

KBO Baseball Projections Project Proposal

Problem Statement

Based on past season and current year performance, which hitters in the KBO will have the highest performance statistics (on base plus slugging, batting average) at the end of the 2020 regular season?

Context

The hypothetical client for this project is a prospective advertiser that is looking for a player spokesperson. They want an everyday player, so they are looking at hitters in the KBO. Before they select one, the advertiser has asked me to make a shortlist of the players that will perform at the highest level for the remainder of the season. Based on my analysis and subsequent shortlist, the advertiser will select a spokesperson.

Criteria for Success

At the end of the 2020 regular season, player statistics should match forecasted performance with a mean absolute percentage error of less than 15%.

Scope of Solution Space, Deliverables

The solution space will be limited to the projected performance of KBO hitters.

Deliverables, all of which will be contained in a GitHub repo, will include:

- forecast of end-of-year individual hitter statistics
- project report detailing conclusions and the process that lead to those conclusions
- slide deck presenting forecast methods and results

Stakeholders

Advertising Company Executive Team, who will make final decisions based on my recommendations

Data Sources

Baseball Reference has individual player statistics for the 2020 (and past) KBO seasons:

[2020 Korean Baseball Organization](https://www.baseball-reference.com/league/2020/kbo/)

This data can be pulled directly from the website in CSV format

Outline of Problem Solving Steps

1. Problem Identification

- a. making this project proposal
- b. identifying the specific scope of the project

2. Data Wrangling

- a. collect data from Baseball Reference
- b. organize data into one cohesive DataFrame
- c. clean data to deal with “na” entries
- d. make sure data definitions within dataset are clear

3. Exploratory Data Analysis

- a. create plots and charts to look for relationships within dataset
- b. see which features of the dataset might have surprising relationships

4. Pre-Processing and Training Data Development

- a. standardize dataset
- b. split into training data for modeling

5. Modeling

- a. select various potential models
- b. check performance metrics of potential models
- c. deploy the model with the best performance

6. Documentation

- a. create a project report detailing process and conclusions
- b. create a slide deck presenting forecast results and the methodology therein