Predicting MLB Player Value

Linear Regression and Web Scraping

August 2022

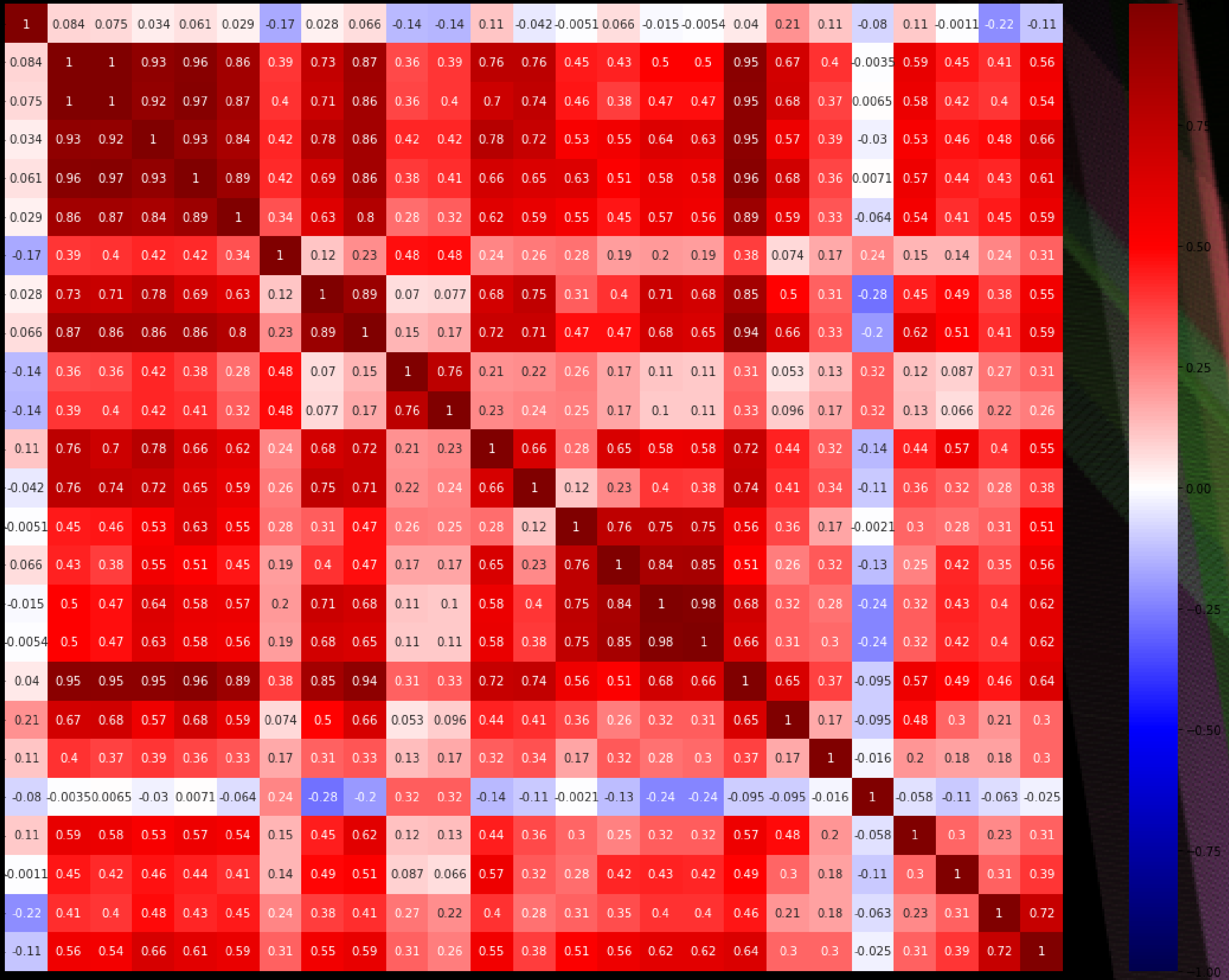
Background

- 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



1	0.084	0.075	0.034	0.061	0.029	-0.17	0.028	0.066	-0.14	-0.14	0.11	-0.042	-0.0051	0.066	-0.015	-0.0054	0.04	0.21	0.11	-0.08	0.11	-0.0011	-0.22	-0.11
0.084	1	1	0.93	0.96	0.86	0.39	0.73	0.87	0.36	0.39	0.76	0.76	0.45	0.43	0.5	0.5	0.95	0.67	0.4	-0.0035	0.59	0.45	0.41	0.56
0.075	1	1	0.92	0.97	0.87	0.4	0.71	0.86	0.36	0.4	0.7	0.74	0.46	0.38	0.47	0.47	0.95	0.68	0.37	0.0065	0.58	0.42	0.4	0.54
0.034	0.93	0.92	1	0.93	0.84	0.42	0.78	0.86	0.42	0.42	0.78	0.72	0.53	0.55	0.64	0.63	0.95	0.57	0.39	-0.03	0.53	0.46	0.48	0.66
0.061	0.96	0.97	0.93	1	0.89	0.42	0.69	0.86	0.38	0.41	0.66	0.65	0.63	0.51	0.58	0.58	0.96	0.68	0.36	0.0071	0.57	0.44	0.43	0.61
0.029	0.86	0.87	0.84	0.89	1	0.34	0.63	0.8	0.28	0.32	0.62	0.59	0.55	0.45	0.57	0.56	0.89	0.59	0.33	-0.064	0.54	0.41	0.45	0.59
-0.17	0.39	0.4	0.42	0.42	0.34	1	0.12	0.23	0.48	0.48	0.24	0.26	0.28	0.19	0.2	0.19	0.38	0.074	0.17	0.24	0.15	0.14	0.24	0.31
0.028	0.73	0.71	0.78	0.69	0.63	0.12	1	0.89	0.07	0.077	0.68	0.75	0.31	0.4	0.71	0.68	0.85	0.5	0.31	-0.28	0.45	0.49	0.38	0.55
0.066	0.87	0.86	0.86	0.86	0.8	0.23	0.89	1	0.15	0.17	0.72	0.71	0.47	0.47	0.68	0.65	0.94	0.66	0.33	-0.2	0.62	0.51	0.41	0.59
-0.14	0.36	0.36	0.42	0.38	0.28	0.48	0.07	0.15	1	0.76	0.21	0.22	0.26	0.17	0.11	0.11	0.31	0.053	0.13	0.32	0.12	0.087	0.27	0.31
-0.14	0.39	0.4	0.42	0.41	0.32	0.48	0.077	0.17	0.76	1	0.23	0.24	0.25	0.17	0.1	0.11	0.33	0.096	0.17	0.32	0.13	0.066	0.22	0.26
0.11	0.76	0.7	0.78	0.66	0.62	0.24	0.68	0.72	0.21	0.23	1	0.66	0.28	0.65	0.58	0.58	0.72	0.44	0.32	-0.14	0.44	0.57	0.4	0.55
-0.042	0.76	0.74	0.72	0.65	0.59	0.26	0.75	0.71	0.22	0.24	0.66	1	0.12	0.23	0.4	0.38	0.74	0.41	0.34	-0.11	0.36	0.32	0.28	0.38
-0.0051	0.45	0.46	0.53	0.63	0.55	0.28	0.31	0.47	0.26	0.25	0.28	0.12	1	0.76	0.75	0.75	0.56	0.36	0.17	-0.0021	0.3	0.28	0.31	0.51
0.066	0.43	0.38	0.55	0.51	0.45	0.19	0.4	0.47	0.17	0.17	0.65	0.23	0.76	1	0.84	0.85	0.51	0.26	0.32	-0.13	0.25	0.42	0.35	0.56
-0.015	0.5	0.47	0.64	0.58	0.57	0.2	0.71	0.68	0.11	0.1	0.58	0.4	0.75	0.84	1	0.98	0.68	0.32	0.28	-0.24	0.32	0.43	0.4	0.62
-0.0054	0.5	0.47	0.63	0.58	0.56	0.19	0.68	0.65	0.11	0.11	0.58	0.38	0.75	0.85	0.98	1	0.66	0.31	0.3	-0.24	0.32	0.42	0.4	0.62
0.04	0.95	0.95	0.95	0.96	0.89	0.38	0.85	0.94	0.31	0.33	0.72	0.74	0.56	0.51	0.68	0.66	1	0.65	0.37	-0.095	0.57	0.49	0.46	0.64
0.21	0.67	0.68	0.57	0.68	0.59	0.074	0.5	0.66	0.053	0.096	0.44	0.41	0.36	0.26	0.32	0.31	0.65	1	0.17	-0.095	0.48	0.3	0.21	0.3
0.11	0.4	0.37	0.39	0.36	0.33	0.17	0.31	0.33	0.13	0.17	0.32	0.34	0.17	0.32	0.28	0.3	0.37	0.17	1	-0.016	0.2	0.18	0.18	0.3
-0.08	-0.0035	0.0065	-0.03	0.0071	-0.064	0.24	-0.28	-0.2	0.32	0.32	-0.14	-0.11	-0.0021	-0.13	-0.24	-0.24	-0.095	-0.095	-0.016	1	-0.058	-0.11	-0.063	-0.025
0.11	0.59	0.58	0.53	0.57	0.54	0.15	0.45	0.62	0.12	0.13	0.44	0.36	0.3	0.25	0.32	0.32	0.57	0.48	0.2	-0.058	1	0.3	0.23	0.31
-0.0011	0.45	0.42	0.46	0.44	0.41	0.14	0.49	0.51	0.087	0.066	0.57	0.32	0.28	0.42	0.43	0.42	0.49	0.3	0.18	-0.11	0.3	1	0.31	0.39
-0.22	0.41	0.4	0.48	0.43	0.45	0.24	0.38	0.41	0.27	0.22	0.4	0.28	0.31	0.35	0.4	0.4	0.46	0.21	0.18	-0.063	0.23	0.31	1	0.72
-0.11	0.56	0.54	0.66	0.61	0.59	0.31	0.55	0.59	0.31	0.26	0.55	0.38	0.51	0.56	0.62	0.62	0.64	0.3	0.3	-0.025	0.31	0.39	0.72	1

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 R2 on val

Final Model: Ridge

Features (Scaled)	Coefficient
Current Year WAR	0.85
Age	-0.40
Doubles	0.39
Hits	-0.34
Runs	.30

- R-squared: .756
- Compare to naive model:
  - Next year WAR = Current year WAR
  - R-squared: .159

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