



Predicting MLB Player Value

Linear Regression and Web Scraping
August 2022

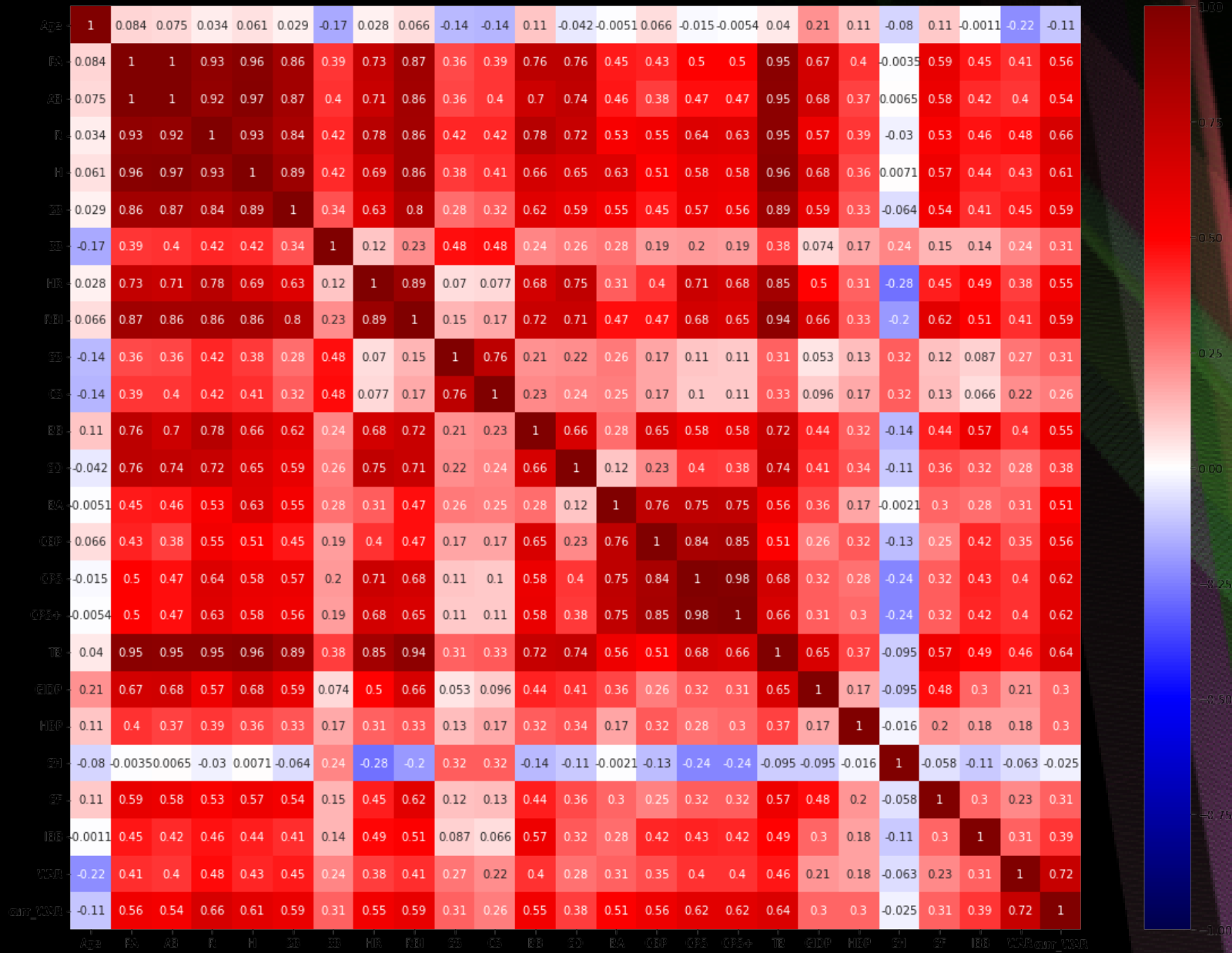
Background

- MLB player value metric: Wins Above Replacement
- Derived from net runs added
- Can we predict next-season WAR using current-season WAR?

Dataset

- 2016, 2017, 2018 full-season batting statistics
- 2017, 2018, 2019 WAR (“next-season”)
- baseball-reference.com
- 23 features, 1 dependent variable (next season WAR)

Multicollinearity



Model Development

- Train/Test split, Train/Val split: 60/20/20
train/val/holdout
- Feature scaling
- Polynomial and linear models
- RidgeCV, LassoCV, ElasticNetCV
- Reduced feature selection improves R^2 on val

Final Model: Ridge

Features (Scaled)	Coefficient
Current Year WAR	1.37
Age	-0.28
Was Current Year WAR > 4	0.01

- R-squared: .539
- Compare to naive model:
 - Next year WAR = Current year WAR
 - R-squared: .456

Future Improvements

- True time-series model based on past two or three years
 - Greatest residuals included players with “down” years
- Dummy variables: position, all-star