Technical Report: Predicting 2024 Pitch Mix Proportions

1. Objective

The goal of this analysis was to predict the proportions of pitch types that each batter will see during the 2024 MLB season based on historical data for the 2021-2023 MLB seasons. This data set provided pitch-level data for all players that faced 1,000 pitches from the mentioned seasons and predictions were made for said players in the 2024 season.

2. Data Summary

The data set includes over 1,000,000 pitch-level observations for 314 players from the 2021-2023 MLB seasons. I then created a subset containing only player name, batter id, pitch type, game year, and PitchCat. PitchCat categorized all pitch types into one of three categories:

* Fastballs
* Breaking Balls
* Off-Speed

Each players historical pitch mix was used for estimating their 2024 pitch mix.

3. Modeling Approach

I decided to apply different methods for veteran players and rookies because each required a different approach:

* Rookies (first year players with limited data)

For rookies, there was limited sample size so I decided to use a random forest model. This machine learning technique is good for non-linear relationships and smaller datasets.

Random forest combines the forecasts from numerous decision trees, which helps prevent overfitting and provides resilience to noisy data..

* Veterans (players with historical data)

For veterans I chose a moving average model. This approach captures the historical pitch mix trends by calculating a smoothed average over the last three years. The moving averages are effective at showing gradual shifts in pitch mix tendencies. For veteran players this model is well suited to make reliable predictions.

4. Results

The predictions were made for each player’s 2024 season for estimated proportions of fastballs, breaking balls, and off-speed

* Mean Predicted Proportions for the Population:
  + - Fastballs: 57.67%
    - Breaking Balls: 30.72%
    - Off-Speed: 11.58%

The full predictions can be found in the attached predictions.csv

5. Limitations

* Limited Data for Rookies: Due to limited observations for rookies, the random forest model may not capture full variability for the 2024 season.
* Changes in Strategies: The model assumes that past trends will continue into 2024. Any opposing strategies or significant changes are not accounted for.
* Context: Since we’re only using historical data, there’s no contextual factors such as pitcher tendencies or batter improvement over the season.

6. Conclusion

By utilizing a moving average model for veterans and a random forest model for rookies, I’ve provided estimates for the pitch mix that batters are likely to face in the 2024 season. These predictions can help inform coaching staffs, coaching decisions, matchups, and game strategies.