Machine Learning Project(CS-584-01)

Intermediate Report

Date: 04/02/2024
Team members:
Tirumalesh Nagothi (A20520569)
Tarun Sai Varanasi (A20526965)
Mohana Uma Sushmanth (A20525576)

Goal: Baseball Home Run Prediction Project(Analysis to predict whether the baseball batter will hit a home run or not based on different attributes)

Introduction

The goal of the Baseball Home Run Prediction project is to create a machine learning model that can forecast, using a variety of game variables and player characteristics, whether a baseball hitter will hit a home run. The objective of this project is to classify each at-bat as either producing a home run or not. It is handled as a classification problem.

Pitch speed, launch angle, launch speed, and other variables are among the many features in the dataset that are utilized to train and assess the predictive model. Missing value handling, outlier detection and imputation, and feature selection employing the SelectKBest model to determine the top 15 pertinent features are among the preliminary data preprocessing procedures.

Because of their propensity to attain high accuracy rates, a number of machine learning algorithms are taken into consideration for model training, including logistic regression, random forest classifier, and decision tree classifier. The project also includes pipeline design, performance metric monitoring, and hyperparameter tuning as essential components to improve the model's predictive powers and expedite the development process.

The ultimate goal of this project is to develop a reliable and accurate predictive model that will help analyze and comprehend the elements that go into hitting a home run in baseball, offering players, coaches, and baseball fans insightful information.

Tasks performed

Data cleaning:

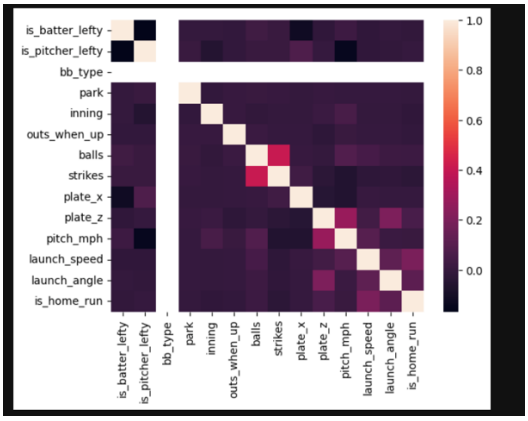
- Removed redundant columns such as bip_id, batter_id, and pitcher_id as they were identical in each row and provided no value for machine learning processing.
- The dataset does not contain any duplicate entries.
- After scaling the values, there was no noticeable change in the distribution of the pitch_mph and launch_speed features. We will investigate alternative methods if needed.
- There are some missing values in the columns: launch_speed, launch_angle, and bb_type.

	Column	Percentage
19	launch_speed	0.255276
20	launch_angle	0.254844
8	bb_type	0.000130

- Since imputing the null values with the mode of launch_speed and launch_angle did not result in an increase in outliers, we proceeded with filling the missing values using the mode of the respective feature.
- Additionally, for the categorical feature bb_type, we replaced the missing values with the most frequently occurring label in the column.

Outliers detection:
There are outliers in the continuous features such as plate_x, plate_y, launch_speed, and launch_angle. Despite attempting to remove the outliers using scaling techniques and transformations, there was not a significant change. Therefore, we plan to closely examine this issue and implement proper outlier imputation, possibly using fence values for detection.

Correlation:
There is minimal correlation among the numerical features in the dataframe, with the maximum correlation being approximately 0.24, which is positive. Additionally, we did not observe any significant negative correlations that could be utilized in a predictive model.



Feature Engineering:
For the most common and practical approach, we have opted for label encoding. This assigns a numerical value to each label, facilitating computation in the machine learning model. We extracted unique values from the object feature and assigned numerical values starting from 1. Starting with 0 could potentially result in 0 values, which would not be meaningful in this context.

Data transformation:
Initially, we applied the StandardScaler technique to scale the values, aiming to make the model more flexible during training compared to using raw values. We will explore alternative scaling methods if necessary.

Feature Selection:
We utilized the SelectKBest model to identify the top 15 features from the existing columns in the dataset. As a result, we will consider the following values:

```
['home_team' 'away_team' 'batter_team' 'batter_name' 'pitcher_name'
 'is_batter_lefty' 'is_pitcher_lefty' 'bb_type' 'bearing' 'pitch_name'
 'park' 'inning' 'outs_when_up' 'balls' 'strikes' 'plate_x' 'plate_z'
 'pitch_mph' 'launch_speed' 'launch_angle']
```

When it comes to model selection, we have considered several common algorithms, including logistic regression, random forest classifier, and decision tree classifier. These models were chosen due to their ability to achieve a minimum accuracy of 90%. However, we cannot solely rely on accuracy as a performance metric. We need to further analyze the models using various plots and classification metrics to assess their overall performance effectively.

Things to be added/ change:
- We will conduct hyperparameter tuning to optimize the algorithms further.
- We will monitor the performance metrics across the training, testing, and validation datasets to ensure the models generalize well.
- We plan to design the final solution using pipelines to streamline the data processing and modeling steps.
- If required for the project submission, we will deploy the finalized model.

References:
https://pandas.pydata.org/pandas-docs/stable/index.html
https://numpy.org
https://www.python.org
https://scikit-learn.org/stable/
https://matplotlib.org
https://plotly.com