MLB Pitch Mix Prediction Analysis Report

- Aditya Kingrani

1.Introduction:

This report presents the analysis and results of predicting pitch mixes for MLB batters in the 2024 season, based on historical data from 2021-2023. The project aims to forecast the proportions of fastballs (FB), breaking balls (BB), and off-speed pitches (OS) that each batter is likely to face.

2.Background:

Pitch mix prediction is crucial for strategic decision-making in baseball. Understanding the likely pitch types a batter will face can inform game preparation, player development, and scouting strategies.

3.Purpose:

The primary objective is to develop a model that accurately predicts the pitch mix for MLB batters in the upcoming season, providing valuable insights for team strategy and player preparation.

4.Methods:

4.1. Data Preprocessing:

- Mapped pitch types to categories (FB, BB, OS)
- Created composite features (e.g., count, runners on base)
- Encoded categorical variables using LabelEncoder

4.2. Feature Engineering:

- Aggregated data by batter
- Calculated pitch type percentages for each batter

4.3. Model Development:

- Utilized Random Forest Regressor
- Split data into 80% training and 20% testing sets
- Used 100 trees in the Random Forest ensemble

4.4 Evaluation Metrics:

- R-squared (R^2) score for each pitch type and overall
- Mean Squared Error (MSE) for each pitch type

5.Results:

5.1. Model Performance:

- Overall R^2 score: 0.4743 (47.43% accuracy)
- R^2 scores by pitch type:
  i. Fastballs: 0.3013
  ii. Breaking Balls: 0.4818
  iii. Off-speed: 0.6399
- Overall MSE score: 0.0008
- MSE scores by pitch type:
  i. Fastballs: 0.0010
  ii. Breaking Balls: 0.0007
  iii. Off-speed: 0.0006

5.2. Feature Importance (Top 5):

i. PLATE_X (37.29%)
ii. PLATE_Z (19.69%)
iii. LAUNCH_SPEED (9.74%)
iv. BALLS (6.70%)
v. OUTS_WHEN_UP (6.24%)

5.3. Predictions:

- Generated for 314 batters for the 2024 season.
- Proportions for FB, BB, and OS pitches provided for each batter.

Analysis:
The Random Forest model demonstrates moderate predictive power, with varying performance across pitch types. Off-speed pitches are predicted with the highest accuracy, while fastballs show the lowest predictive accuracy. The model's overall accuracy of 47.43% suggests room for improvement.

Limitations:
1. Moderate overall accuracy (47.43%) indicates potential for model enhancement.
2. The model doesn't account for potential changes in player performance or team strategies in 2024.
3. External factors like weather conditions or ballpark effects are not considered.
4. The broad categorization of pitches into three types may oversimplify complex pitching strategies.

Recommendations:
1. Explore advanced machine learning techniques (e.g., Gradient Boosting, Neural Networks) to improve model accuracy.
2. Incorporate additional features, such as pitcher-specific data or more granular pitch classifications.
3. Implement time-series analysis to capture trends and seasonality in pitch mix data.
4. Consider ensemble methods to combine predictions from multiple models.
5. Gather and integrate more recent data to capture latest trends in pitching strategies.

Summary:
This analysis provides a foundation for predicting pitch mixes in MLB, with moderate accuracy. While it offers insights into factors influencing pitch selection, there's substantial room for improvement. The identified limitations suggest several avenues for enhancing the model's predictive power and reliability for future seasons.

References:

1. Scikit-learn documentation for Random Forest Regressor.
2. MLB Statcast data documentation.
3. Previous studies on pitch prediction in baseball.