# Group 3: Predicting 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
- Interpreted as the number of additional wins a player adds to a team

# 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, strike outs, batting average, on base percentage, slugging percentage, OPS percentage, OPS+, rOBA, Rbat+, total bases, double plays grounded into, hit by pitch, sacrifice hits, sacrfice flies, intentional walks

# Models Used

- OLS
  - Baseline linear model
  - Benchmark for comparing models
- LASSO
  - Feature selection through L1 regularization
- Boosting
  - Tree based method
  - Good at capturing non-linear patterns

# OLS Models

- OLS identifies and measures the relationship between a response variable and predictor variables.
  - Finds a best-fitting line through a set of data points
- Pros: Convenient, 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

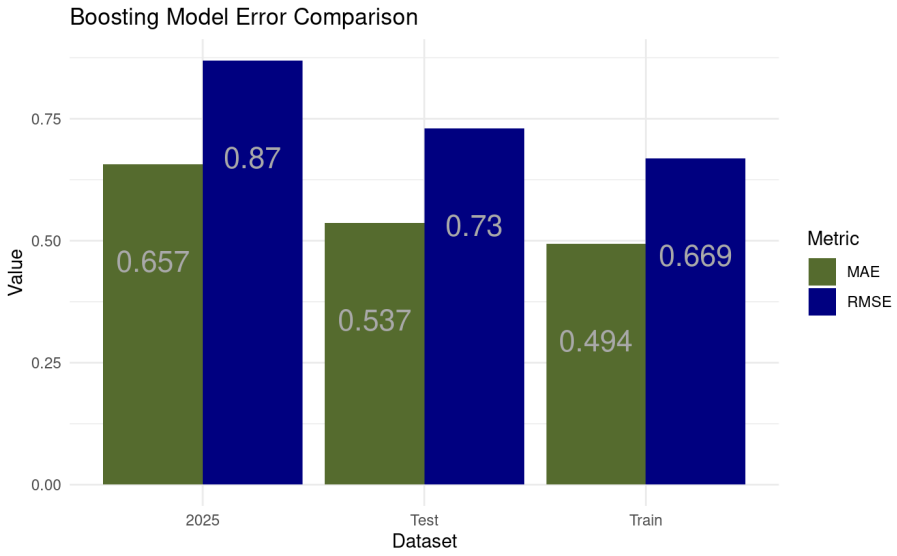
- Error metrics for both final testing data set (2025) and the training data split into training and testing sets
- Shrunk terms:
  - Plate appearances, home runs, RBIs, batting average, on base percentage, OPS+, rOBA

# Boosting Models

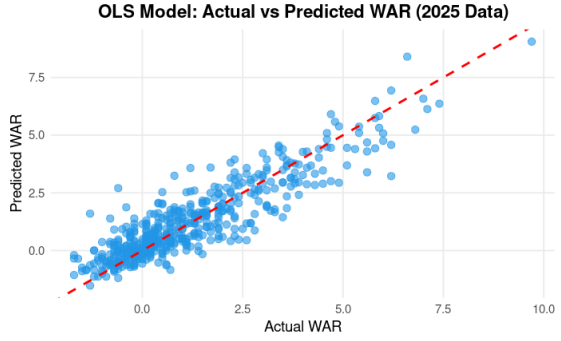
- Boosting grows trees sequentially using information from previously grown trees
  - Each tree fit on a modified version of the original data set
- Pros: High predictive accuracy and captures complex, nonlinear relationships automatically.
- Cons: Prone to overfitting and requires careful tuning of hyperparameters to perform well.

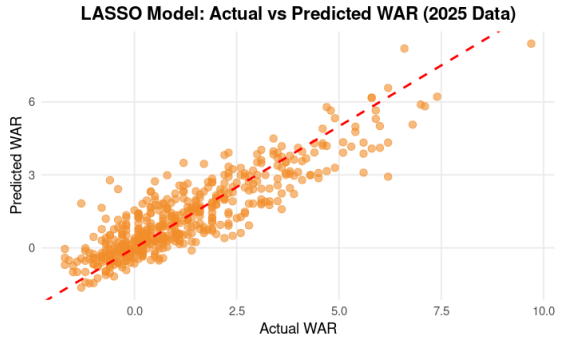
# Boosting Metrics Plot

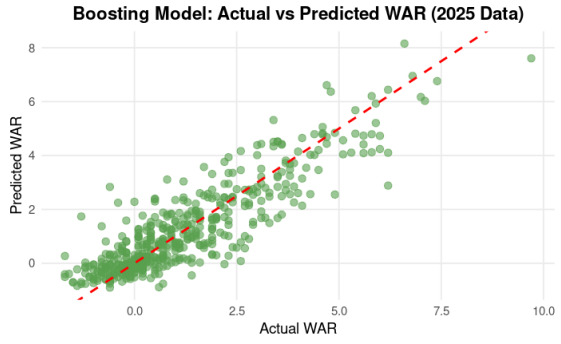
- Error metrics for both final testing data set (2025) and the training data split into training and testing sets



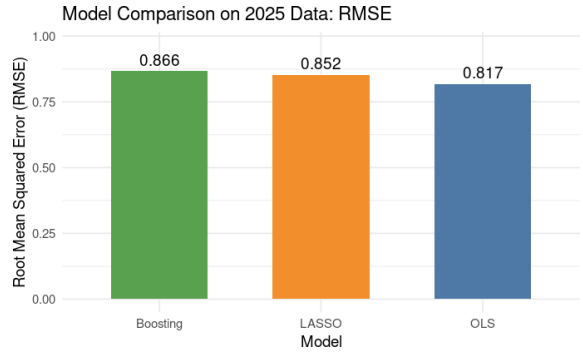
# Model Comparison: OLS, LASSO, and Boosting

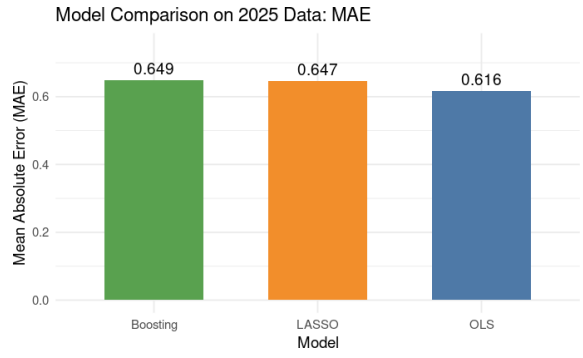


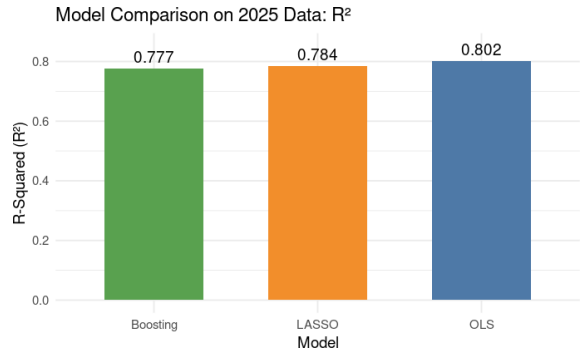



# Model Comparison: RMSE, MAE, and $R^2$



# Player Examples Using OLS model

| Player | Prediction | Actual |
|---|---|---|
| Aaron Judge | 9.051733 | 9.7 |
| Hunter Goodman | 3.174746 | 3.7 |
| Michael Toglia | -1.032315 | -1.7 |
| Bobby Witt Jr. | 6.125421 | 7.1 |

# Questions?