Austin Kingsley

(773) 341-8026

akingsley@regis.edu | akingsley319@yahoo.com

https://github.com/akingsley319/MLB_xBA_Prediction

MLB Expected Batting Average (xBA) Prediction

In an effort to better understand the ability for a major league baseball player to hit a baseball, I will be attempting to predict the expected batting average of players based on matchups and performance history.

This data science task will involve the statistics of games going back to 2017. Statcast was introduced to all ballparks in 2015, but the currently used measurements were adopted by all ballparks in 2017 according to the baseball savant api documentation. As such, this project will involve classification, time-series analysis, and a great deal of analysis for missing data and handling superfluous entries.

The data involved in this project is incredibly large, involving data from every game played between 2017 and 2021 for initial model training and testing. There are also over 50 columns of potentially relevant data.

I plan to use dimensionality reduction processes, clustering algorithms, various evaluation metrics, and time-series analysis models. I will be weighing the various options for filling various missing numerical categories based on analysis of existing data. I will be filling missing event and player data using the ‘des’ category, which has a written description of the play. I might also seek out another resource for connecting missing names to the ‘batter’ fields.

Known problems I will overcome include inconsistent pitcher schedules, the necessity of time-series analysis from multiple parties (pitchers and hitters), missing knowledge of rosters, and the potentially different number of pitchers on team rosters.

For the pitcher schedules, I will use blocks of time for those models which need it. I would consider using five day blocks due to this being the common length of time between starts. I would do this for relief pitchers as well for consistency and ease.

For the introduction of multiple parties to the prediction, I am looking at ways to stack the models. This might involve inserting predicted performance and matchup analysis into a supervised learning model such as random forest.

For roster knowledge, I hope to find a resource that will provide this information. If this does not work, I will gather players used in a predetermined number of prior games and construct a probable roster from that.

For the roster differences in terms of hitters and pitchers, I intend to use a predetermined number of hitters/pitchers from each team based on predicted performance and the expected starter.

I plan on performing data cleaning, analysis, and table separation for the first two weeks. Week three and four will be devoted to analysis, clustering, and model analysis for pitchers and batters. I expect this process to possibly take more time, due to the incorporation of new techniques and possible complications arising here which I did not foresee. I will then spend week five catching up and putting the pieces together for final prediction. Week six will be spent attempting any further model tuning, making it usable with current MLB play, and putting together the presentation of the project.