Batter Pitch Mix Modeling Technical Report

Introduction

The goal of this project is to predict the pitch mixes that batters will face in the 2024 season, specifically broken down into fastballs, breaking balls, and off-speed pitches. To answer this question, I was provided pitch-by-pitch Statcast data containing information about every pitch from the 2021, 2022, and 2023 seasons. Since I did not have access to data from the 2024 season to test any model I built, I decided to use the data from the 2021 and 2022 seasons to predict the outcomes for the 2023 season. Once the model was validated, I applied the same model to the data from the 2022 and 2023 seasons to predict the results for 2024.

Data Preparation

Before training the model, I handled data cleaning by filtering out pitches labeled as "Pitch Outs" or "Other". With these pitches excluded, I categorized each pitch type into one of the designated categories: fastballs, breaking balls, and off-speed pitches. The mapping was as follows:

- Fastballs: 4-Seam Fastball, Cutter, Sinker
- Breaking balls: Slow Curve, Curveball, Forkball, Knuckle Curve, Screwball, Slider, Sweeper, Slurve
- Off-speed: Change-up, Eephus, Split-Finger, Knuckleball,

Once the pitches were mapped to their respective categories, I calculated rolling percentages for each batter that measured how frequently they saw each type of pitch. After experimenting with different rolling window sizes, I determined that a window of 3,000 pitches provided the most accurate testing results. This window size also made intuitive sense, as it captures how pitchers were approaching a batter in recent at-bats while also maintaining a large sample size. Using rolling percentages also allows us to look at how a batter's pitch mix changed over the course of the season, particularly for batters who faced well over 3,000 pitches, as their percentages could shift meaningfully as the rolling window moved.

Additionally, I experimented with a time decay method that weighted recent pitches more heavily than older ones, but I found that the rolling percentages method provided better results.

Model Training

Using these rolling percentages, I trained a compositional regression model to predict batters' pitch mixes in 2023 based on their pitch mixes from 2021 and 2022. I implemented an 80-20 train-test split, reserving 20% of the data for testing. Notably, the model was trained only using players who faced at least 1,000 pitches in 2023, as that subset was the focus of the analysis. The model performed quite well, with mean absolute errors of around 2% within each category.

Model Shortcomings

One possible limitation of the model is that I wasn't able to fully account for the specific game situations that a player found himself in, which could influence the pitch mix that he faced. For example, a player may be more likely to see a fastball if they are batting with the bases loaded. However, because we are dealing with players who faced a large sample size of pitches across various situations, I anticipate that this effect would likely balance out over the course of a full season. Additionally, some players may just be more likely to bat in situations where they are more likely to face certain pitches, which could remain consistent across seasons.

It's also possible that some batters faced pitchers that were more likely to throw certain types of pitches, skewing the pitch mixes they faced due to factors beyond their control.

I also considered using a more advanced modeling technique that could have captured non-linear relationships between the variables. However, given the small sample size of players that I was working with, I chose to avoid training a more complex model to reduce the risk of overfitting. While my model performed quite well, it's possible that a different machine learning model would have been able to account for more complex relationships between past and future pitch mix percentages and potentially improved performance.