Swing Probability Prediction in Major League Baseball Pitches

Olubayode Ebenezer

4/23/2024

Problem Statement

In baseball analytics, understanding the probability that a pitch will be swung at by a batter is crucial for both game strategy and player performance analysis. Currently, there is a gap in predictive analytics for pitches without a direct description of their outcome. The dataset includes comprehensive pitch data over three seasons, but the third season lacks descriptive outcomes for each pitch. This project aims to fill that gap by predicting the swing probability for each pitch in the third season based on the available data.

Additional Objectives

Identify key factors for specific pitch types by defining and analyzing middle-middle pitches—pitches within 0.5 feet of the center of the strike zone—and determine which variables most significantly influence the decision to swing at these pitches.

Develop comparative player metrics using the model's insights to establish metrics that compare individual players' performance to league averages, focusing on their decisions to swing, and providing a deeper analysis of player tendencies and skills.

Approach

1. Data Analysis and Preparation: Analyzed the dataset from the first two seasons to understand patterns and trends associated with swing decisions. Prepared the data by cleaning and selecting relevant features that influence a batter's decision to swing.
2. Feature Engineering: Conducted feature selection and engineering using polynomial interactions to build the model for swing probability. Selected features found most informative using feature importance, exploring interactions of features to improve model performance.
3. Model Development and Tuning: Developed a predictive model using the data from the first two seasons to estimate the likelihood of a swing for each pitch, including some hyperparameter tuning. The model considers various factors such as pitch type, count, pitcher, batter stance, and the physical characteristics of the pitch.
4. Model Validation and Evaluation: Validated the predictive model using a subset of data (e.g., cross-validation) to ensure accuracy and robustness, evaluated it using the F1 Score.
5. Prediction and Application: Applied the validated model to the third season's data to predict swing probabilities and evaluated the model's performance in terms of its practical utility for baseball teams and analysts.
6. Results Documentation and Presentation: Documented the methodology, model development, validation process, and results in a comprehensive manner.

Methodology

Data Loading and Merging

Load the datasets for Year1 and Year2, and the validation data for Year 3. Merge the Year1 and Year2 data into a single dataframe.

Data Exploration

- Numerical Distribution Plots
- Data Quality Reports:
  - Observations on Missing Values: Significant findings were noted regarding missing values and most suggested randomness in missing data patterns.

Data Cleaning

- For normally distributed data, impute missing values using the mean imputation grouped by related features of the group.
- For non-normally distributed data, consider median imputation grouped by related features (e.g., pitch type).
- For categorical data, mode imputation.

Data Preprocessing

- Dropped rows that were missing.
- Converted descriptions into a new SwingProbability I called SwingType category based on logical grouping derived from baseball rules and typical game scenarios.

Feature Engineering

Developed feature interactions to capture complex relationships within the data by creating polynomial and interaction features that reflect strategic elements of baseball pitching and batting.

Modeling and Results

Baseline Models

The two baseline models used are CATBOOST and LGBM Classifier. Utilized Smote for Sampling.

- Trained Catboost Modeling using the top best 23 predictive features gotten from Features importance (Feature Selection) that was built using LGBM.
- Built the second model using LGBM without the features interactions features generated through Polynomial.

Final Model and Submission

- Model Prediction: Computed the mean probabilities of the classes for swing prediction by averaging the probabilities obtained from two different models: LightGBM and CatBoost Classifier.
- Probability Array Example: The array mean probabilities holds the averaged class probabilities for each pitch. Each row in the array corresponds to a pitch, and each column corresponds to one of the four classes of swing probability: No Swing, Unlikely Swing, Attempt to Swing (bunt), and Definite Swing.
- Use argmax to determine the Predicted Class and Then used the year3 data to predict my Swing Likelihood and I then mapped the No Swing(0) and Unlikely Swing(1) to No Swing in my Swing Probability and Mapped Attempt to Swing (bunt)(2), and Definite Swing(3) as Swing.

Evaluation

Understanding the Baseball Swing Prediction Model

- Metrics Used: F1 Score, combining precision and recall to evaluate model performance.
- Why Use the F1 Score: Ensures both the accuracy of swing predictions and their comprehensiveness, critical in game-changing decisions.
- Predictions & Interpretations: Utilized argmax to select the most likely type of swing based on model probabilities.
- Model Reliability: Evaluated through cross-validation to ensure robust performance across different data sets.

Further Analysis

Understanding Middle-Middle Pitches

- Definition and criteria for middle-middle pitch identification.
- Impact of pitch location and speed on swing probability.

Metric Methodology: Swing Decision Efficiency (SDE)

Detailed explanation of the calculation and application of the SDE metric to season 2 data, including the classification of swing decisions and aggregation to determine efficiency.

Results and Implications

Discussion of top and bottom performers based on SDE scores, highlighting differences in swing decision-making skills among players and its implications for coaching and player development.

Conclusion

This is just an overview of my project. And my data are gotten from Miami Marlins Baseball team.