Pitching Repertoire of Trevor Bauer - STAT 432 Analysis I

Justin Kim (yundong2@illinois.edu (mailto:yundong2@illinois.edu))

05/04/2021

Abstract

This analysis is being done to develop a method to help fans know what pitches are being thrown in a setting such as a live broadcasted MLB game. Machine learning methods such as KNN and decision trees were used to build model and predict what pitches were being thrown. The results indicate that it is easier more accurate to predict pitches by pitchers, instead of predicting using data from all pitchers. This means that with more data available from pitches, and knowing what pitches a specific pitcher has thrown before, the more accurate prediction we weill ble able to make from the classifier.

Introduction

It is extremely difficult to identify what pitch a pitcher is throwing by looking at the pitch. This is because no pitcher in the world throws the same pitch as another pitcher. Each pitcher has their own unique qualities, such as their delivery form, their grip and release point, the velocity and spin rate of their pitch, whether the ball moves horizontally or vertically, and much more. Each pitcher's unique delivery makes it hard to classify a pitch with an overall criteria. However, a common theme all pitchers share, is that they will have consistent qualities in throwing certain pitches, making it easier to classify pitches by pitchers. This means it becomes easier to classify pitches by a pitcher, instead of several pitchers.

Even though it becomes easier to classify pitches by a pitcher, it is still difficult for average fans to identify what pitches a pitcher is throwing. Instead, we will train a machine to detect pitch type thrown by the pitcher. This will be useful in live MLB games, where there could be deep learning computer live that could tell what pitch a pitcher has thrown by analyzing the pitch thrown.

Methods

To prove that it is easier to classify a specific pitcher's pitches given the pitcher's data, we will look at Trevor Bauer, one of MLB's best pitchers currently. First, we subset pitches thrown by last name "Bauer," since there are no other pitchers with last name "Bauer" currently playing in the league. We also remove unwanted columns such as batter name and date, and remove and NA values before working with the data.

We first split the data into test and training data, by randomly selecting 80% of Bauer pitching data to be the training data, and the remaining 20% to be test data. Then we use knn3 function to fit a model to classify by “pitch_types.” Because we do not know the best K value for fitting, we fit several k values from 1 to 100 by increments of 2, and choose the model with the highest test accuracy. We find that the highest test accuracy comes from k value 1, with test accuracy of 97.2%.

Accuracy of 97.5% is very good, but we see if fitting a Decision Tree on the data can yield better results. By following the same procedures as KNN on the train data, we fit several cp values to the decision tree and choose the model which had the highest test accuracy for predicting pitcher types. We see that cp value of 0.00001 had the highest accuracy of 99.85%.

Data

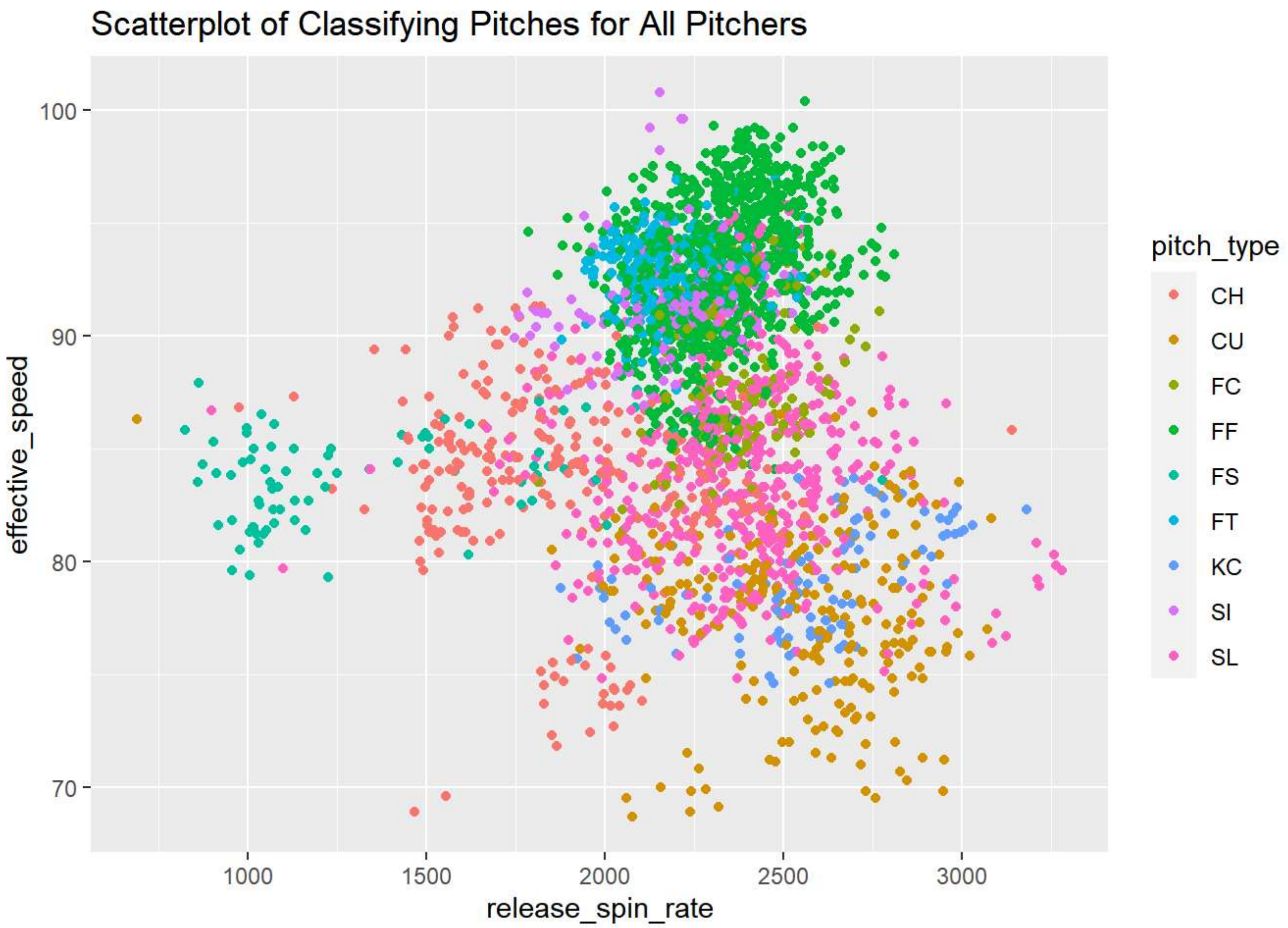
The dataset used is a dataset of pitches from a particular month of the 2019 regular season. Each row is a specific pitch thrown by a pitcher, with information of the pitch, such as speed and location. Below are the important columns of the dataset that will be used for modeling.

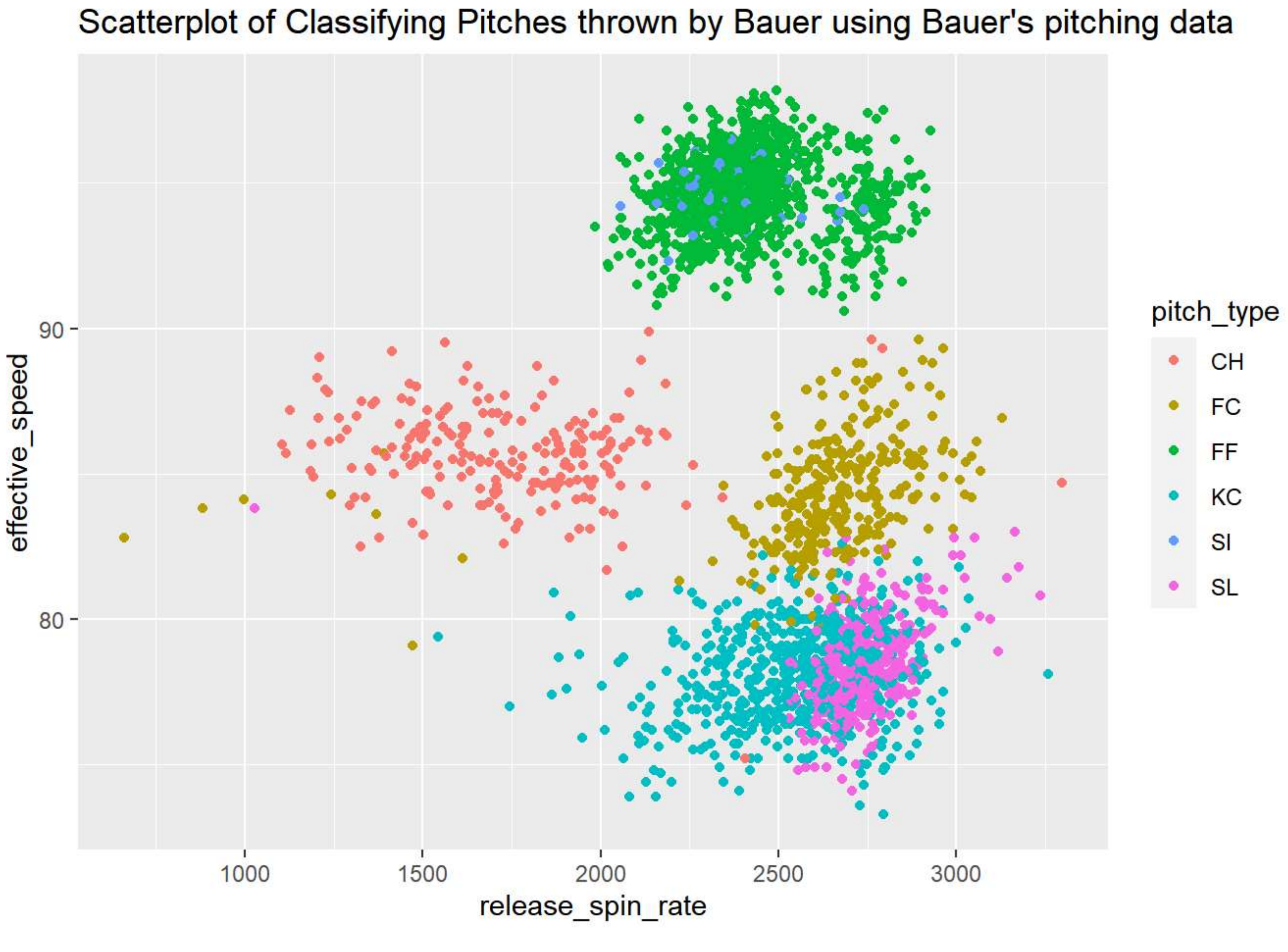
- pitch_type : Type of pitch thrown
- release_speed: Pitch velocity
- release_pos_x: Horizontal Release Position of the ball measured in feet from the catcher’s perspective.
- release_pos_y: Release position of pitch measured in feet from the catcher’s perspective.
- release_pos_z: Vertical Release Position of the ball measured in feet from the catcher’s perspective.
- pfx_x: Horizontal movement in feet from the catcher’s perspective.
- pfx_z Vertical movement in feet from the catcher’s perpsective.
- plate_x: Horizontal position of the ball when it crosses home plate from the catcher’s perspective.
- plate_z: Vertical position of the ball when it crosses home plate from the catcher’s perspective.
- vx0, vy0, vz0:The velocity of the pitch, in feet per second, in x,y,z-dimension, determined at y=50 feet.
- ax, ay, az: the acceleration of the pitch, measured at the initial point in 3D (ft/s)
- effective_speed: Derived speed based on the the extension of the pitcher’s release.
- release_spin_rate: Spin rate of pitch
- release_extension: Release extension of pitch in feet as tracked by Statcast.

The information of these variables were pulled from the Statcast Search CSV documentation.

Results

```
## [1] 0.9970545 0.6954712
```


Scatterplot of Classifying Pitches for All Pitchers
effective_speed
100
90
80
70
release_spin_rate
1000
1500
2000
2500
3000
pitch_type
CH
CU
FC
FF
FS
FT
KC
SI
SL


Scatterplot of Classifying Pitches thrown by Bauer using Bauer's pitching data
effective_speed
90
80
release_spin_rate
1000
1500
2000
2500
3000
pitch_type
CH
FC
FF
KC
SI
SL

Our results show higher accuracy for classifying and predicting pitches when subsetting for pitches. When predicting Bauer's pitch using pitches thrown by Bauer, we reached accuracy up to 99.85%. When predicting pitches by classifying pitches thrown by several pitchers, we were not able to obtain high accuracy. The scatterplots shown reinforces this. When plotting release spin rate vs effective speed for pitches thrown by Bauer, we do not see many overlapping pitches, in big contrast with the first scatter plot. We see how much easier it is to classify a pitch type than the 1st graph of all pitchers.

Discussion

Our results show that it is easier and much more accurate to predict a pitcher's pitch when subsetting by specific pitchers. When predicting pitches thrown by Trevor Bauer, we were able to reach high accuracy. However, when predicting pitchers without subsetting by specific pitchers, it became harder to predict what pitch was thrown, and we achieved low accuracy. This is because, as stated before, each pitcher is unique, and each pitch is unique. For example, is more to a regular four-seam fastball than its velocity and its straight delivery. Some pitchers throw four-seam fastballs with high spin-rate, which may result in higher vertical movement than others. Another example is how pitchers throw sliders and curve-balls differently. They may throw it at different speeds and less spin rate resulting in different type of movements at the plate. Some pitches have obscure boundaries of what type of pitch it is, even if defined by the pitches. Some pitchers may throw a slider that is similar to a cutter, and some may throw a two seam fastball that is similar to a cutter, which shows how important it is to obtain data of what pitch is thrown, identified by the pitcher.

An obvious weakness of this method of classifying a pitch is that it would not work as well if there are no prior hand information of the pitcher. Well established pitchers such as Trevor Bauer and Gerrit Cole have been in the major leagues for a long time and thus many data collected about their pitches and what type of pitches thrown. For pitchers that are new to the big leagues, such as pitchers called up from the minor leagues, we would have to rely on the less accurate method of predicting pitches using all pitchers data.