Predicting All Stars Based of a Full Year of Baseball

Team Members: Chris Spartz, Kevin Gnaster, Zach Josten
Team Number: 31

Data:

The data we are using for this project is Baseball statistics from the last 20 years (from 1999-2019), with an emphasis on more advanced Baseball statistics such as WAR, WRC+, BABIP.

Data source:

https://www.fangraphs.com/leaders.aspx?pos=all&stats=bat&lg=all&qual=0&type=8&season=2019&month=0&season1=1999&ind=1&team=0&rost=0&age=0&filter=&players=0&startdate=&enddate=

Heilmeier's Questions:

What are you trying to do? Articulate your objectives using absolutely no jargon.

Based on MLB position players performances in the last 20 years, we are trying to create a logistic regression model to predict if they will be an All-Star that season. The All Star Game takes place in the summer every year, so they are only judged based on their performance halfway through the season. We are going to look at players' full seasons instead to fit the model and then use the model to find who should be All-Stars in a given season.

How is it done today, and what are the limits of current practice?

Players are selected today by multiple rounds of fan voting. Fans can be very biased based on their favorite players, teams and how to measure players. Some fans pick players based on who they like to watch and others base it on their favorite statistics. Some of the most popular statistics are not representative of how good a player actually is. This can skew the data to favor some of the players better in more traditional statistics.

What is new in your approach and why do you think it will be successful?

The change in our approach is that in Baseball the All Star game is halfway through Baseball's 162 game season, whereas we will be basing our selections on the full season's worth of statistics. The statistics of each player is the focus instead of Fan vote to decide who should be in the all star game.

Who cares? If you are successful, what difference will it make ?

Players get paid bonuses for whether or not they make the All Star game, and some players might feel like they have lost out on some money since their performance should have gotten them into the All Star Game.

What are the risks?

There are a few risks involved with the final logistic model we will come up with. There could be biases in the model because players can slump in the second half of the season or heat up. There can also be injuries or other unforeseen things that can't be predicted that can influence the model.

How much will it cost?

There is no cost associated with the analysis of the data, as it is open source data.

How long will it take?

The analysis of the data will most likely take around 3 weeks for in-depth predictions and information to be presented.

What are the mid-term and final "exams" to check for success?

The midterm is the completion of the python code with in depth analyses and the logistic regression model. The final checkpoint is getting our data in a good place to present to non-expert audiences.