Capstone Project 1 - Building a Fastball Prediction Model

Milestone Report

I. Problem Statement

Strategy in Major League Baseball is ever-involving. With hitters increasingly focused on launch angles and elevating the ball, it is unsurprising that teams are becoming more reliant on home runs to score. In fact, in 2019 the teams collectively hit an eye-popping 6,776 home runs, which is roughly a 10% increase over the previous record of 6,105 home runs hit in 2017.[1] This trend is very likely to continue in the coming seasons.

It also is generally-accepted that fastballs are particularly ripe for batters to hit for home runs given their increased velocity and lack of movement relative to other pitch types. Indeed, Fantasy Labs, a website focused on providing strategic advice to fantasy sports players, noted that four-seam fastballs were the most common pitch type to be hit for home runs.[2] And simple common sense dictates that a batter will have an even greater chance of hitting a home if he ***knows*** that the pitcher is going to throw a fastball - one only needs to watch batting practice or the home run derby to see this confirmed.

The goal of this project is to build a model that can accurately predict when a pitcher will throw a fastball. The targeted clients for this model are professional, semi-professional, and collegiate baseball teams, including the players, managers, and other personnel. The organizations could use this data to inform their batters to anticipate fastballs under certain circumstances, which in turn should lead to the team hitting more home runs. Conversely, organizations could use this information to inform their pitchers as to when fastballs are most often thrown and coach the pitchers to throw a different pitch under those circumstances. Professional sports gamblers could also use the model to inform their wagers. These are just a handful of what is likely many uses for an accurate fastball prediction model.

---

[1] Data compiled and published by Baseball Almanac at **https://www.baseball-almanac.com/hitting/hihr6.shtml**

[2] **https://www.fantasylabs.com/articles/home-run-trends-part-3-pitch-type/**

II. The Datasets

This project uses the MLB Pitch Data 2015-2018 dataset that is publicly-available on Kaggle at (https://www.kaggle.com/pschale/mlb-pitch-data-20152018). I have chosen to build my model using two .csv files from that dataset. The first file, 'pitches.csv', charts various data for each pitch thrown during each of the four seasons from 2015 through 2018. The second file, 'atbats.csv', contains various static data for each at-bat from each of the four seasons from 2015 through 2018. The two files have the following dimensions:

'pitches.csv' == 2,870,000 x 40
'atbats.csv' == 740,000 by 11

One of the categorical variables in the 'pitches' file, 'pitch_type', classifies the pitch type for each pitch thrown. The variable contains over fifteen distinct labels, including 'FF' for a four-seam fastball. I will be able to use these labels to create a new binary variable to classify each pitch as a fastball or non-fastball, which will become the target of my analysis.

III. Data Cleaning/Wrangling

The data in its original form was in pretty good shape given that it came from Kaggle. That being said, I still needed to perform some cleaning and wrangling on the respective files in order to prepare the data for analysis.

Step 1 - Merging the Datasets

I first combined the two .csv files into one dataset. The two .csv files were merged (using 'pd.merge()') on the common 'ab_id' column, which provided a unique numeric identifier for each at-bat in the 'atbats' file and each pitch thrown during each at-bat in the 'pitches' file.

Using the 'nunique()' method in pandas, I was able to determine that the 'pitches.csv' dataset has data recorded from 740,241 unique at-bats whereas the 'atbats.csv' has data recorded from 740,389 unique at-bats. Because the 'pitches.csv' file is the dataset with the 'pitch_type' variable that is the focus of this analysis, I decided to left-join the 'atbats.csv' dataset to the 'pitches.csv' so that the extra at-bats in the 'atbats.csv' file would drop out. Given the large volume of data in the 'pitches.csv' data file (approximately 2.87 million unique pitches), I was comfortable only joining the corresponding at-bats from the 'atbats.csv' file.

Once merged, the resulting dataframe has a dimension of (2867154, 50).

Step 2 - Addressing the Missing Data

A number of columns in the newly merged dataset contained some 'NaN' values. Some of the columns in the dataset are continuous numerical datatypes (floats and ints). There were approximately 14,000 pitches that were missing almost all of their continuous numerical data. Because all of the continuous numerical variables are tracking characteristics of the pitch after the pitcher has released it, it appears that the data collection device simply failed to track these pitches.

For this missing data, I elected to replace the 'NaN' values with the mean value of each variable. I did this because I wanted to maintain the ability to perform exploratory data analysis on these variables, which would not be possible if I replaced the 'NaN' values with a non-numeric value (such as a 'MISSING' string) because it would convert the columns to an 'object' datatype. I also did not believe that replacing the 'NaN's with the mean values would skew any calculations significantly given that the missing data makes up substantially less than 1% of the entire dataset.

I took a different approach regarding the 'NaN' values for the categorical variables. The first categorical variable with 'NaN' values was our target 'pitch_type' variable, which contains string values, including 'FF' for four-seam fastballs. I investigated into whether other columns of the dataset may inform whether the unlabeled pitches were fastballs (especially the 'start_speed' and 'end_speed' columns). Unfortunately, all but five of those entries had 'NaN' values for both start and end speeds, and the five that did have entries appeared to be unreliable because they contained abnormally slow outliers. I therefore decided to replace the 'NaN' values with the string 'MISSING' in order to make it clear that there is no pitch_type label for those particular pitches.

I also replaced the 'NaN' values with 'MISSING' string values in the 'code' column because it similarly contains categorical string values.

Step 3 - Creating Dummy Variables for Certain Categorical Variables

.The dataset contains six columns of categorical data of interest that may need dummy variables - 'pitch_type', 'p_throws,' 'stand,' 'top,' 'type', and 'code.'

The 'pitch_type' column is our target variable and it contains a number of different string labels, including 'FF' for four-seam fastballs. Because the goal of the project is to

predict when a four-seam fastball will be thrown, I created a new 'fastball' column that contains a value of 1 for every pitch that was labeled with the pitch type 'FF' and a value of 0 for every non-'FF' pitch type category.

The 'p_throws', 'stand', and 'top' columns are all binary categorical data. Both the 'p_throws' and 'stand' columns contain either an 'R' or 'L' depending on whether the pitcher and batter are right or left handed, respectively. The 'top' column is a boolean datatype that is 'True' for every pitch thrown in the top half of an inning and 'False' for every pitch thrown in the bottom half of the inning. I therefore created binary numerical columns for each of these columns (1 for 'R' and 'True' values, 0 for 'L' and 'False' values) titled 'p_throws_num', 'stand_num', and 'top_num.'

The final two columns, 'type' and 'code', track the same information regarding the outcome of the pitch. The 'type' column provides only three different categories ('B' for ball, 'S' for strike, and 'X' for ball in play), while the 'code' column contains more detailed categorical data (e.g., hit on ground, popped up, foul ball, etc.). In the interest of limiting the total number of columns in the dataset to keep the file a manageable size, I elected to only add three dummy columns ('type_B', 'type_S', and 'type_X') to the dataset using the .get_dummies() method and pd.concat(). If it becomes potentially relevant to the analysis, I will add dummy columns for all of the more specific string labels contained in the 'code' column.

IV. Exploratory Data Analysis

As noted above, the dataset has many features to explore. It contains a number of continuous variables documenting characteristics of each pitch once the pitcher has released it, including, among other things, each pitch's speed, location, spin/rotation, and movement. I will run some calculations to determine whether each of these variables is statistically significant to determining whether a pitch is a four-seam fastball.

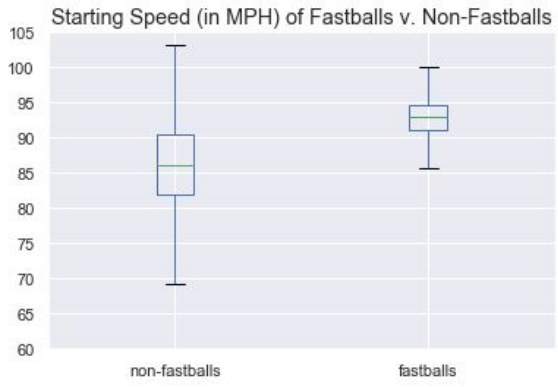
The dataset also provides some categorical variables regarding the game circumstances under which each pitch is thrown, including the inning of the game, the current score for each team, the number of outs, the current batting count, whether any runners are on base, the pitch sequence number of the current at-bat, and the dominant hand of both the pitcher and batter. I will do some initial exploration until whether any of these variables is statistically significant to whether a pitcher elects to throw a fastball. The goal of this project is to build a model that can accurately predict when a pitcher will throw a four-seam fastball based on these pre-pitch circumstances.

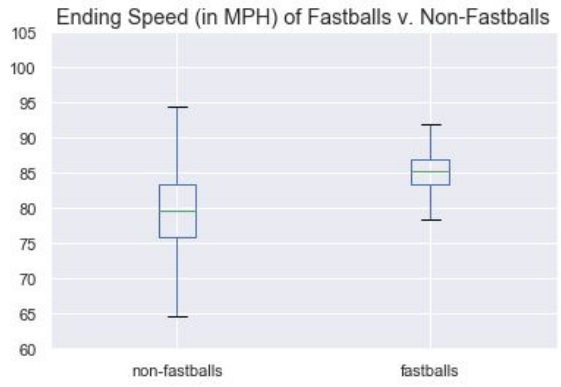
A. Continuous Variables

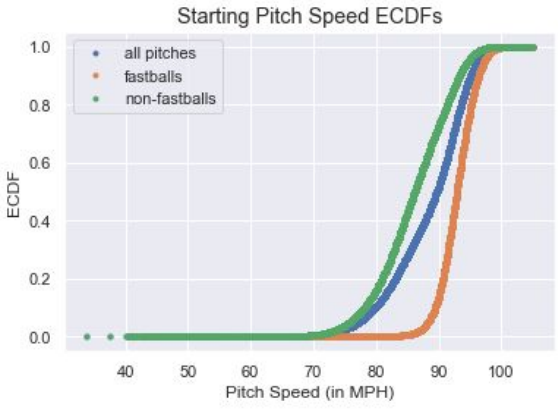
I will start by analyzing the various characteristics of each pitch once the pitcher has thrown it, which are continuous variables. For each of these variables, I calculated the mean, median, and std for the entire dataset as well as for each of the fastball and non-fastball datasets for an initial comparison. I also calculated the Pearson correlation coefficient between each variable and the fastball variable. I then ran permutation and bootstrap p-tests to confirm whether each variable is statistically significant to whether a pitch is a fastball.

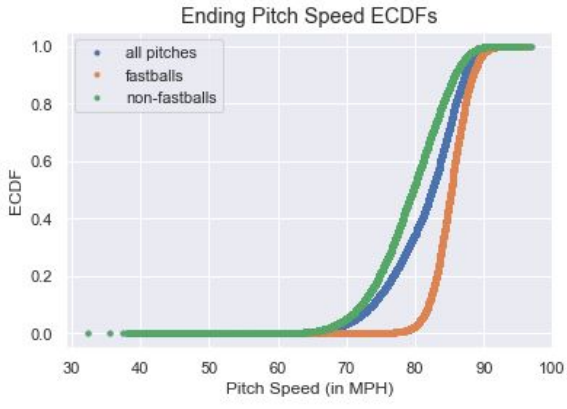
Pitch Speed

The dataset has two variables tracking pitch speed. The variable 'start_speed' provides the miles per hour (MPH) of the pitch at the time it leaves the pitcher's hand. The variable 'end_speed' provides the MPH of the pitch at the time it crosses home plate. Here are the results of my initial analysis:









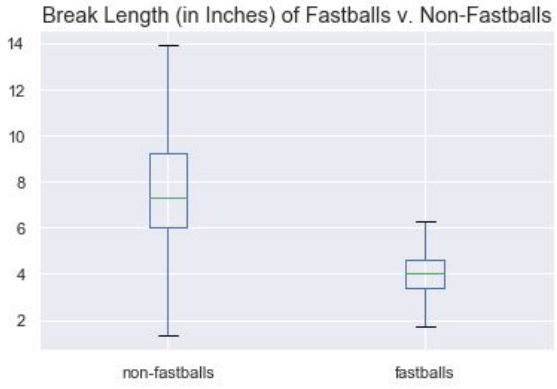
The graphs above demonstrate that pitch speed appears to be quite significant to whether a pitch is a four-seam fastball, which of course makes logical sense. Both the

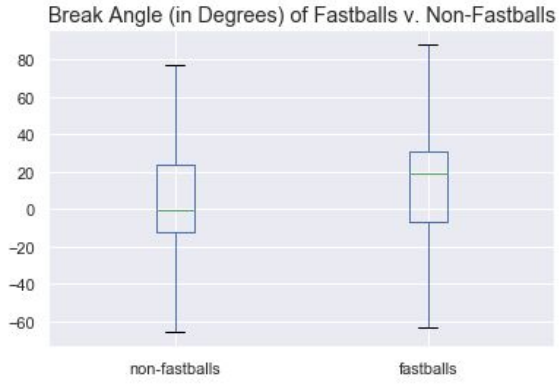
box plots and the ECDF plots indicate that fastballs have a higher average velocity than non-fastballs.

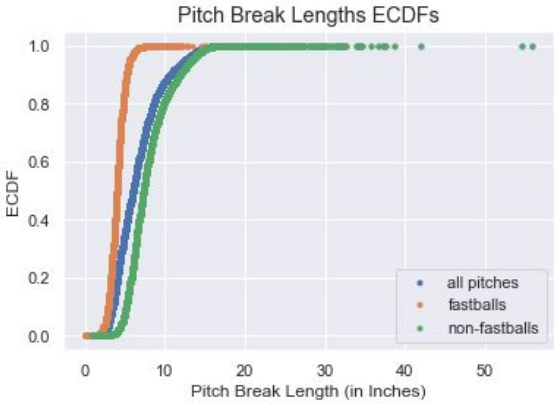
Indeed, the calculations confirmed this conclusion. The mean values of fastballs are a starting speed of 92.86 MPH and ending speed of 85.17 MPH, in comparison to a starting speed of 88.38 MPH and ending speed of 81.36 for the entire dataset. The Pearson coefficient between the fastball/start_speed is 0.55 and fastball/end_speed is 0.53, and running both permutation and bootstrap p-tests on the difference in means with both variables using sample sizes of 1000 generated p-values of 0 for each test.

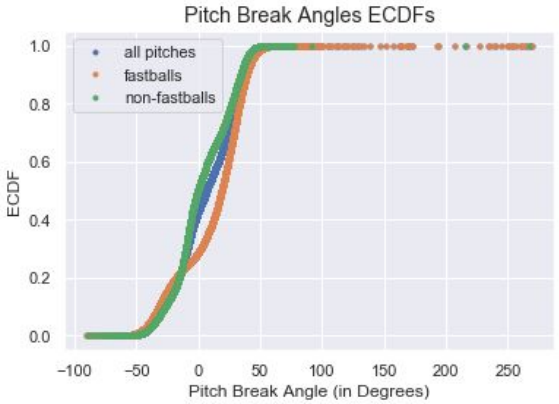
Pitch Movement

The dataset has two variables tracking the pitch movement from the time it is thrown until the time it reaches the plate. The 'break_length' variable measures the amount of movement (in inches) of the pitch. The 'break_angle' variable measures the angle at which each pitch moves as it approaches the plate. Here are the results of the analysis of these variables:





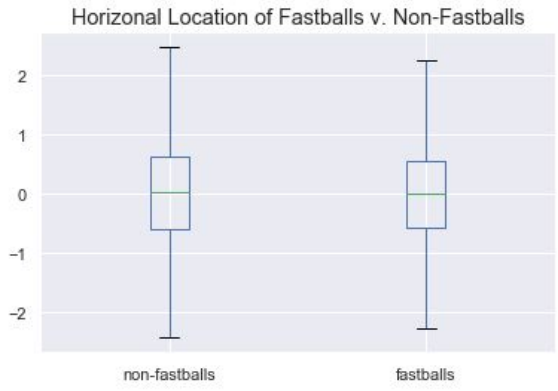


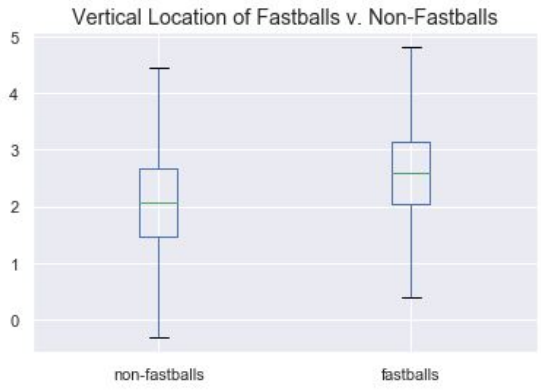


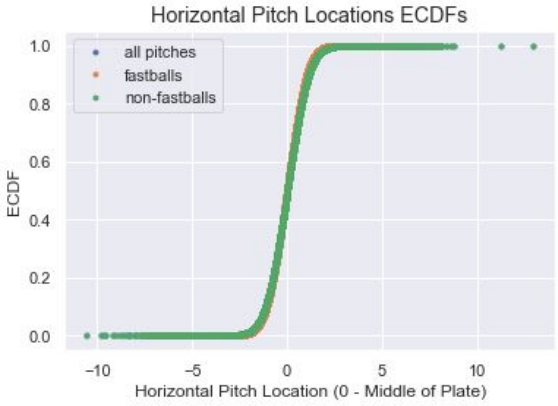
The graphs above show that both the pitch break length and angle appears to be statistically different for fastballs than for the entire dataset. To confirm, the mean values of fastballs are a break length of 4.01 inches and break angle of 11.08, in comparison to a break length of 6.52 and break angle of 5.85 for the entire dataset. The Pearson coefficient between the fastball/break_length is -0.65 and fastball/break_angle is 0.16, and running both permutation and bootstrap p-tests on the difference in means with both variables using sample sizes of 1000 generated p-values of 0 for each test.

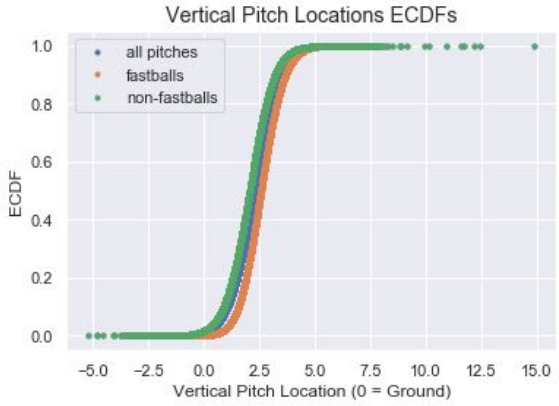
Pitch Location

The dataset has two variables that track the vertical and horizontal location of the pitch when it crosses home plate. The variable 'px' tracks the horizontal pitch location, with a value of 0 representing the middle of the plate. The variable 'pz' tracks the vertical pitch location, with a value of 0 representing ground. Here is the result of the analysis of pitch location:





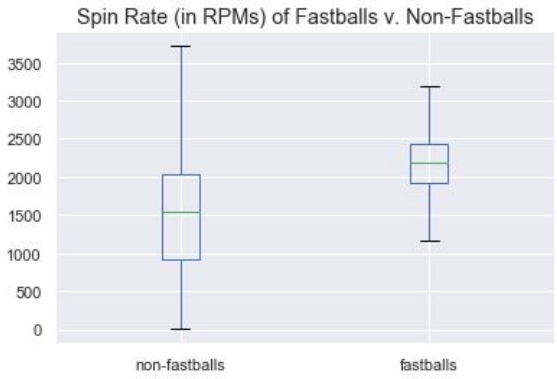


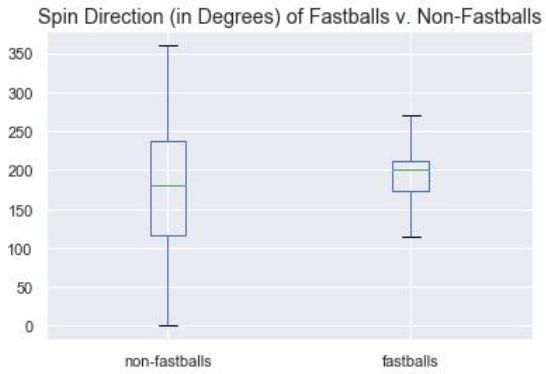


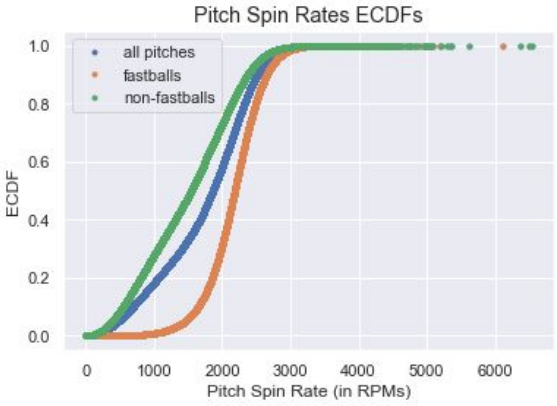
The graphs above demonstrate that the vertical pitch location for fastballs appears to be slightly higher than average, whereas the horizontal location is unclear. Investigating further, the mean values of fastballs are a horizontal pitch location of -0.0085 and vertical location of 2.6064, in comparison to a horizontal pitch location of 0.0066 and vertical pitch location of 2.2549 for the entire dataset. The Pearson coefficient between the fastball/horizontal location is -0.013 and fastball/vertical location is 0.276, and running both permutation and bootstrap p-tests on the difference in means with both variables using sample sizes of 1000 generated p-values of 0 for each test. Therefore, I can comfortably conclude that both variables are statistically significant, though the horizontal location is the less significant of the two variables.

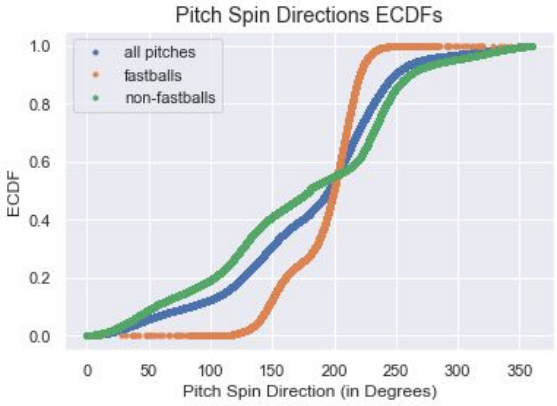
Pitch Spin

The dataset has two variables that track the rotation of a pitch once it is thrown. The variable 'spin_rate' tracks the rotations per minute (RPM) of each pitch. The variable 'spin_dir' tracks the angle (in degrees) that each pitch is spinning. Here is the result of the analysis of pitch spin:





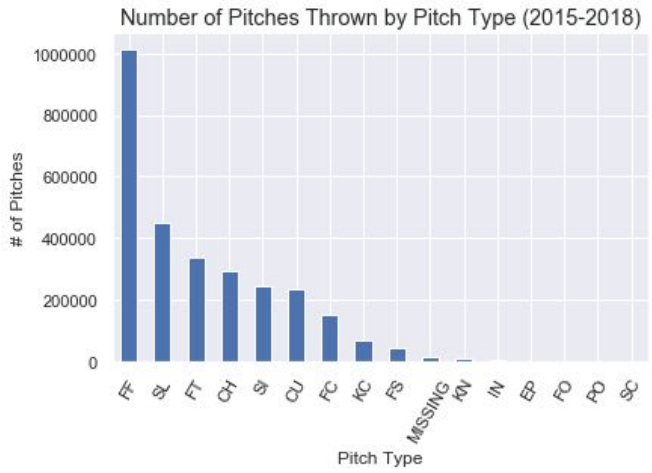


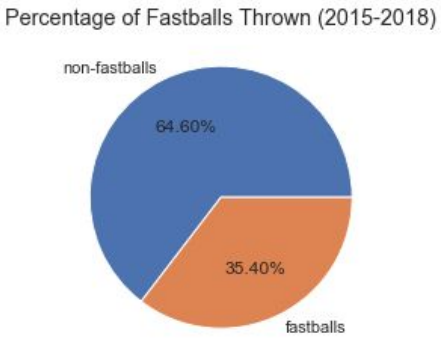


The graphs above demonstrate that both variables appear to be significant. Fastballs appear to spin at a much higher rate and at a much more centralized angle. Investigating further, the mean values of fastballs are a spin rate of 2161.10 RPMs and spin direction of 191.58 degrees, in comparison to a spin rate of 1731.17 RPMs and spin direction of 180.23 degrees for the entire dataset. The Pearson coefficient between the fastball/spin rate is 0.467 and fastball/spin direction is 0.125, and running both permutation and bootstrap p-tests on the difference in means with both variables using sample sizes of 1000 generated p-values of 0 for each test. Therefore, I can comfortably conclude that both variables are statistically significant.

B. Pre-Pitch Categorical Variables

I will analyze the significance of the pre-pitch circumstances by calculating the percentage of four-seam fastballs that are thrown when each variable is present. In order to determine the significance of those percentages, we will need to compare them to the default percentage of how often a pitcher throws a four-seam fastball on average:

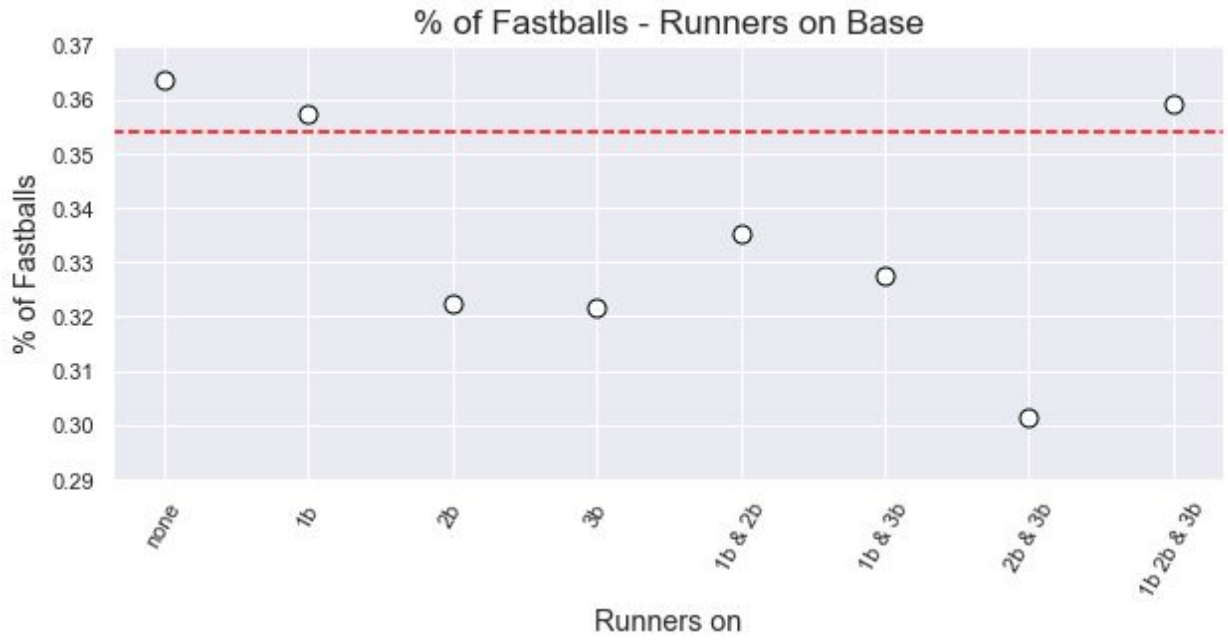




As we can see, the default likelihood that a pitcher will throw a four-seam fastball knowing nothing about the pre-pitch circumstances is 35.40%. Therefore, I can analyze how statistically significant each categorical variable is by comparing the frequency under which fastballs are thrown when that variable is present to this default 35.40%. I will conclude that a variable is significant only if deviates substantially from this default percentage.

Runners on Base

One categorical variable that may influence a pitcher's fastball usage is whether there are any runners on base. Given that fastballs are generally considered the easiest pitch for a batter to hit, it may make sense that a pitcher throws his fastball more often when there are less runners on base (so the resulting hits score less runs). Alternatively, I would expect a pitcher to throw more fastballs when he cannot afford to walk a batter (most importantly when the bases are loaded). Let's see if the data uncovers any trends:

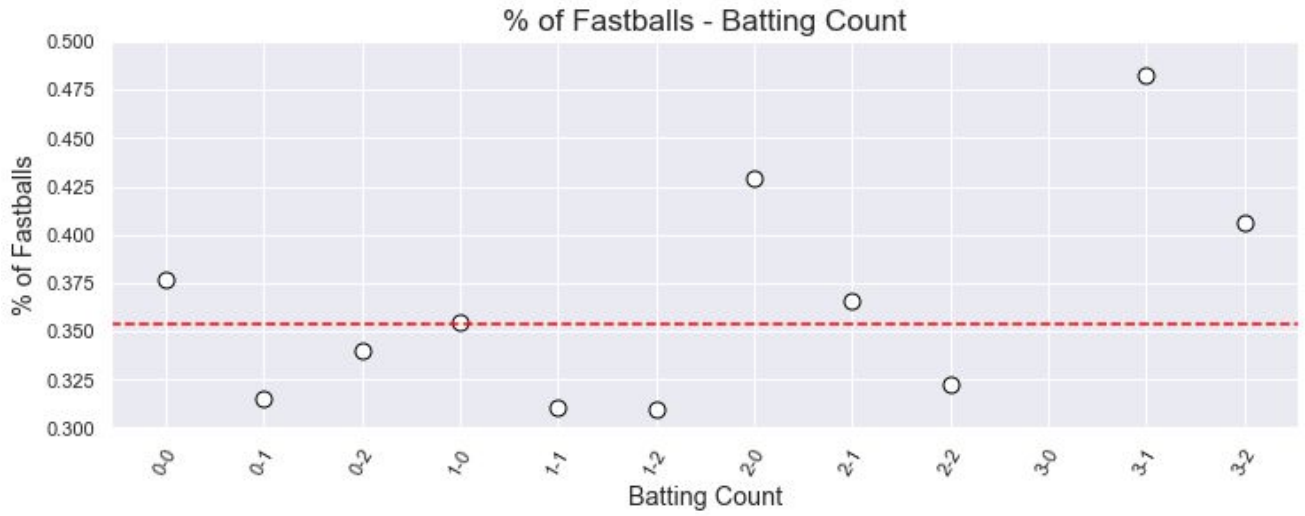


It appears that the most significant influence on fastball usage is whether there is at least one runner in scoring position (on second or third base). Fastball usage dips to 32-33% when at least one runner is in scoring position and all the way down to 30% when runners are on second and third (with first base open). The fastball usage spikes back up to 36% when the bases are loaded, however, likely because a walk in those circumstances results in a run scored.

Batting Count

Another categorical variable that may be correlative to fastball usage is the current batting count on the hitter. I surmise that a pitcher is more likely to throw his fastball when he is "behind" in the count (*i.e.*, more balls than strikes) as he is trying to avoid walking the batter. Conversely, I would expect less fastballs when he is "ahead" in the

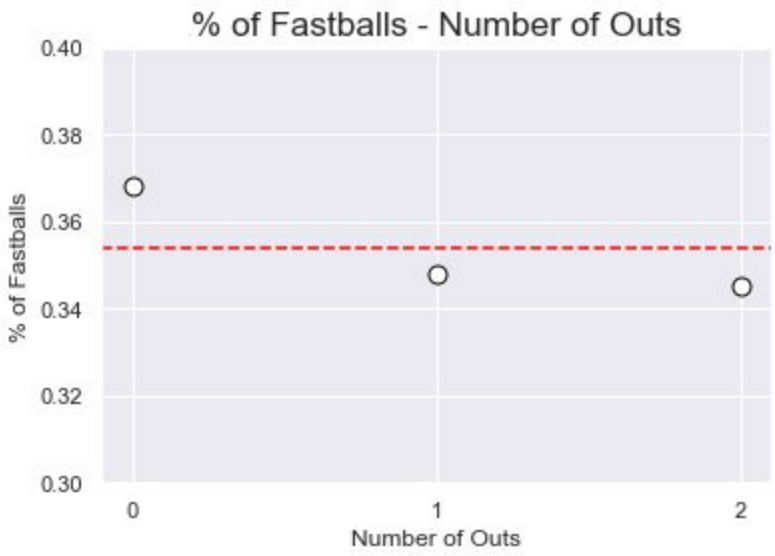
count (*i.e.*, more strikes than balls) and has more freedom to miss the strike zone. Here is the data:



As expected, the batting count appears to be strongly correlated to fastballs. A pitcher is much more likely to throw his fastball when he is behind in the count and much less likely to when he is ahead in the count. Interestingly, however, a pitcher is just over 2% more likely to throw a fastball on the first pitch. The free swingers will be happy to hear that!

Number of Outs

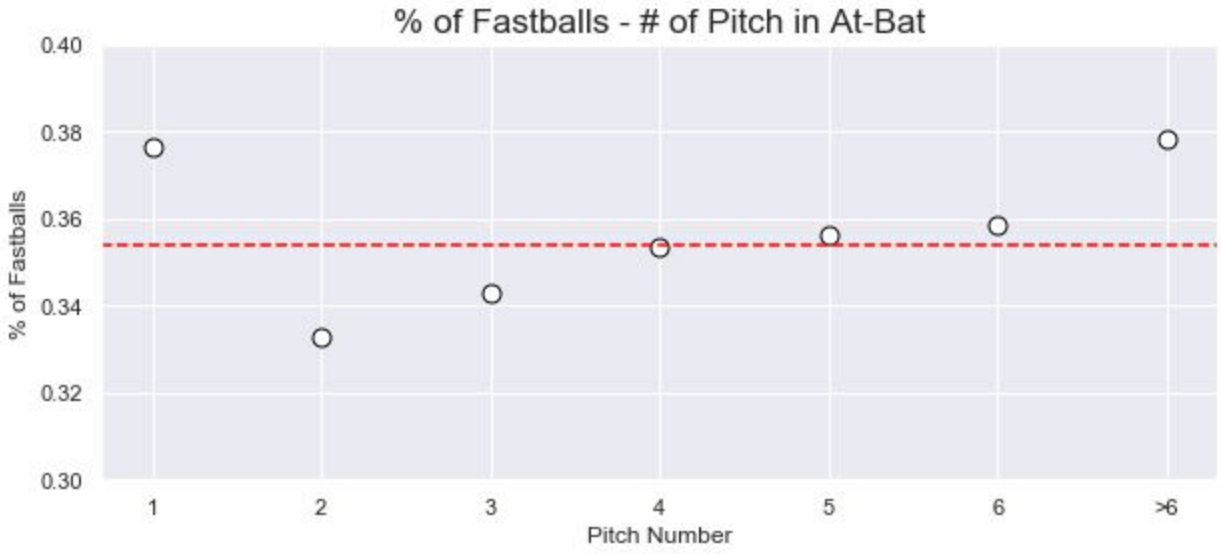
Another categorical variable to analyze is the number of outs in the inning. I suspect that this will be less correlative than some of the other variables:



On the whole, it appears that pitchers throw more fastballs than normal with no outs and less fastballs than normal with one or two outs. The fastball usage frequency does not appear to vary materially between one and two outs.

Pitch Sequence

Another interesting categorical variable in the dataset is the pitch number for each at-bat. I would speculate that a pitcher is more likely to throw a fastball on the first pitch as well as late in the batting count to avoid a walk. Let's check it out:

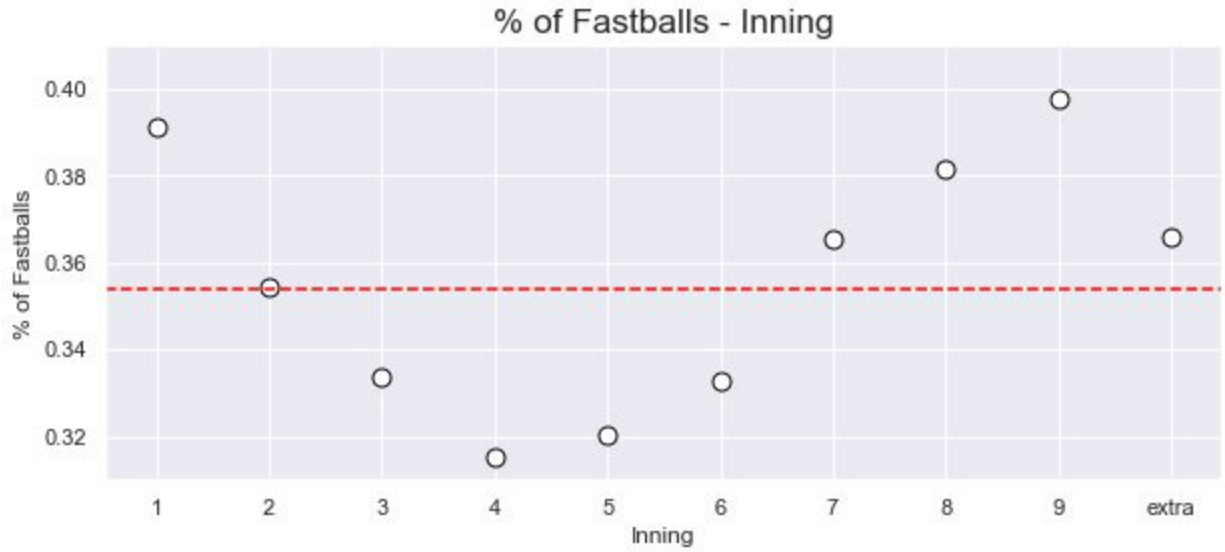


The data appears to validate this hypothesis. A pitcher throws his fastball approximately 2% more often on the first pitch and less often than average on the second and third pitches. Pitches four through six are approximately at the default average of 35.4% (though increasing slightly at each subsequent pitch). The frequency then increases significantly on pitches seven and beyond, likely to avoid walking the batter.

This categorical dataset has the potential to be even more significant in a pitch-by-pitch analysis, as it may be that a significant indicator of fastball usage is what type of pitches have been previously thrown during the same at-bat.

Inning

The inning of the game could also be correlative to fastball usage. For example, early in the game the starting pitcher may elect to use more fastballs if he perceives that surrendering a home run is less significant than walking a batter and upping his pitch count. Fastball usage also could be influenced by relief pitchers entering the games in the later innings, as they tend to have less variety of pitch types in their arsenal and can throw harder given they will be throwing less pitches overall in the game. Here is the data:

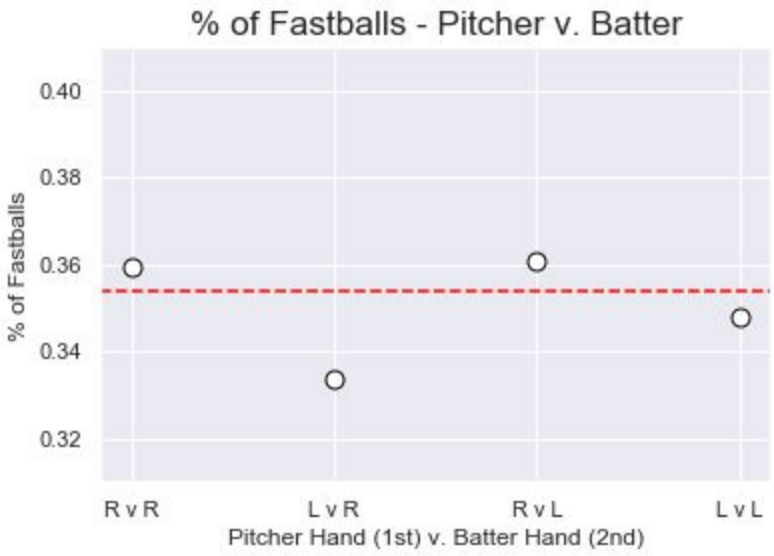


Interestingly, the data demonstrates a significant increase in fastball usage in the first inning, followed by a consistent dropoff in usage from the second inning through the sixth inning. Fastball usage then spikes dramatically in the later innings. The inning of the game therefore appears to be quite statistically significant (though this also could be tied to whether the starting pitcher has remained in the game). Perhaps the real

conclusion to draw is that starting pitchers throw their fastballs early and then rely on their breaking balls as the game progresses.

Pitcher/Batter Dominant Hand

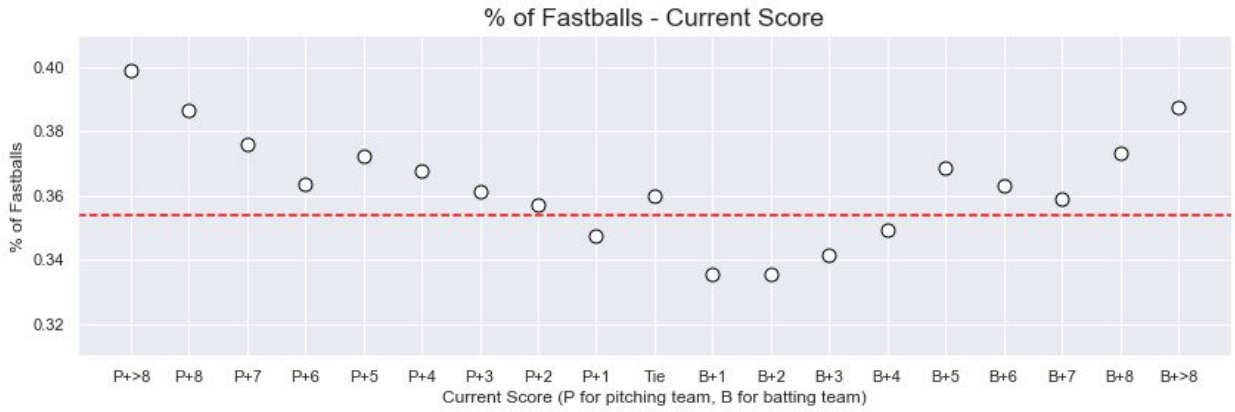
Another variable to consider is whether the pitcher and batter are right- or left-handed. This variable may bear out to be quite significant under certain circumstances, as pitchers often employ different strategies depending on whether the batter is right- or left-handed. Let's see what the numbers say:



The main takeaway from this analysis is that right-handed pitchers throw fastballs more often than average and left-handed pitchers throw their fastballs less often than average regardless of the batter. This seems like a reliable rule of thumb that could come in handy.

Game Score

My intuition tells me that the number of fastballs thrown will directly correlate with the size of the lead by either team. The reason for this hypothesis is straightforward - the more the outcome of the game becomes settled, the more likely a pitcher will be to just throw fastballs to prevent walks and get the game completed. Here is the data:



The results are quite interesting. Pitchers appear to throw fastballs in a tie game at a rate of 36%, which is only slightly higher than the default 35.40% rate. The percentage dips slightly when the pitcher has a one-run lead, but then gradually increases to well above the default rate as the pitcher's lead increases.

If the batting team has a lead of four or less runs, however, a pitcher throws fewer than the default fastball rate. Once the lead increases to five or more, then the pitcher changes course and throws above the default rate of fastballs. It therefore is fair to speculate that teams feel that a game is more or less out of reach once a deficit reaches five runs.

V. Conclusions

Based on the above analysis, I have drawn the following conclusions about the dataset:

- Fastballs very clearly have distinct characteristics once released by the pitcher. Every post-release continuous variable in the dataset analyzed is statistically significant to determining whether the pitch is a fastball;

- The batting count appears to be one of the most significant indicators of fastballs - batters should expect to face significantly more fastballs when ahead in the count (and a full count) and significantly less when behind in the count;

- Pitchers are less likely to throw fastballs with runners in scoring position (unless the bases are loaded);

- Pitchers throw more fastballs when there are no outs in an inning;

- Pitchers throw more than an average amount of fastballs in the first, seventh, eighth, ninth, and extra innings. Pitchers throw less than an average amount of fastballs in the third, fourth, fifth, and sixth innings;

- Right-handed pitchers throw significantly more fastballs than left-handed pitchers; and

- Batters should expect to face less than an average amount of fastballs when their team is winning by 4 or less runs. Pitchers will gradually throw more fastballs as the pitching team's lead increases or the batting team gets out to a 5 or more run lead

VI. Summary/Next Steps

Based on the above analysis, I am quite confident that I could build a model that could determine whether a pitch is a fastball with near-100% accuracy once the pitcher has thrown the ball. This ultimately is not the goal of this project, however, and a model of that type would have much less utility (and likely already exists).

Regarding the pre-pitch circumstances, my analysis demonstrated that a number of variables correlate with fastball likelihood percentage, though no one single variable appears to be outcome-determinative of when pitchers throw fastballs (even the batting count only moved the percentage by at most a little over 10% from the default fastball rate). The next step is to apply machine learning techniques to the dataset to see if I can build an accurate pre-pitch fastball prediction model.