Reds Assessment: Predicting Pitch Types Using Random Forest

Deven Chatterjea

October 2024

Overview

This report focuses on predicting the pitch mix (Fastball, Breaking Ball, Off-Speed) that batters will face in the 2024 MLB season. A Random Forest classification model was developed, trained using data from the 2021-2023 MLB seasons. The selected features for the model include game context, player tendencies, fielding alignment, and hit quality. This assessment provides insight into how these features influence pitch selection.

Data Preparation

The raw data includes pitch-by-pitch records for several seasons. To ensure meaningful predictions, the data was carefully cleaned and processed:

- **Pitch Categorization**: All pitch types were grouped into three categories: Fastballs (FB), Breaking Balls (BB), and Off-Speed (OS). This allowed for a simplified target variable. Pitch type 'PO' was excluded because it does not correspond to regular pitching strategies.
- **Feature Engineering**: A *hit quality score* was engineered to avoid multicollinearity using launch speed and launch angle. The score is calculated as:

$$\text{Hit Quality} = \text{Launch Speed} \times \max(0, \cos(\text{Launch Angle} - 20^\circ))$$

  This score emphasizes optimal hit conditions between 10° and 30°.
- **Handling Missing Data**: Missing values in the base runner information and other key variables were filled appropriately, ensuring no critical data gaps.
- **One-Hot Encoding**: Categorical variables, such as infield/outfield alignment and batter/pitcher handedness, were one-hot encoded to ensure compatibility with the Random Forest model.

Exploratory Data Analysis (EDA)

To guide feature selection and model choice, various aspects of the data were explored:

- **Pitch Category Distribution**: Fastballs comprised the majority of pitches thrown (approximately 70%), while Breaking Balls made up around 20%, and Off-Speed pitches were the rarest at 10%. This imbalance was expected and confirmed the need to handle class imbalance during model training.
- **Feature Correlations**: Preliminary analysis showed limited linear relationships between the features and the target pitch categories, justifying the choice of a non-linear model like Random Forest.

Modeling Approach

The prediction problem was modeled as a multiclass classification task, where the goal was to predict whether the next pitch would be a Fastball, Breaking Ball, or Off-Speed pitch. Given the complexity of strategies and the non-linear nature of the feature, Random Forest was chosen for the following reasons:

- **Non-linearity**: Random Forest models do not assume a linear relationship between input features and the target variable, making them suitable.
- **Handling Categorical Data**: The model easily processes categorical variables without extensive preprocessing, which is important given the diversity of baseball data.
- **Scalability**: With over a hundred-thousand rows of data, Random Forest efficiently handles large datasets using parallelism.

Model Performance and Evaluation

The model achieved the following results:

- **Training Accuracy**: 97%, indicating that the model performed very well on the training data.
- **Testing Accuracy**: 52%, suggesting significant overfitting, where the model is learning noise from the training data and struggling to generalize to new, unseen data.

The high training accuracy and lower testing accuracy suggest overfitting. The model captured the patterns in the training data, but these patterns did not generalize as well to the test data. This could be due to several factors, including:

- **Class Imbalance**: Fastballs dominate the dataset, which may cause the model to skew predictions toward Fastballs, making it difficult to predict the minority classes like Breaking Balls and Off-Speed pitches.
- **Feature Complexity**: Although the selected features provided information, certain features may introduce noise or redundancy. Similarly, the derived hit quality score could be a factor for introduced noise.

Limitations and Future Work

Although the Random Forest model showed promise in its ability to predict pitch categories, several limitations remain:

- **Overfitting**: The gap between training and testing accuracy indicates that further work is needed to improve generalization. Tuning hyperparameters such as `max_depth` and reducing the number of features might help address this issue.
- **Class Imbalance**: Further investigation into more robust methods for handling class imbalance (such as SMOTE or other resampling techniques) may improve predictions for minority pitch types.
- **Model Complexity**: While Random Forest provides interpretability in terms of feature importance, future work could involve experimenting with simpler models like Logistic Regression or more complex ones like Gradient Boosting or Neural Networks.