Predicting Pitch Types for Each Batter: A Machine Learning Approach

Introduction

Hank Aaron, the legendary Hall of Fame outfielder and former Home Run King, once stated, “Guessing what the pitcher is going to throw is 80% of being a successful hitter; the other 20% is just execution.” This highlights the crucial role that anticipating pitch types plays in a batter's success. In baseball, accurately predicting whether a batter will face a Fastball (FB), Breaking Ball (BB), or Off-Speed (OS) pitch can significantly impact in-game decision-making. The objective of this analysis is to develop a predictive model that forecasts pitch categories for each batter in the 2024 season, using data from previous seasons to inform batters about their expected pitch distribution. Specifically, I aim to provide batter-specific probabilities for facing different pitch types based on historical data from 2021 to 2023.

Data

The dataset was compiled from the 2021 to 2023 MLB seasons and includes data on 314 MLB players observed multiple times. It contains over one million rows and 56 columns, but due to the study's objectives, I selected only relevant columns for analysis. Observations in the dataset are not independent, as all players were studied multiple times, leading to an unbalanced dataset with heterogeneous variances.

Several features, such as post-pitch events (e.g., BB_TYPE, HIT_LOCATION, LAUNCH_SPEED), game scores, and win expectancy variables, were excluded from the analysis. Although the presence of base runners may influence pitch selection—especially for off-speed pitches to prevent stolen bases—many of these variables had significant missing values (ON_1B, ON_2B, and ON_3B had missing rates of 69.15%, 81.21%, and 90.71%, respectively). Imputing these values could introduce uncertainty and potentially lead to incorrect assumptions, negatively affecting model performance. Consequently, I retained only those features deemed relevant to pitch type prediction. Additionally, pitch types were mapped to their respective categories, with fastballs being the most frequently occurring.

Approaches to Prediction

The Naïve Approach

The naïve prediction method calculates the average occurrence of each pitch type—Fastballs, Breaking Balls, and Off-Speed pitches—for every batter using historical data from 2021 to 2023. By mapping the pitch types to their respective categories and creating indicators, I computed the mean probabilities for each batter facing each type of pitch. This straightforward approach serves as a baseline for evaluating more complex models. However, it assumes that past averages will predict future outcomes, which may not always hold true, and overlooks contextual factors that could influence pitch selection.

Modeling Approach

The simplest method for addressing this problem is multinomial logistic regression, which models the probability of each pitch type (FB, BB, OS) based on game features such as the pitcher's handedness, the batter's stance, and the count (balls and strikes). However, logistic regression assumes that all observations are independent. Given that each batter faces multiple pitches in different contexts, this assumption is violated. Consequently, I explored more advanced models capable of handling repeated measures and within-batter correlations.

To account for these repeated measures, I considered linear mixed-effects models (LME), which allow for random effects that capture batter-specific variability. These models are robust to unbalanced data and heterogeneous variances, as different batters exhibit varying patterns of pitch selection. I implemented a random intercept for each batter (BATTER_ID) to account for these correlations. However, the computational cost of fitting mixed-effects models on a large dataset with many predictors proved to be a significant challenge. Thus, I turned to modern machine learning methods, which offer greater computational efficiency and flexibility.

Adopting Machine Learning Approaches

Due to the computational limitations of mixed-effects models, I shifted focus to extreme gradient boosting (XGBoost) and random forest classifications—two powerful machine learning algorithms well-suited for handling large datasets and capturing complex interactions between features without the need for explicit random effects.

Random Forest is an ensemble learning method primarily used for classification and regression tasks. It builds multiple decision trees during training and combines their predictions to enhance accuracy and control overfitting. Each tree is trained on a random subset of the data, leading to diverse predictions.

In contrast, XGBoost builds trees sequentially, where each tree learns from the errors of its predecessors. It incorporates L1 (Lasso) and L2 (Ridge) regularization techniques to prevent overfitting, making it more robust when working with complex datasets and capable of handling missing data effectively.

Modeling Steps

1. I trained two models (Random Forest and XGBoost) on the 2021/2022 data and tested them on the 2023 data.
2. Both models produced satisfactory classification reports, with accuracy scores of 0.67 for XGBoost and 0.65 for Random Forest. Other metrics, such as precision and F1-score, were also evaluated.
3. To improve model performance, I applied SMOTE (Synthetic Minority Over-sampling Technique) to ensure a balanced representation of pitch types during model training. However, this did not yield better metrics than the initial approach.
4. I selected the best model (XGBoost) and used it to make predictions based on all data from 2021 to 2023 to forecast for 2024.
5. Finally, I aggregated the probabilities for each batter by calculating the mean across the three pitch categories and created an interactive dashboard for each player.

Limitations

1. **Naïve Prediction Method**: This method assumes that past performance will directly predict future outcomes, ignoring potential changes in player performance and situational factors.
2. **Model Limitations**: While Random Forest and XGBoost have their advantages, they may not fully capture the nuances of pitch selection influenced by game context or specific batter-pitcher matchups. Additionally, they might not adequately address the correlations and repeated measures inherent in the dataset.
3. **Data Quality**: The dataset contains biases due to class imbalances, which, although mitigated by SMOTE and advanced modeling techniques, still affect model performance.
4. **Feature Selection**: Excluding certain features (e.g., base runners due to significant missing values) and other factors not captured in the dataset may overlook critical context influencing pitch selection.

This report presents a comprehensive analysis of pitch mix predictions for the upcoming 2024 MLB season, utilizing advanced modeling techniques to enhance predictive accuracy. The XGBoost model has been identified as the most effective method, providing actionable insights into pitch mix strategies for individual batters. Future work should focus on refining these models and exploring additional features that may further improve prediction accuracy.