Collin Hough
2023 Baseball Analytics Trainee Take Home
Part 2 A

Pitcher A should develop a cutter and Pitcher B should develop a circle changeup.

Logic

My goal with these choices was to try and accentuate the strengths of each player. I asked myself the following questions, “What pitch types do they each predominantly throw?”, “How often does each pitch type generate a swing and miss?”, “How often does each pitch type result in a hit?”, and “When a hit occurs, how many bases are given up on average?”. I felt that if I could answer all of these questions for each type of pitch the players throw, then I could determine how effective each pitch type is. To answer each question, I partitioned the pitch-by-pitch data of each player by the types of their pitches and calculated the following terms: swing and miss rate, usage rate, hit rate, bases per hit. Each can be defined as such:

$$Swing\ And\ Miss\ Rate = \frac{Amount\ of\ swinging\ strikes}{Amount\ of\ times\ pitch\ type\ was\ thrown}$$

$$Usage\ Rate = \frac{Total\ amount\ of\ times\ pitch\ type\ was\ thrown}{Total\ amount\ of\ pitches\ thrown}$$

$$Hit\ Rate = \frac{Total\ amount\ of\ hits}{Total\ amount\ of\ time\ pitch\ type\ was\ thrown}$$

$$Bases\ Per\ Hit = \frac{(1 * Number\ of\ Singles) + (2 * Number\ of\ Doubles) + (3 * Number\ of\ Triples) + (4 * Number\ of\ Home\ Runs)}{Number\ of\ hits\ from\ one\ pitch\ type}$$

The calculations for each player can be found at “*data/processed/part2/pitcher_{A/B}_stats.csv*” in the project. When analyzing each player’s statistics, I eliminated data from pitches with a usage rate less than 10%. After this, I manually analyzed the relationship between each players’ statistics to determine which pitch type each was most successful with. For the purposes of this analysis, I determined a pitch type’s success by comparing its swing and miss rate, usage rate, and hit rate to the other pitch types. I reserved the bases per hit statistic to be used as a tiebreaker if necessary. I found Player A had the most success throwing their fastball and Player B with their curveball and changeup. I decided to group Player B’s curveball and changeup together due to their lower usage rates in comparison to Player B’s fastball and their significantly higher swing and miss rates.

I believe the players should develop new pitches that are similar those they find the most success with, which is why Player A should develop a cutter and Player B a circle changeup.

Methodology

For predicting the swing and miss rates of each new pitch type, my methodology was to build a logistic regression model to predict pitch results, and then to create an example dataset of pitches to be used for predictions. Regarding the model, I chose logistic regression since each pitch results in a distinct classifier. I used the initial dataset of pitch-by-pitch data to create training and testing sets for the model. To create prediction datasets for each pitcher, I first partitioned the initial dataset by each pitcher's identifier. Since we will not have data on the new pitch types until next season, I used the data of the similar pitch types to make projections for the new ones. Because I used this preexisting data in the training and testing of the model, I had to create a new dataset based off it. To do this, I found 95% confidence intervals for each feature in the similar pitch data and created the new dataset by uniformly sampling each confidence interval. With the new dataset, my goal was to use the predicted pitch results to calculate swing and miss rate. Unfortunately, my model was ineffective, and I was unable to find a finite prediction for each new pitch type.