Forecasting Batting Performance using Machine Learning

Larry Jackelen

amateur sabermetrician

long-suffering Minnesota Twins fan

Problem Statement

Major League Front Offices want to know which players to target in free agency. Two major factors in that are:

1. How good they will be next year
2. How much they are going to cost in salary

A popular metric for a player's value is Wins Above Average (WAA).
Using this and other metrics available from Baseball-Reference.com, utilize machine learning to provide a forecast of a player's next season WAA per game.

Success Criteria for the Model

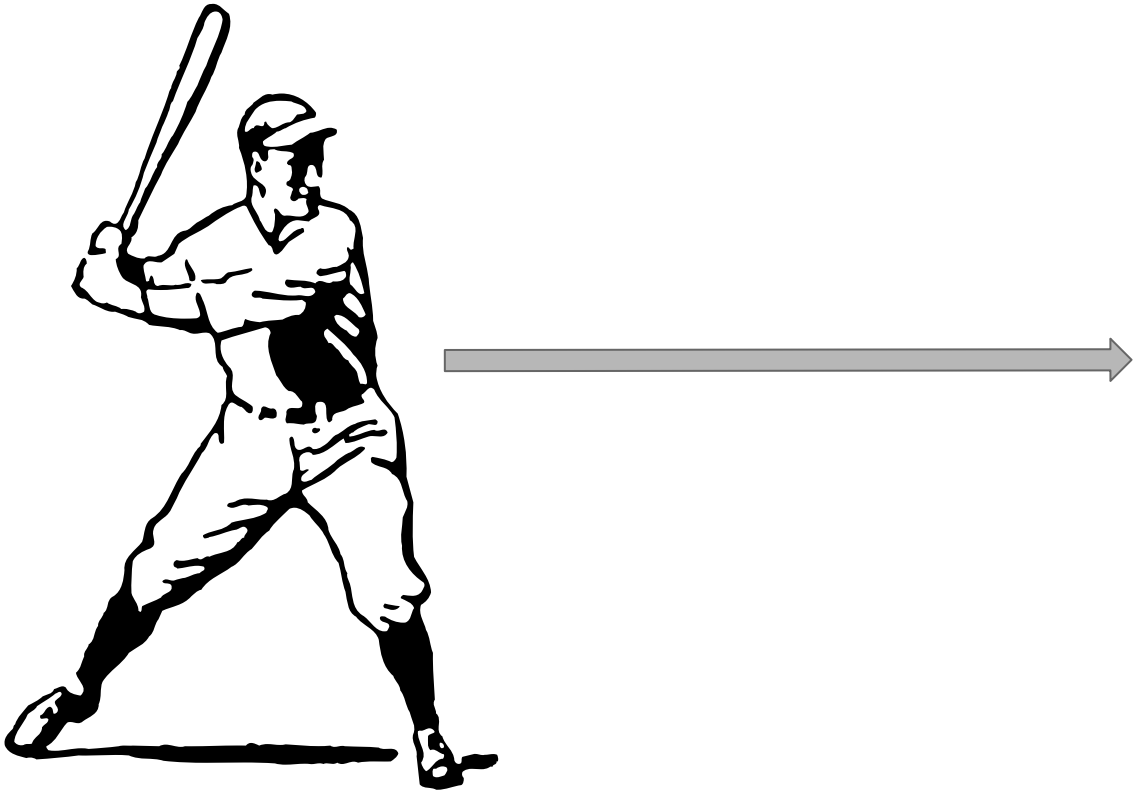
We want to be able to forecast better than using a weighted average of a player's three previous WAA per game.

Get to the World Series!

An Intro to Batting Performance

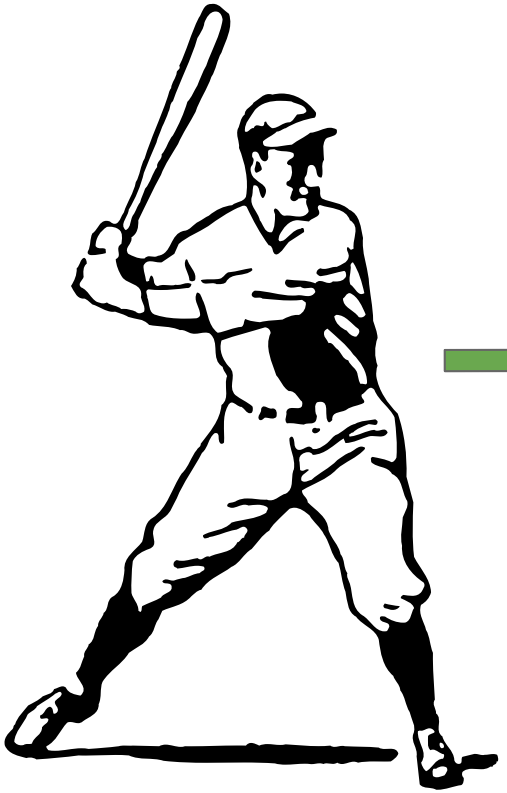
Meet our player:

Homer B. Gone



Meet our player:

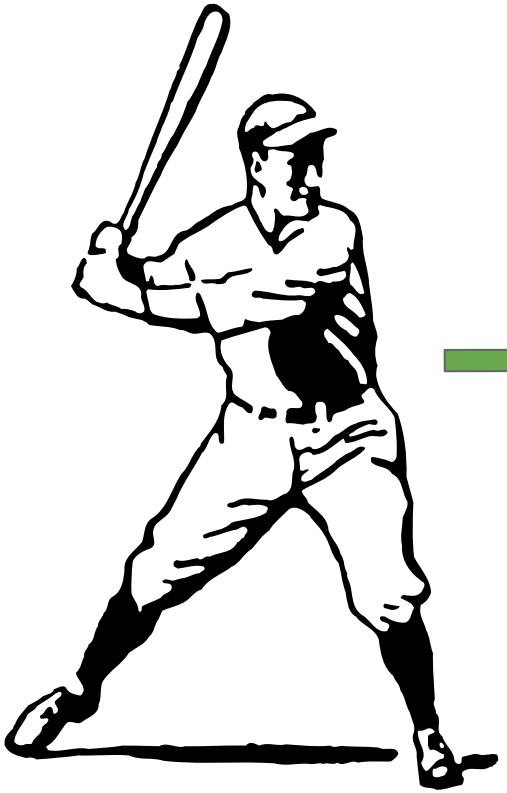
Homer B. Gone

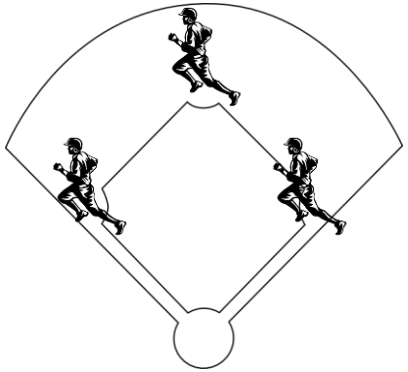


Nice Double!

Meet our player:

Homer B. Gone





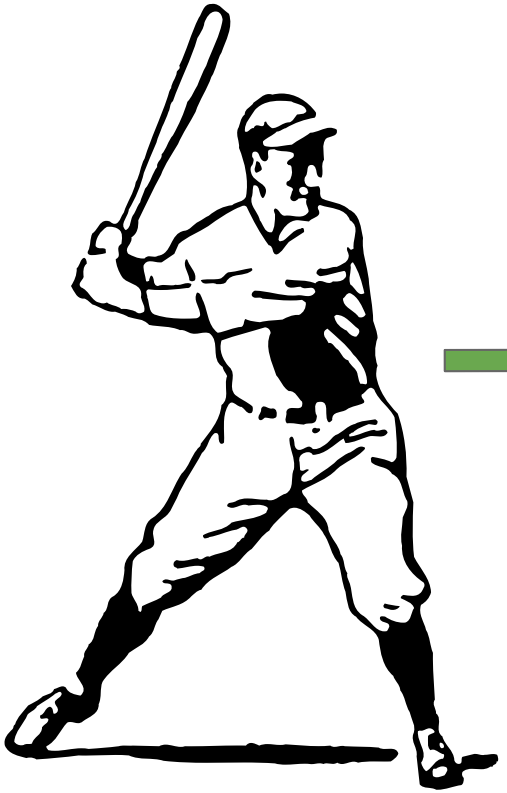
Your team scores

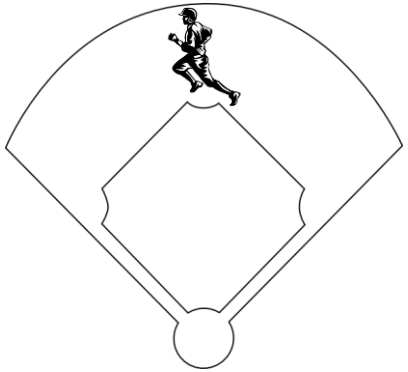
3 Runs

Nice Double!

Meet our player:

Homer B. Gone





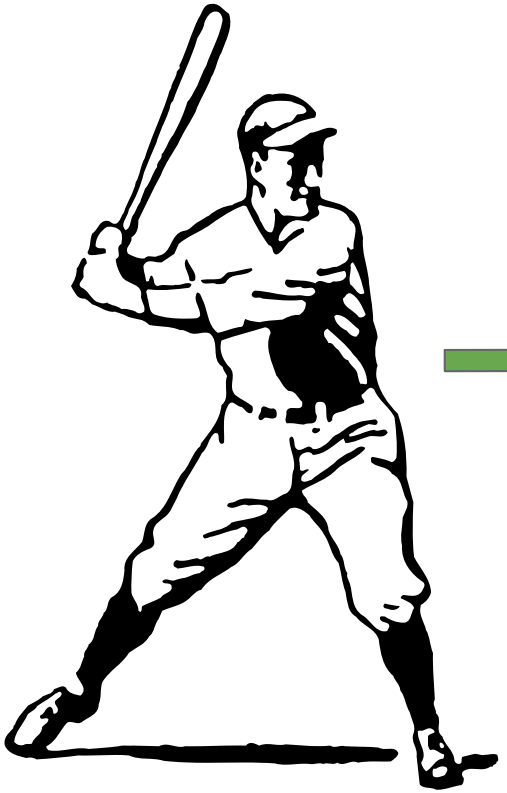
Your team scores

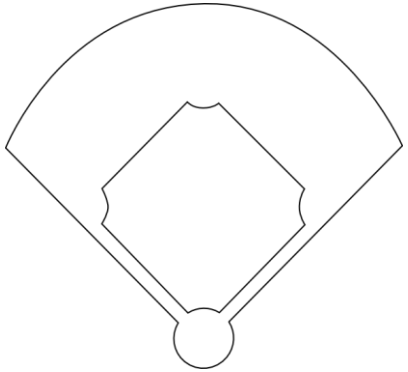
1 Runs

Nice Double!

Meet our player:

Homer B. Gone





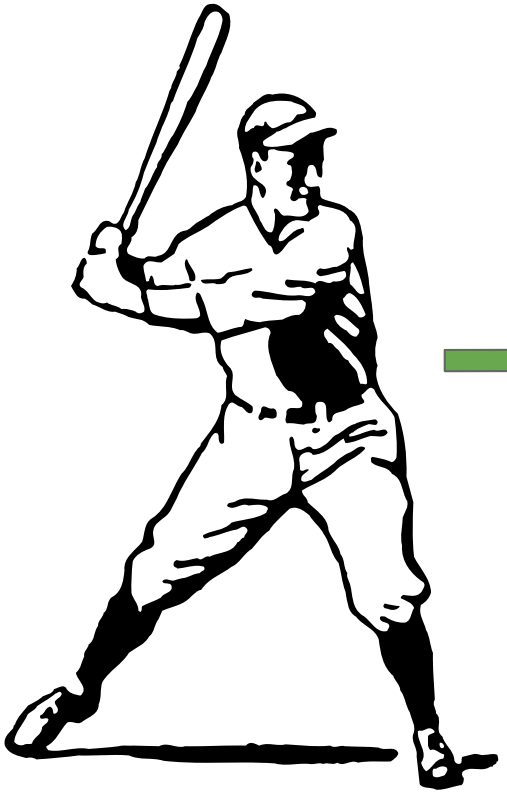
Your team scores

0 Runs

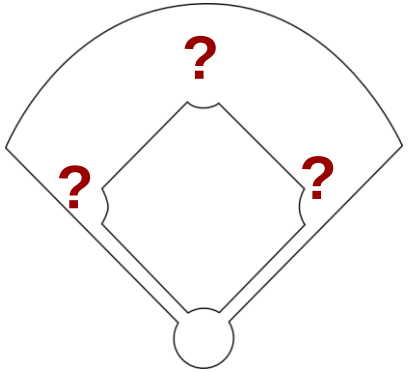
Nice Double!

Meet our player:

Homer B. Gone



Your team scores



?

Nice Double!

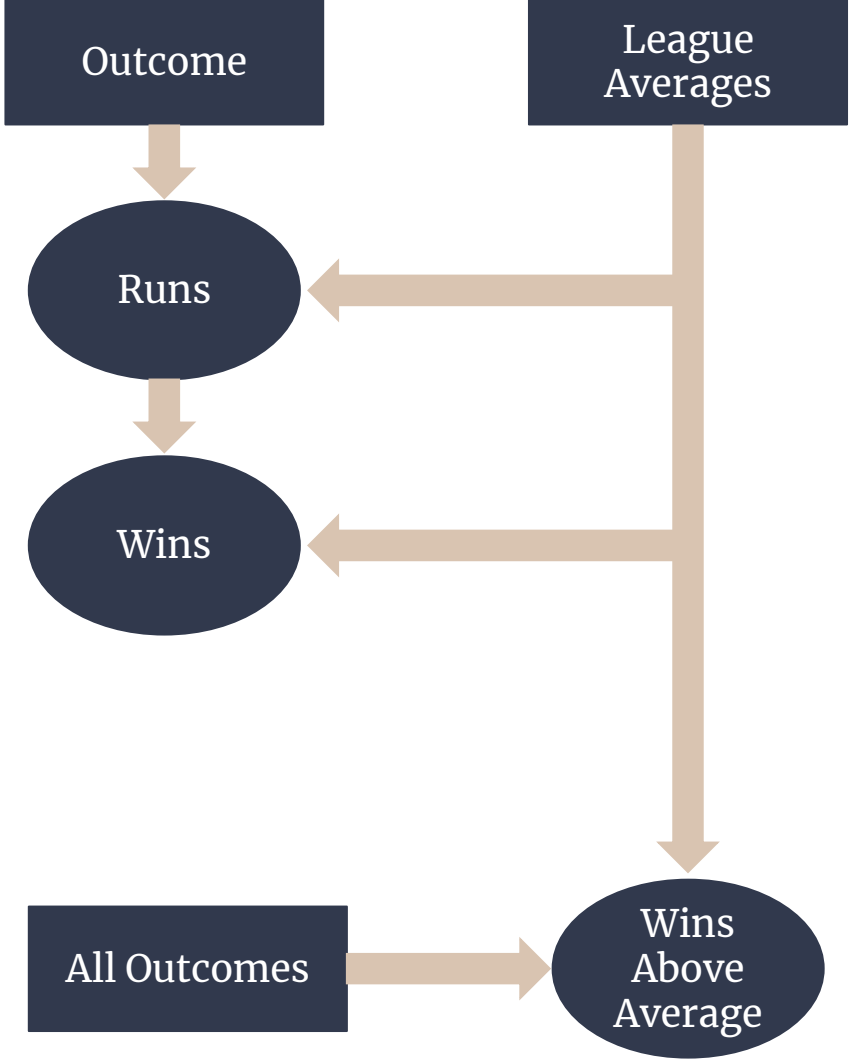
The batter cannot control who is on base, so it leads us to ask:

How **Valuable** *is* a double??

Outcomes to Value: Calculating Runs and Wins

Enter Baseball Reference and the World of Sabermetrics!

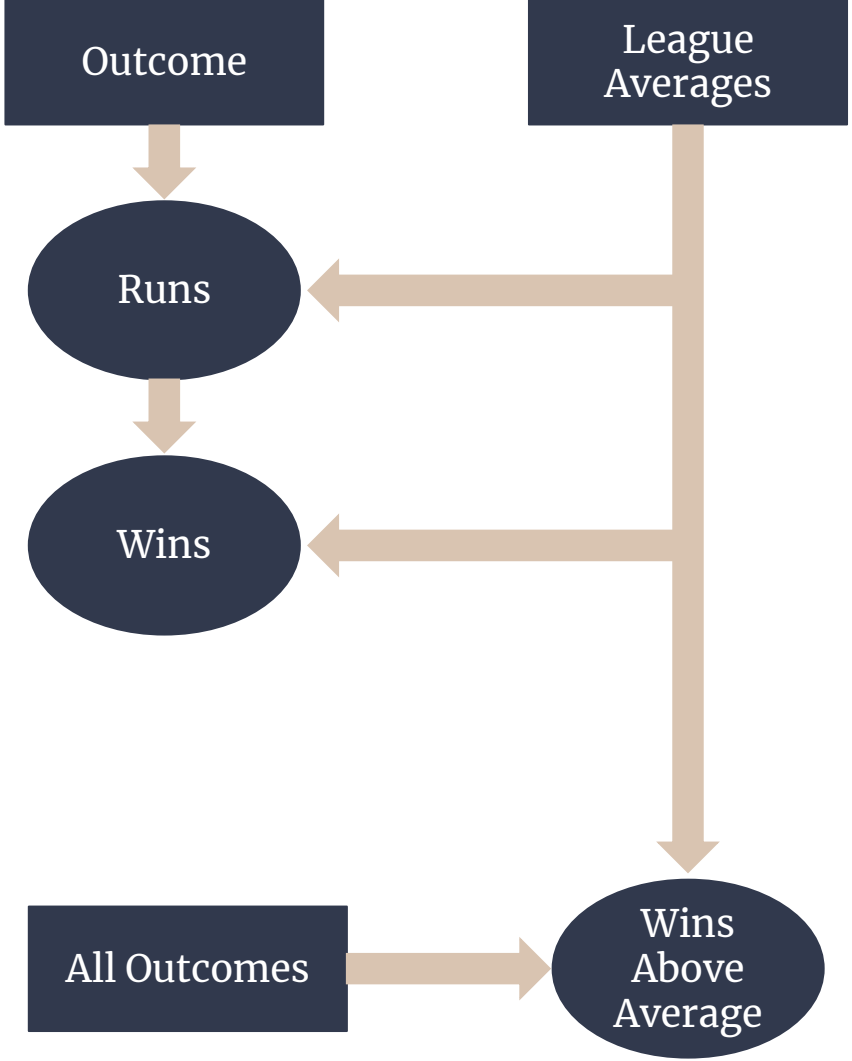
Baseball Reference is an online platform that provides historical statistics for all things major league baseball. One of their main features is their generation of value statistics.



Enter Baseball Reference and the World of Sabermetrics!

It's all about context!

Baseball Reference is an online platform that provides historical statistics for all things major league baseball. One of their main features is their generation of value statistics.



The Problem:
Forecasting Future Value

The Data

Statistic and value information for each player from the past three years

- Games Played
- Positions Played
- Runs and Wins created
- Player Age

The Model

Linear Regression

Baseline of weighted average of the past three seasons.

The Results

Baseline Score: **29%**

Our Model Score: **39%**

Where will this be used?

Finding free agents and evaluating what salary to offer, given they:

- Have 3 consecutive years of playing metrics available
- Are Positional Players (non-pitchers)
- Are 21 - 38 years of age

These requirements are somewhat stringent and are the limitations of this model.

How can this be improved?

- Include more traditional stats like home runs, hits, RBIs, OBP, OBP+ to see if they can help predict runs
- Decompose Baseball-Reference's logic even more and try to model those pieces individually
- Forecast beyond just one year, train to get the next 2-3 years
- Add pitchers!

Any Questions?