Group 3: Predicting Baseball WAR

Gioia Bonanno-Garcia, Ari Crumley, Vincent West

2025-12-03

Overview

- Goal: build a model to predict WAR (Wins Above Replacement)
- WAR measures a player's total contribution compared to a replacement-level player
  - It combines multiple aspects of performance: hitting, base running, defensive value, positional difficulty, and their playing time, and puts it into a single number.
  - Interpreted as the number of additional wins a player adds to a team
  - Since it provides a single encompassing measure of a player's value, WAR is greatly relied on by MLB front offices

Motivation

- Can a player's current-season performance statistics be used to predict their next season Wins Above Replacement (WAR)?
- Major League Baseball teams rely heavily on WAR to evaluate a players value, make contract decisions, and project roster needs.
- It provide a competitive advantage for a team by:
  - Helps identify declining players
  - Allows for budgeting and contract planning
  - Encourages player development and roster optimization
- Our goal is, by using statistical and machine learning models, we will:
  - Identify which player statistics best predict future WAR
  - Compare the performance of three different statistical learning models: OLS, LASSO, and BOOSTING
  - Build a model that maintains high predictive accuracy

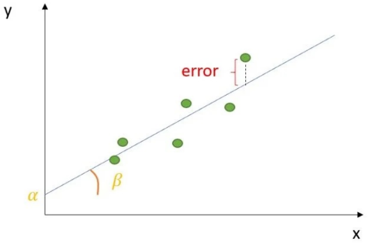
Data used

- Our data came from baseball-reference.com
- We used standard batting data from 2020-2025 to build our models (500 obs each year)
  - 2020-2024 was used for training
  - 2025 was our test data set
- Variables:
  - WAR, age, **games played**, plate appearances, at bats, runs scored, **hits**, doubles, triples, **home runs**, RBIs, stolen bases, caught stealing, walks, strikeouts, **batting average**, on base percentage, slugging percentage, OPS percentage, OPS+, rOBA, Rbat+, total bases, double plays grounded into, hit by pitch, sacrifice hits, sacrifice flies, intentional walks

Models Used

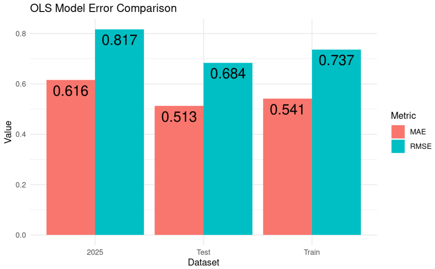
OLS Models

- An Ordinary Least Squares (OLS) model identifies and measures the relationship between a response variable and predictor variables
  - It finds the best-fitting linear trend that minimizes squared error.
- Pros:
  - Convenient
  - Has accurate regression results for linearly related data
- Cons:
  - May be too simplistic for real world examples
  - Assumptions of Linear Regression



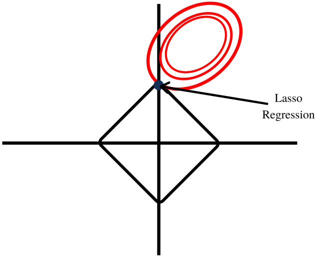
OLS Metrics Plot

- Error metrics for both final testing data set (2025) and the training data split into training and testing sets
- Significant terms:
  - age, games played, plate appearances, at bats, runs, hits, doubles, triples, home runs, stolen bases, caught stealing, walks, strikeouts, OPS+



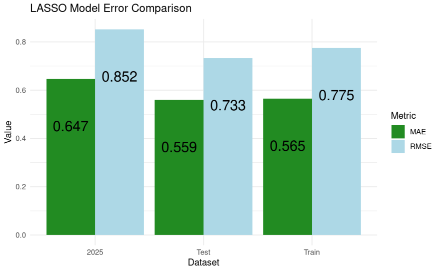
LASSO Models

- LASSO models perform regularization (L1), which shrinks some coefficients to exactly zero
  - Essentially feature selection
- Pros: Produces a more interpretative model, prevents over fitting
- Cons: LASSO performs poorly when predictors are highly correlated



LASSO Metrics Plot

- Error metrics for both final testing data set (2025) and the training data split into training and testing sets
- Shrunk terms:
  - Plate appearances, home runs, RBIs, batting average, on base percentage, OPS+, rOBA



Boosting Models

- Boosting grows trees sequentially using information from previously grown trees
  - Each tree fit on a modified version of the original data set
  - Good at capturing non-linear patterns
- Pros: High predictive accuracy and captures complex, nonlinear relationships automatically.
- Cons: Prone to over fitting and requires careful tuning of hyper parameters to perform well.

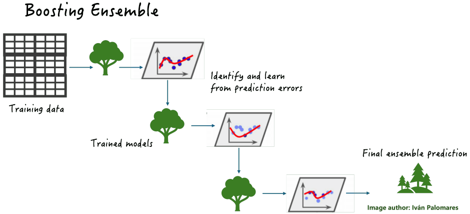
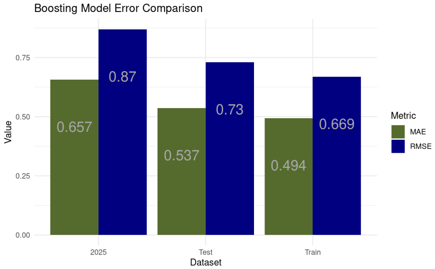


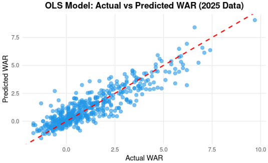
Image author: Iván Palomares

Boosting Metrics Plot

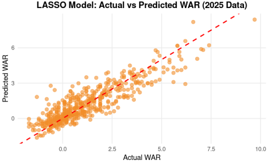
- Error metrics for both final testing data set (2025) and the training data split into training and testing sets



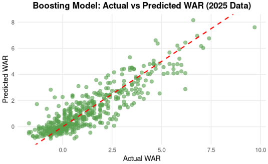
Model Comparison: OLS, LASSO, and Boosting



OLS Model: Actual vs Predicted WAR (2025 Data)

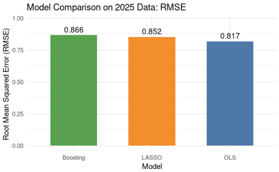


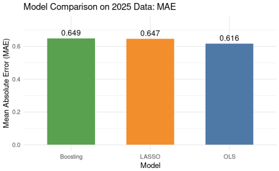
LASSO Model: Actual vs Predicted WAR (2025 Data)

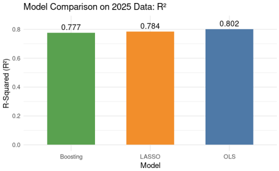


Boosting Model: Actual vs Predicted WAR (2025 Data)

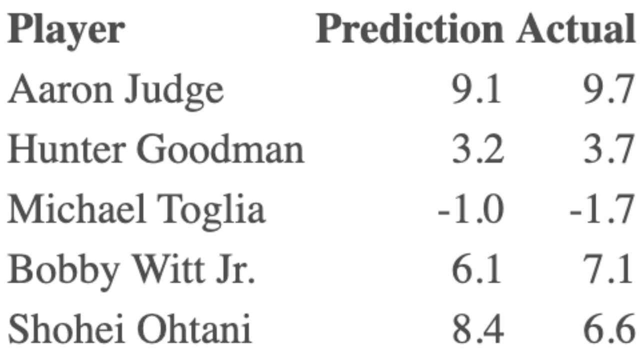
Model Comparison: RMSE, MAE, and R^2







Player Examples Using OLS model



Player	Prediction	Actual
Aaron Judge	9.1	9.7
Hunter Goodman	3.2	3.7
Michael Toglia	-1.0	-1.7
Bobby Witt Jr.	6.1	7.1
Shohei Ohtani	8.4	6.6

Key Findings

- OLS performed the best overall
  - Lowest prediction error
  - Highest explained variance
- LASSO selected a subset of meaningful predictors, making it easier to understand which player stats drive WAR
- Variables with a strong predictive value: plate appearances, home runs, hits, OPS+, walks, and strikouts
- WAR prediction is challenging
  - some components are hard to obtain from batting-only statistics
  - player injuries, playing time, or other external factors produce noise

Questions?