Reds Take-Home Assessment Writeup

Lucca Ferraz

Predicting Pitch Mixes Faced In Major League Baseball

Importance

Baseball analytics for years has focused on measuring batter performance in a vacuum, trying to estimate batter true talent through a variety of methods and metrics (WAR, BABIP, wOBA, etc.). However, in evaluating these hitters too little has been explored about the *context* in which different players bat. As we all know, not all pitches are created equal, and different hitters will face a different set of pitches over the course of the season. But can we predict these pitch "mixes"? It can be extremely useful to players and managers alike to know what a given batter can expect on a daily basis when he steps up to the plate. Hitters can spend their offseason working on their performance against certain pitches to remain one step ahead of opposing pitchers, and managers can make better lineup decisions knowing which pitches are likely to come against which batters. For this project, the goal was to predict 2024 pitch mixes seen by hitters who faced at least 1,000 pitches both from 2021-2023 and in 2024 (although no 2024 data was available to me), categorized into three pitch types: fastballs, breaking balls, and off-speed pitches.

Data

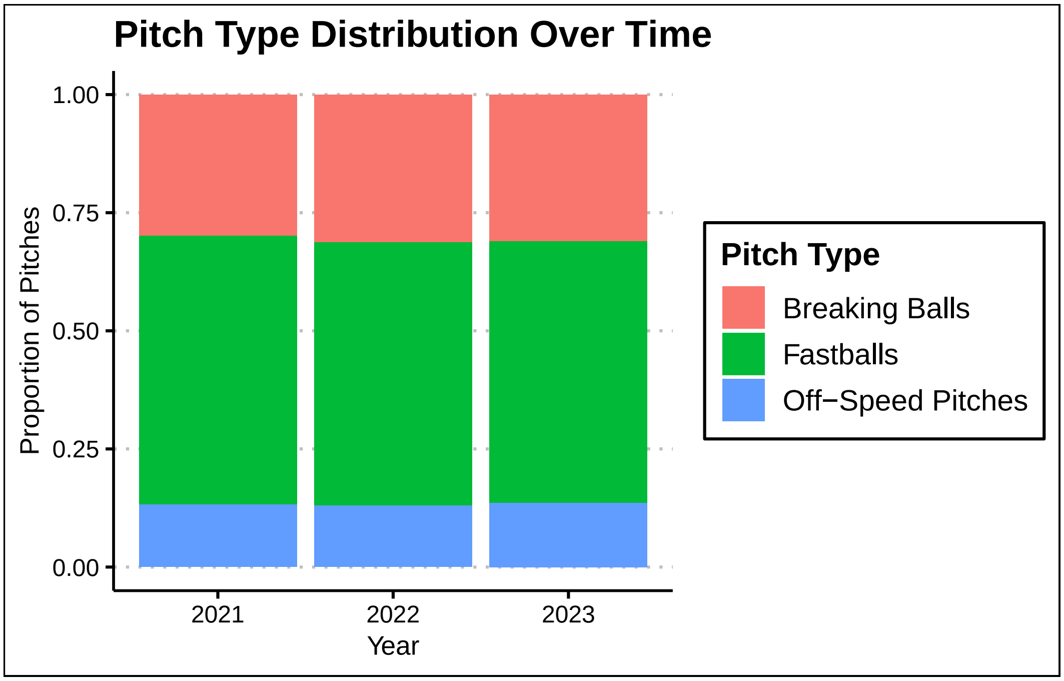
For this project, I had access to pitch-by-pitch data from the 2021-2023 seasons. Each pitch was labeled with the Statcast pitch type as well as other relevant information (batter ID, pitcher ID, bat side, throw side, game info, etc.) as well as the exact characteristics of the pitch and detailed results of the play (xWOBA, BABIP, launch speed/angle, ISO, etc.)

Formatting Data

In order to make predictions for 2024, I first had to transform the data provided from a pitch-by-pitch basis into a season-by-season basis. To accomplish this, I grouped the data by batter for each season, so I had statistics on each hitter for each season in the dataset. I grouped the different pitch types into the three target categories and calculated each batter's pitch mix faced in every given season. I also computed each batter's performance against each pitch type to provide context for how good different hitters were against different pitches.

Have Pitch Mixes Changed Over Time?

My first question when tackling this project was how pitch mixes across the MLB as a whole vary from year-to-year. If there is high volatility across seasons, predicting a future year's pitch mix using previous years would be extremely difficult and likely inaccurate. The plot below demonstrates the distributions of the three pitch types from 2021-2023:



As we can see above, the distribution of pitches across the MLB is relatively consistent from year-to-year, with fastballs always in the majority and off-speed pitches being utilized the least. The assumption of consitency across seasons is important for the modeling process later.

Training the Models

Since I had access to three season's worth of data, I decided to train the models using the 2021 pitch data to predict 2022 pitch mixes. Once I trained these models, I then applied them on the out-of-sample 2022 pitch data and evaluated their performance in predicting 2023 pitch mixes. I chose to create three model and compare their performance in order to choose the one that fits the data best. My first and baseline "model" was simply applying the previous year's pitch mix for each batter, e.g. predicting Player X to have the exact same pitch mix in 2022 as he did in 2021, and so on. Another model I tried and evaluated was a multinomial regression, since I needed three separate percentages to all add to 1. My final model was an XGBoost model, trained separately for each pitch type (fastball, breaking ball, off-speed) and then adjusted so the percentages all added to 1 (and no batter was projected to have over 100% pitches thrown to them). The multinomial and XGBoost models were trained using a combination of the previous year's pitch mix as well as the batter's performance against each type of pitch (strike rate, contact rate, average xWOBA, average BABIP, launch speed/angle zone).

Model Results

Below is a table displaying the performance of the three models:

model	train_rmse	test_rmse
Baseline	0.0659288	0.0585112
Multinomial	0.0476769	0.0644763
XGBoost	0.0011694	0.0580379

We can see that the XGBoost model had the best performance (lowest RMSE) on both the training and test data. As such, I chose this model to use for my final predictions.

Limitations

Although not a huge problem here given the 1,000 pitch threshold for the data provided, one potential hurdle in this process is a lack of data. For example, 4 players in the 2024 predictions dataset did not have any pitch data for the 2023 season. For these players, I estimated their pitch mix for 2024 using their 2022 numbers. A huge limitation on this work is the fact that teams get new data throughout the season. If a batter performed poorly against a certain pitch type in 2022, teams might try to throw him more of that pitch in 2023. However, if one month into the 2023 season that same batter is dominating his previous Achilles Heel pitch, opposing pitchers are going to adjust and throw that pitch to him less. As such, this model should be adjusted adn re-calculated semi-frequently to capture recent trends as the season goes on. A portion of the batter pitch mix faced could be due to game situational factors as well, which was not captured by my model. The reason for this is predicting situational factors for individual players in a new season(2024) would be too difficult and noisy with the data provided. Ultimately, the predictions attached provide a best estimate, but there is always room for error and they should be treated as estimates, not objective truths.