Modeling A Batter’s Pitch Mix

Model Description

The goal of this project is to predict what proportion of each general pitch types specific batters will see in 2024. General pitch types include Fastballs, Breaking Balls, and Off-speed pitches. The data used to create the model for predictions is pitch by pitch data from the 2021-2023 MLB seasons. A row of data contained contains characteristics about the pitch (including specific pitch type) and situation data (ex: runners on base). Certain rows were excluded from the training data. This included rows with a pitch type of Other, Pitch-out, or Null. A small number of rows had a batters count with either 4 balls or 3 strikes. These are invalid counts, and thus the rows were dropped from the training data. Then, each specific pitch type in the dataset was classified into the appropriate general pitch type. The mapping is shown below:

General Pitch Type	Specific Pitch Type
Fastball	4-Seam Fastball, Sinker, Cutter
Breaking-Ball	Slider, Knuckle Curve, Curveball, Sweeper, Slurve, Slow Curve, Screwball
Off-speed	Changeup, Split-Finger, Eephus, Knuckleball, Forkball

Since the goal is to predict the distribution of pitch types a player sees in 2024, a Bayesian approach was taken to construct the predictive model. The general pitch type from the data can be viewed as a Multinomial random variable where each pitch type has an associated probability of being seen. The Dirichlet distribution is a conjugate prior for a multinomial likelihood. Leveraging this, a Dirichlet posterior distribution can be obtained to model the probability that a pitcher sees a given pitch type in 2024.

The Bayesian approach above is reliant on the response variable being independently and identically distributed. Without conditioning on other variables in the data, this is not true. To achieve independence, the data was first conditioned on the batter’s count at the time of the pitch. This is common in pitch modeling and is intuitive, too. Pitcher’s throw different pitches based on the count at the time of the pitch. The year of the pitch was also conditioned on since the overall distribution of pitch mix changes from year to year.

For each combination of Batter, Year, Balls, and Strikes, a posterior distribution was created using the observed data and a weakly informative Dirichlet prior. The parameters for the weakly informative prior were scaled integers from the population distribution of pitch types given a value for Year, Balls, and Strikes.

This gives a multiple of posterior distributions for each player, year, and batter’s count combination. To combine these distributions into a final probability distribution, the conditional variables need to be integrated out of the conditional posterior distribution. To integrate out the conditioning on batter’s count, the process above was repeated for the distribution of batter’s counts that a batter sees in a given season using a Multinomial likelihood and Dirichlet prior. This results in two sets of posterior distributions—one for each of $\mathbb{P}(\theta \mid Data, Year, Balls, Strikes)$ and $\mathbb{P}(Balls, Strikes \mid Data, Year)$.

To obtain a single posterior distribution for each player, random bootstrap samples were taken from each set of posterior distributions for a given player. Using the values sampled from $\mathbb{P}(Balls, Strikes \mid Data, Year)$, a weighted sum was taken on each pitch type from the $\mathbb{P}(\theta \mid Data, Year, Balls, Strikes)$ samples to obtain $\mathbb{P}(\theta \mid Data, Year)$. For a player that has observations in all years, this results in 9 normal bootstrap distributions —one for each pitch type in each year. To combine the bootstrap distributions for each pitch type across all years, the bootstrapped observation can simply added using a weighted sum where the weights are the true proportion of pitches that a player saw in a given year. This results in 3 normal distributions for each player, associated to one of the 3 general pitch types. The mean of the distributions becomes the predicted percentage of each pitch type that a player sees in 2024.

Model Limitations

With more time, this model can be improved. As was found in exploratory analysis of the data, more recent years have more predictive power on a hitter's future pitch mix. Instead of using the observed proportion of pitches for a player in a given year, a weights that weight 2023 more than 2022 and 2021 should be used. However, since not every player has data for all 3 years, this was not done.

Although conditioning on the batter's count is essential to achieve independence, it likely results in overfitting because we're trying to put a distribution on each row of the dataset.

An alternative modeling technique that would likely suit this project well is Bayesian Hierarchical Regression. A Bayesian Hierarchical model would be able to include a population level effect on the batter's count as well as a specific random effect for the effect of batter's count given a player. This makes the model more generalizable, especially to a new player in 2024. This is not true of the model presented as the distributions are entirely player specific. This was briefly attempted, but the training time of the model was too significant to overcome.

Other considerations could include a clustering element prior to modeling to identify different types of batters in the data. This would help make the model more generalizable to both new batters in 2024 and batters with little data from 2021-2023.