams have used illegal methods to video record, decipher pitch signals, and relay them to hitters. This manuscript explores the use of statistical modeling to predict the type of pitch thrown in various scenarios.

In baseball, the catcher signals for a certain type of pitch by using a serious of hand signals. The signals are typically only seen by the pitcher and pitcher's team, unless a runner is on base. The signals determine the type of pitch that the pitcher will throw on the next pitch.

Baseball is a game with many "unwritten rules", one of which would be stealing signs. Sign stealing is an accepted part of the game; however, using electronic equipment to steal signs is not permitted.

In 2017, the Boston Red Sox were caught using electronic devices, an Apple watch, to send signals from the video replay room to the dugout. The defining feature of the verdict was the use of an electronic device. The New York Yankees were also accused of using the Yankees network to gain a competitive advantage.

On January 13, 2020 MLB imposed one of the largest penalties in history on the Houston Astros for a scandal that involved video recording hitters to steal pitch signs and then relaying signals by banging on trash cans. The Astros were fined 5 million dollars, the maximum fine allowed in MLB; stripped of first and second round draft picks for multiple years; and the manager and general manager (GM) were suspended for one year. The manager and GM were ultimately fired by the organization. The Boston Red Sox manager, Alex Cora, and the New York Mets manager, Carlos Beltran, both former Houston Astros employees were also fired by their respective organizations in wake of the - a former Houston Astro are still under investigation.

In recent years, MLB changed the rules to permit the use of technology in the dugout as teams are now allowed to use league-provided iPad/laptops.

In line with historical advances in analytics in baseball and the sabremetrics movement, we seek to explore using statistical models to predict the next pitch thrown depending on the

count and other scenarios. To be clear, the goal is not to use electronic equipment to decode signs, but rather statistical learning tools are used to decipher patterns in pitch sequencing. Section 2 describes the data used in this analysis. Section 3 highlights the statistical models used for prediction as well as the loss functions used to evaluate different models. Section 4 describes the model results and Section 5 concludes with a discussion.

2 Data

The data used for this analysis come from Pitch Fx and specifically the Pitch Rx package Sievert (2014) is used to scrape data into R. The Pitch Fx data is captured by a camera and contains several variables about each pitch, one of which is the pitch type.

3 Statistical Framework

3.1 Loss Functions

3.2 Model Specification

3.2.1 Polya-Gamma Hierarchical Logistic Regression

3.2.2 Bayesian - CART

3.2.3 BART

4 Results

5 Discussion

References

Baumer, B. and Zimbalist, A. (2014). *The sabermetric revolution: Assessing the growth of analytics in baseball*. University of Pennsylvania Press.

Koseler, K. and Stephan, M. (2017). Machine learning applications in baseball: A systematic literature review. *Applied Artificial Intelligence*, 31(9-10):745–763.

Law, K. (2017). *Smart baseball: The story behind the old stats that are ruining the game, the new ones that are running it, and the right way to think about baseball*. William Morrow New York.

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. WW Norton & Company.

Sievert, C. (2014). Taming pitchf/x data with xml2r and pitchrx. *R Journal*, 6(1).