Hitter Evaluation Technical Report

Alex Rintamaa
10-12-24

Introduction

The dataset contains over 1.2 million plate appearances from MLB seasons 2021-2023, with 56 variables capturing different aspects of game situations. This large dataset enables robust model development. A Random Forest model was selected for this analysis, and the report details the rationale behind choosing it, the process of feature selection, and the model's performance, including its limitations.

EDA and Variable Selection

Due to the dataset's size, LASSO regression for feature selection proved computationally expensive, with processing times exceeding 10 hours. Instead, domain research highlighted key variables affecting pitch selection:

- PITCH_TYPE (type of pitch)
- BATTER_ID (Batter's ID)
- PITCHER_ID (Pitcher's ID)
- BALLS (Ball count)
- STRIKES (Strike count)
- INNING (Inning number)
- INNING_TOPBOT (Top or bottom of the inning)
- BAT_SIDE (Batter's handedness)
- THROW_SIDE (Pitcher's handedness)
- ZONE (Strike zone area)
- PLATE_X and PLATE_Z (Pitch location)

A scatter plot of horizontal and vertical pitch movement helped to categorize pitch types into three groups: Fastballs (Four-seam, Sinker), Breaking Balls (Slider, Cutter, Slurve, Screwball, Sweeper), and Off-speed (Changeup, Knuckleball, Forkball, Euphus, Other, Split-Finger, Slow Curve, Knuckle Curve, Curveball).

Model Selection and Predicting

Random Forest was selected for its ability to manage overfitting and handle imbalanced data, as fastballs were thrown more frequently than other pitch types. Other models capable of handling multiclass classification, such as Neural Networks, Gradient Boosting Machines, Support Vector Machines, and Multinomial Logistic Regression were tested. Still, Random Forest provided the highest accuracy (59.9%) with the best processing times. The model was

trained on an 80%-20% train-test split of the dataset. Hyperparameter tuning involved testing a sample of 70,000 observations with various mtry values. The final model used 100 trees, and predictions were generated for each batter's pitch mix.

Results

The Random Forest Model achieved a 59.9% accuracy. The confusion matrix revealed a tendency to overpredict fastballs, with nearly 60,000 pitches incorrectly predicted as fastballs, compared to 14,000 off-speed misclassifications and 20,000 breaking ball misclassifications. Despite these issues, the model offers insight into pitch trends across batters. Notably, fastballs dominate the pitch selection for all batters, while the relative proportions of off-speed and breaking ball pitches are influenced by the batter's handedness, reflecting nuanced adjustments in pitch selection based on handedness-driven matchups.

Limitations

The primary limitation is the relatively low accuracy (59.9%), which suggests potential overfitting to the fastball category. Given the dataset's size, computational power was a bottleneck in hyperparameter tuning and testing more complex models. With enhanced processing capabilities, model accuracy could reach 75% or higher, especially with more extensive hyperparameter tuning. Additionally, parallel processing was not effectively utilized in the Random Forest model, particularly when using the train() function from the caret library, limiting model optimization. As a result, the model may not fully capture pitch prediction variability across different batters.