Forecasting Batting Performance using Machine Learning

Larry Jackelen

amateur sabermetrician

long-suffering Minnesota Twins fan

Sabermetrics

The empirical analysis of baseball, especially baseball statistics that **measure** in-game activity.

Problem Statement Background

Baseball is a data-rich sport. The number of games, the repeatable trial nature of its pace of play, and its history of tidy record-keeping all contribute. Looking into this bounty of data and forming new metrics to answer questions about how players and teams perform is in itself a pastime.

Thanks to sabermetricians, a family of metrics have been created to try and measure a certain player's value in the context of the season they played.

A powerful tool, the value metric this project focused on was
Wins Above Average (WAA)

In short, how many wins did a player provide above the average player that year?

How is this useful?

Major League Front Offices are always interested in which direction a player is going performance-wise, and that is especially the case when looking for new players during Free Agency. The two major questions they have are:

1. How good will they be next year
2. How much should they cost in salary

Free agency is a market at the end of the day. If a player is forecasted to be more valuable than what the market is saying, there is an opportunity. Finding these inefficiencies allow a team's resources to be maximized, leading to a better chance a championship

Problem Statement

Armed with a player’s past three-years of WAA and other metrics available from Baseball-Reference.com, utilize machine learning to provide a forecast of a player's next season WAA per game.

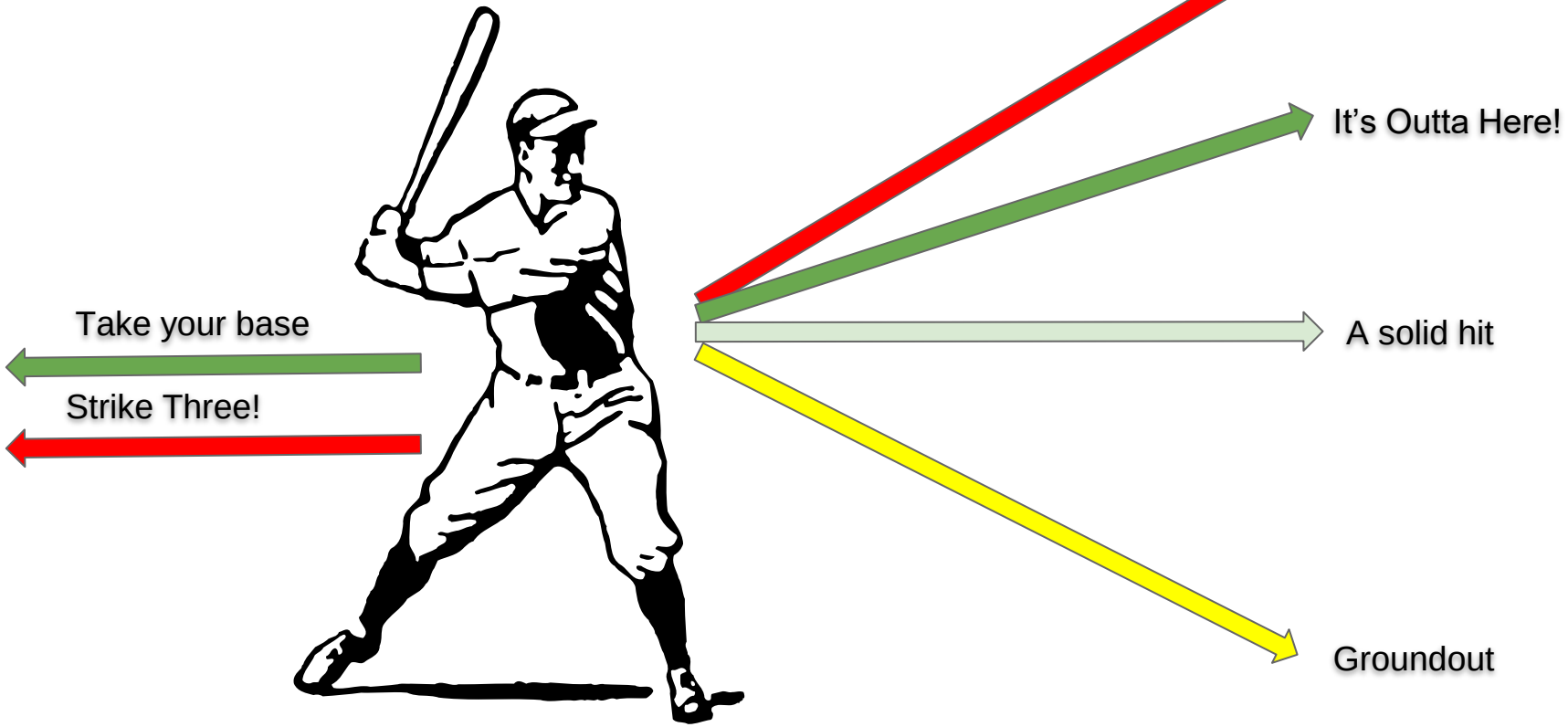
The goal is to outperform the baseline model of the weighted average of a player’s three previous WAA per game.

Get to the World Series!

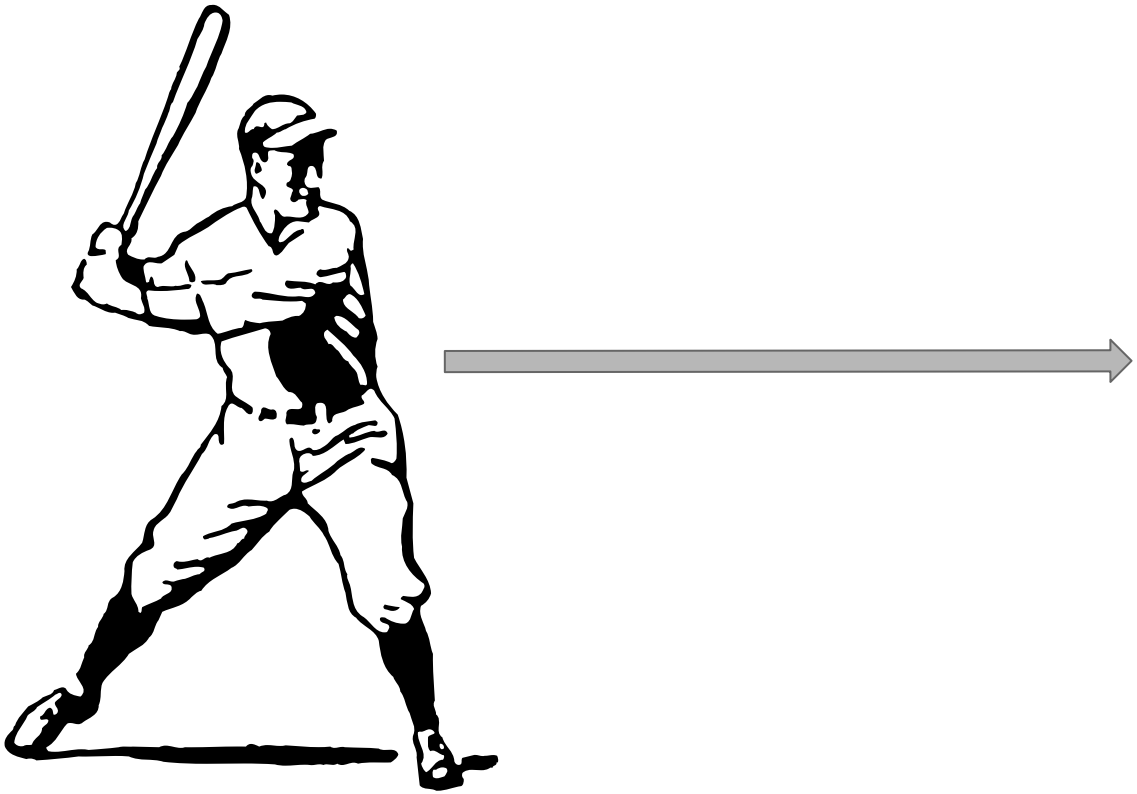
An Intro to Batting Value

Meet our player:

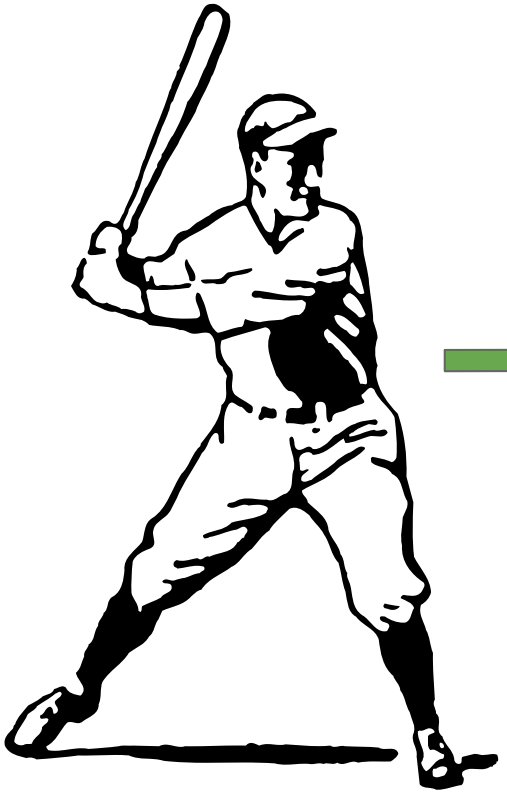
Homer B. Gone



Homer B. Gone

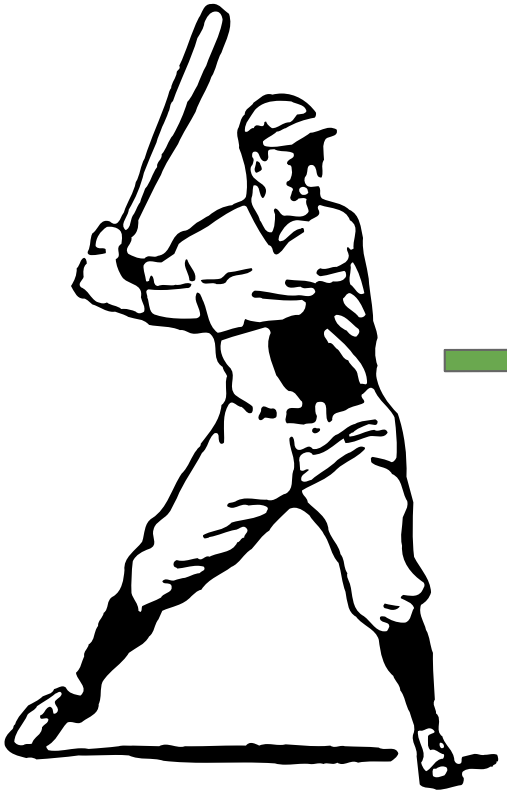


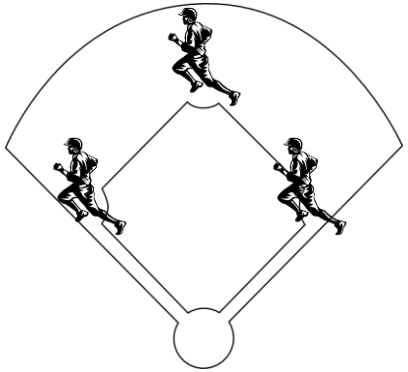
Homer B. Gone



Nice Double!

Homer B. Gone



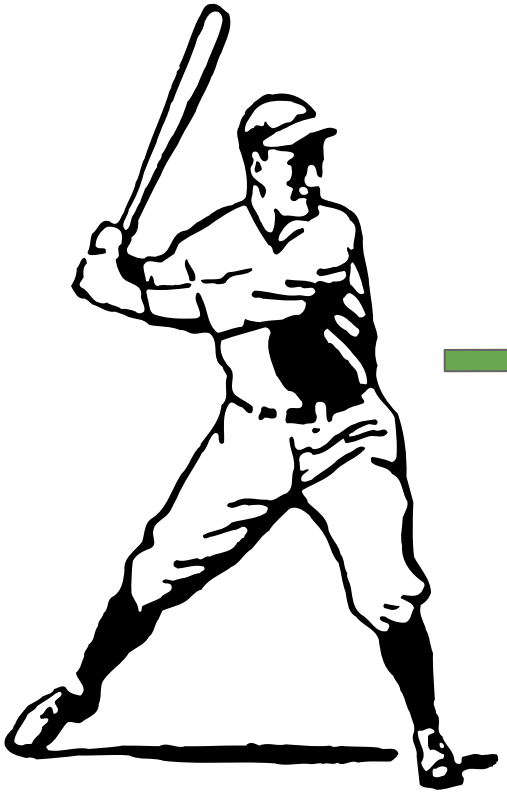


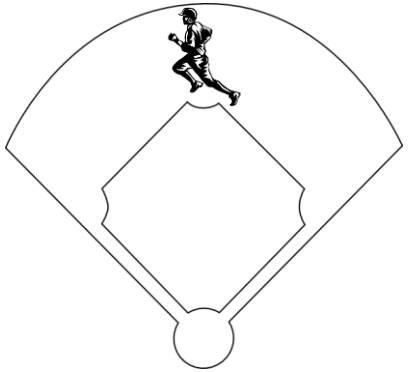
Your team scores

3 Runs

Nice Double!

Homer B. Gone



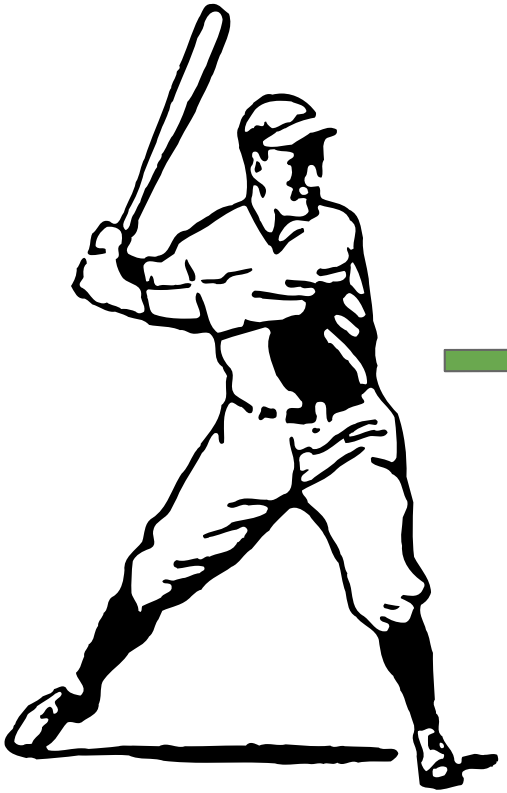


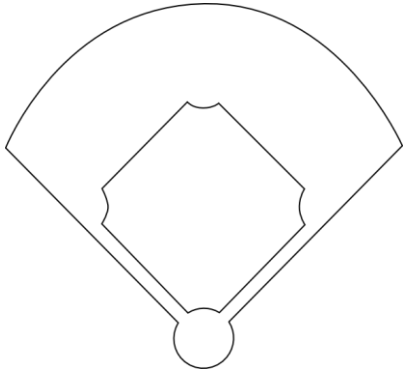
Your team scores

1 Runs

Nice Double!

Homer B. Gone



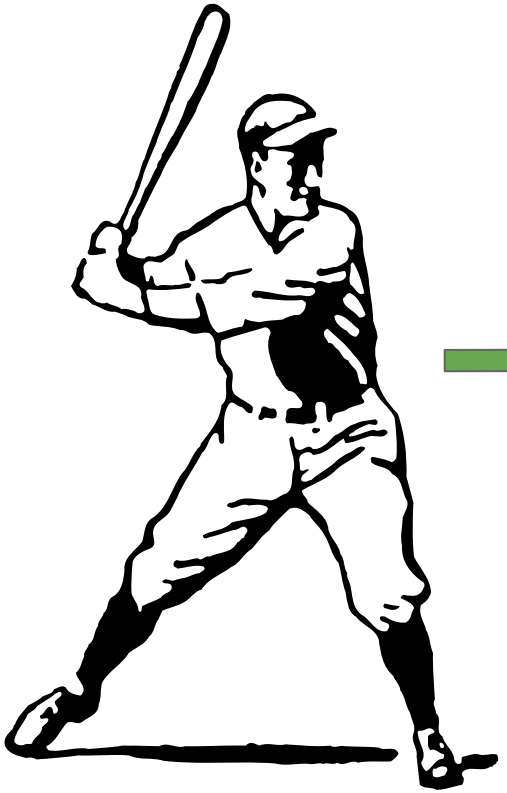


Your team scores

0 Runs

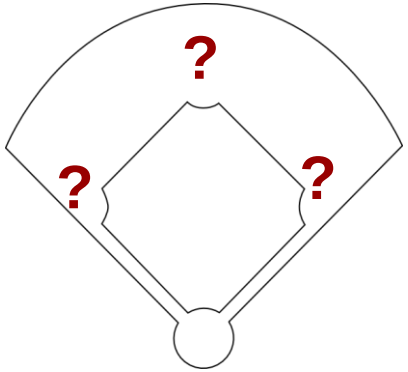
Nice Double!

Homer B. Gone



Your team scores

?



Nice Double!

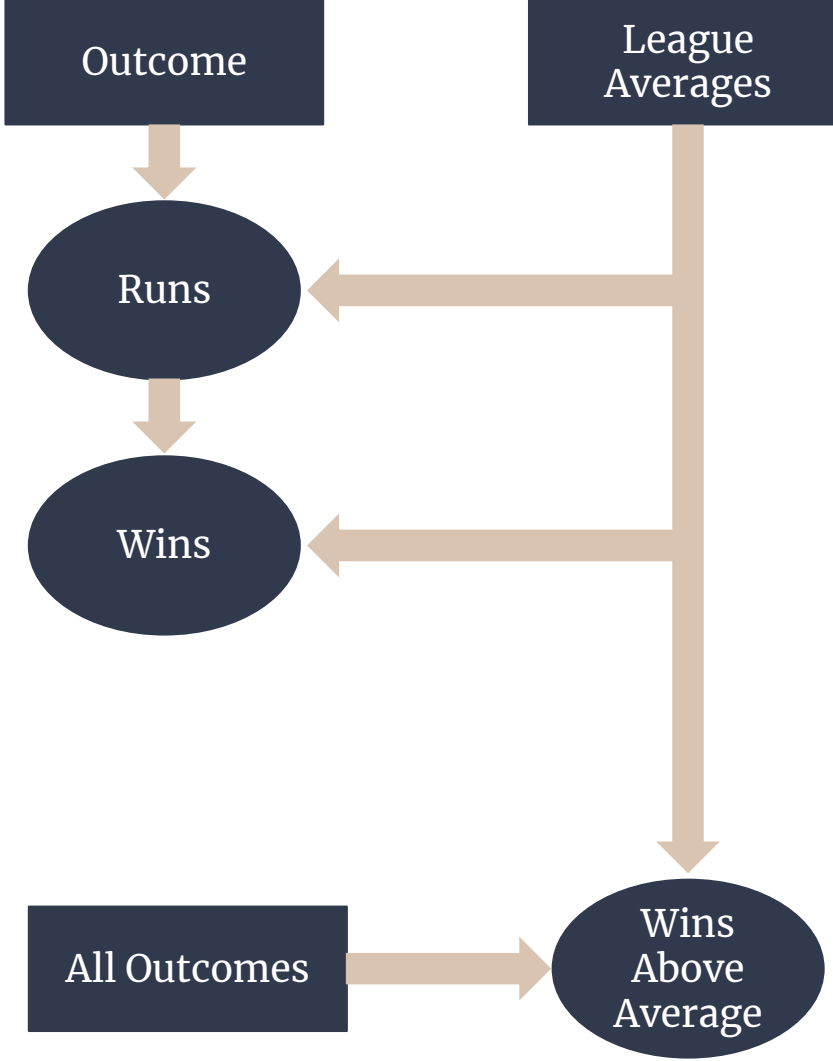
The batter cannot control who is on base, so it leads us to ask:

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

Outcomes to Value: Calculating Runs and Wins

Enter Baseball Reference and the World of Sabermetrics!

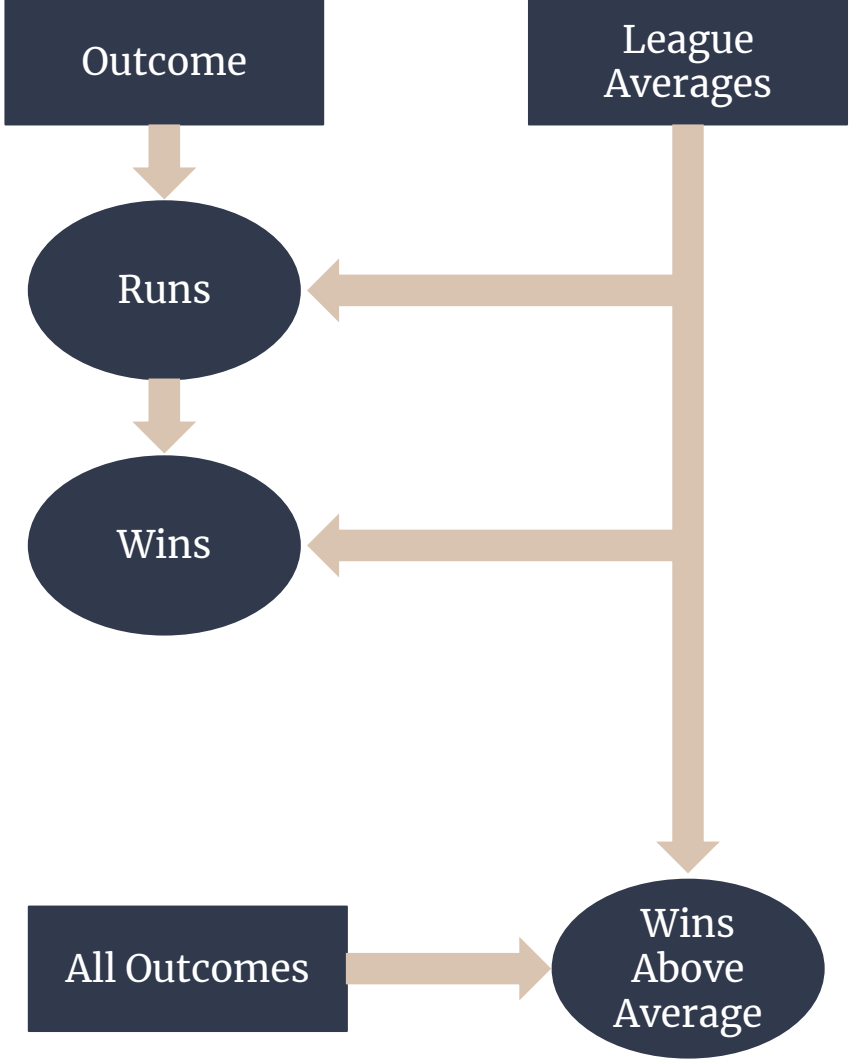
Baseball Reference is an online platform that provides historical statistics for all things major league baseball. One of their main features is their generation of value statistics.



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.



Data Collection

Data Sources

Baseball-Reference

- Value datasets publicly available for download
- Data request to company for scrapable data

Lahman's Baseball Database

- Public download

Data Cleaning

Data Layout and Structure

Team played for,
Order of teams played for intra-season

Age during season,
The Year of the season

Classic Stats

Value Stats

Our Player's Name

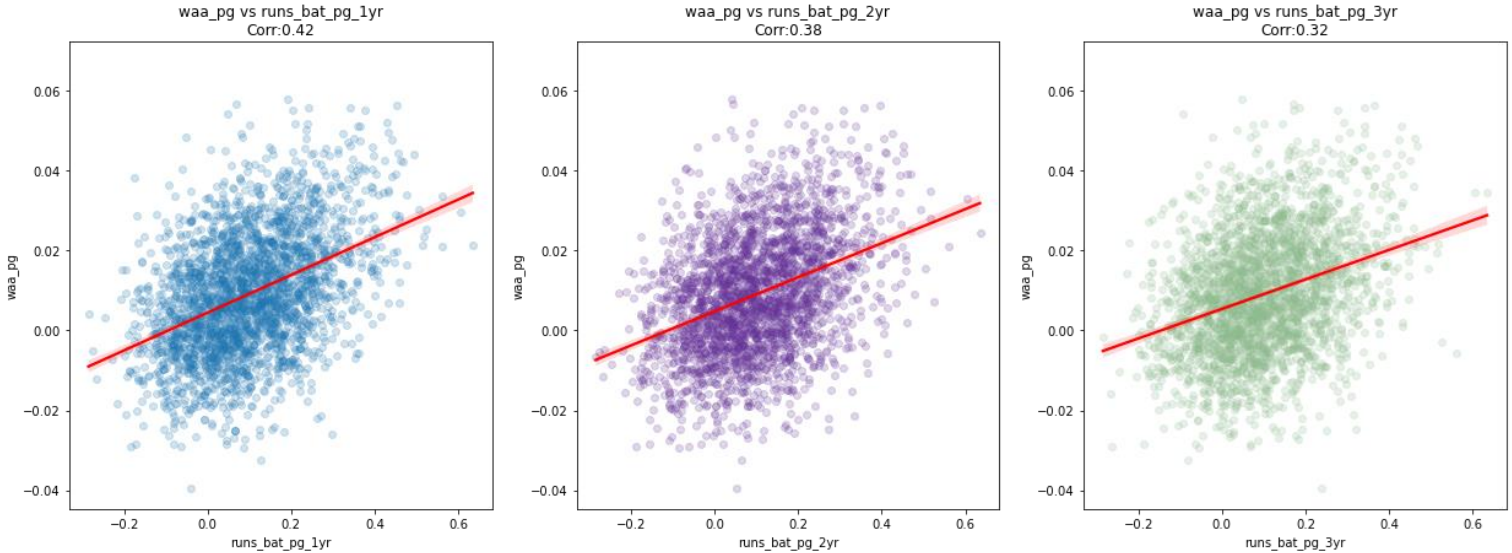
	name_common	age	year_id	team_id	stint_id	pa	g	runs_bat_pg	waa_pg
101325	Sammy Sosa	20.0	1989	TEX	1	88.0	25	-0.214000	-0.028400
101324	Sammy Sosa	20.0	1989	CHW	2	115.0	33	0.059394	-0.012424
101326	Sammy Sosa	21.0	1990	CHW	1	579.0	153	-0.053660	-0.009739
101327	Sammy Sosa	22.0	1991	CHW	1	338.0	116	-0.144310	-0.004397
101328	Sammy Sosa	23.0	1992	CHC	1	291.0	67	-0.019254	-0.003731

Data Cleaning To-Do's

PROBLEM	SOLUTION
Data contains stats from 1871 through 2022. Fundamental things about the game have changed.	Keep seasons from 1962 - 2019 where data keeping was more standardized and the league more stable
Some batters only played a few games	Require 3.1 PA per team game, mirroring the MLB's eligibility rules for rate stats
Multiple stints in a season disrupt the one player-season per row structure	Aggregate data by stint to get to one row for each player-season instance
Not all players play the same amount of games, not all seasons contain the same amount of games	Turn non-rate stats into per-game stats
Features we want to model on are on different rows in the database	Create lookbacks for features 1-3 seasons before target season, keep only rows with full lookback completeness

Using Lookbacks

Lookbacks allowed for flexibility to capture signal across the three previous seasons of the player in the target



The general behavior for each features was a decreasing correlation as lookback increased

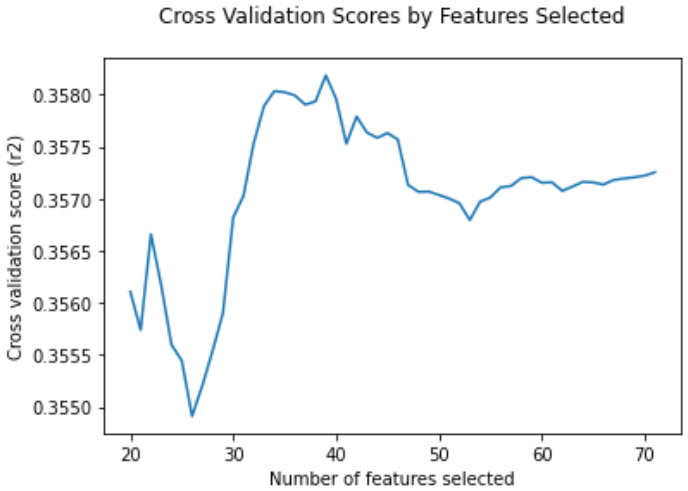
Data Modeling

Preprocessing

1. Standard Scaling

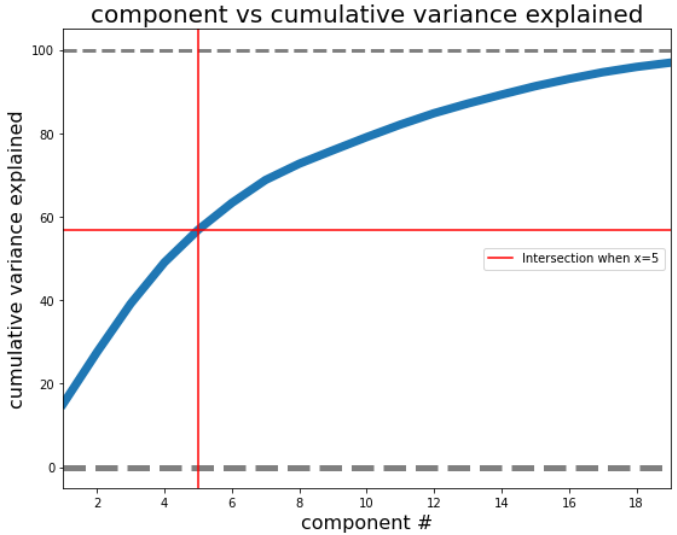
2. Recursive Feature Elimination

Minimum Features Required: 20
BayesianRidge() estimator
39 kept, 32 dropped



3. PCA of Eliminated Features

Add top 5 PC features by variance explained
Capture 56.84% cumulative variance explained



Model Selection (Baseline = 29.6%)

Random Forest Regression

Halving Grid Search to find:

- max_depth **5**
- min_samples_split **2**
- n_estimators **400**

Test Score of **35.9%**

Top Feature Importances:

- waa_pg_1yr
- waa_pg_2yr
- waa_pg_3yr
- age
- various runs features

SVM Regression

Grid Search to find:

- C **0.01**
- epsilon **0.01**
- gamma **scale**

Test Score of **34.0%**

Used default kernel (rbf) which doesn't provide feature importances by default

Linear Regression

Test Score of **39.0%**

Applied Lasso for feature selection

Test Score of …

39.1%

Model Results

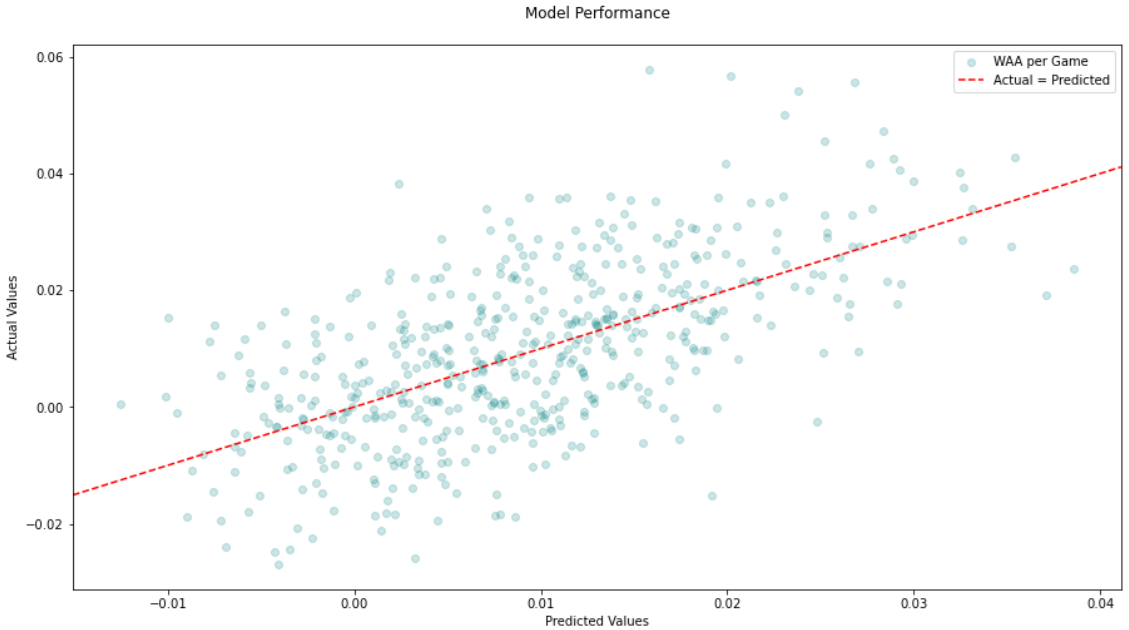
Important Features

waa_pg_1yr , _2yr, _3yr

All had similar standard deviations with 1 yr being 1.05 times more impactful than 2_yr and 2.8 times more than 3_yr

age

Each year of age negatively impacts the predicted waa_pg the same as 0.15 standard deviations of waa_pg_1yr



Results and Conclusions

Using the Model

Things it can do	Things it won't do
Help evaluate the next year performance of a free agent with three consecutive years of playing metrics	Evaluate rookies, players with no playing experience, or players coming off an injury
Help evaluate players presently on the team to forecast existing team value	Evaluate pitchers, players younger than 21 or older than 38
Outperform the baseline of the weighted average of past three seasons of a players waa_pg, as well as a baseline of straight average (29.8%), and just the previous year performance (6.0%)	Build the trophy case it earns

The Future

This model just scratches the surface of using the data available online through resources like Baseball-Reference. Everyday, they and their competitors train and test out new ways of capturing value that will ultimately make it to a Major League Team Office.

Ways to improve in the future:

Include more statistics beyond the value-based ones used in this model. Things like RBI, HR, OBP, OBP+ can all capture signal that may not be captured in Baseball-Reference's process.

Work to further decompose Baseball-Reference's proprietary metrics to see if modeling the components at an even lower level that runs, etc. can yield stronger results

Expand to work for pitchers, the other side of the ball!

Expand forecast to beyond one year to help assist longer free-agent contracts.

Any Questions?