Abstract

The goal of this project is to use a model to predict a batting average of MLB players with given statistics. This project was started to help the predicting batting average of future seasons based on the normal statistics. After data cleaning and multiple visualizations, I made a model to predict the batting averages of the players.

Design

This project was started with the idea of a statistical relationship between many other features and batting average. Data for this project was scraped from baseball-reference.com which has all data of MLB players. Making further prediction model will help the team to build the batting strategy.

Data

The dataset contains a 3 years record of MLB players (2019, 2020, 2021). It has 4633 samples of batters and 11 features (age, game played, plate appearance, at bat, run, hit, double homerun, RBI, strike out, based on balls) with a target(batting average). Nearly ⅘ of data sample has been dropped to make sure to include the batters who have enough at bat.

Algorithms

1. Exclude all batters who had less than 162 at bat in one season
2. Check the correlation of each feature vs target.
3. Build a model based on the dataframe.
4. Exclude uncorrelated columns and calculate statistics to compare.
5. Find out RMSE and MSE.

Tools

* Numpy and Panda for data manipulation
* Scikit-learn and Statsmodel for modeling
* Matplotlib and Seaborn for plotting

Communication