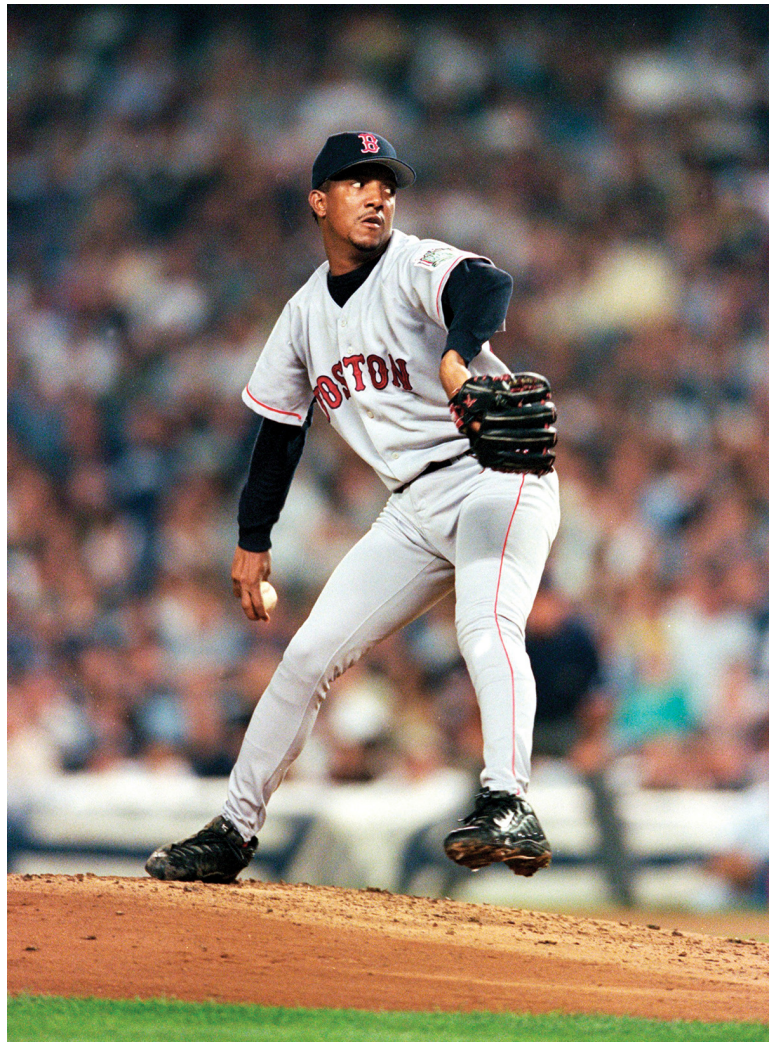
Project Pedro

Andrew Seo



Pedro Martinez – Boston Red Sox. Record holder for highest ERA+ in a single season

Introduction

- "Defense wins championships"
- Question: What are the elements of the best pitchers (best = defined by highest ERA+)?
- How: Web scrape data from Baseball Reference and use Regression to determine most influential features

ERA+ Explained

- Regular ERA: A metric to measure how many runs a pitcher allows per appearance, standardized to 9 innings.
  - Lower = Better
- ERA+: Higher = Better
- Example: Corbin Burnes has a 185 ERA+, which means that the league ERA is 85% higher than Burnes.

$$ERA+ = 100 \cdot \frac{lgERA}{ERA} \cdot PF$$

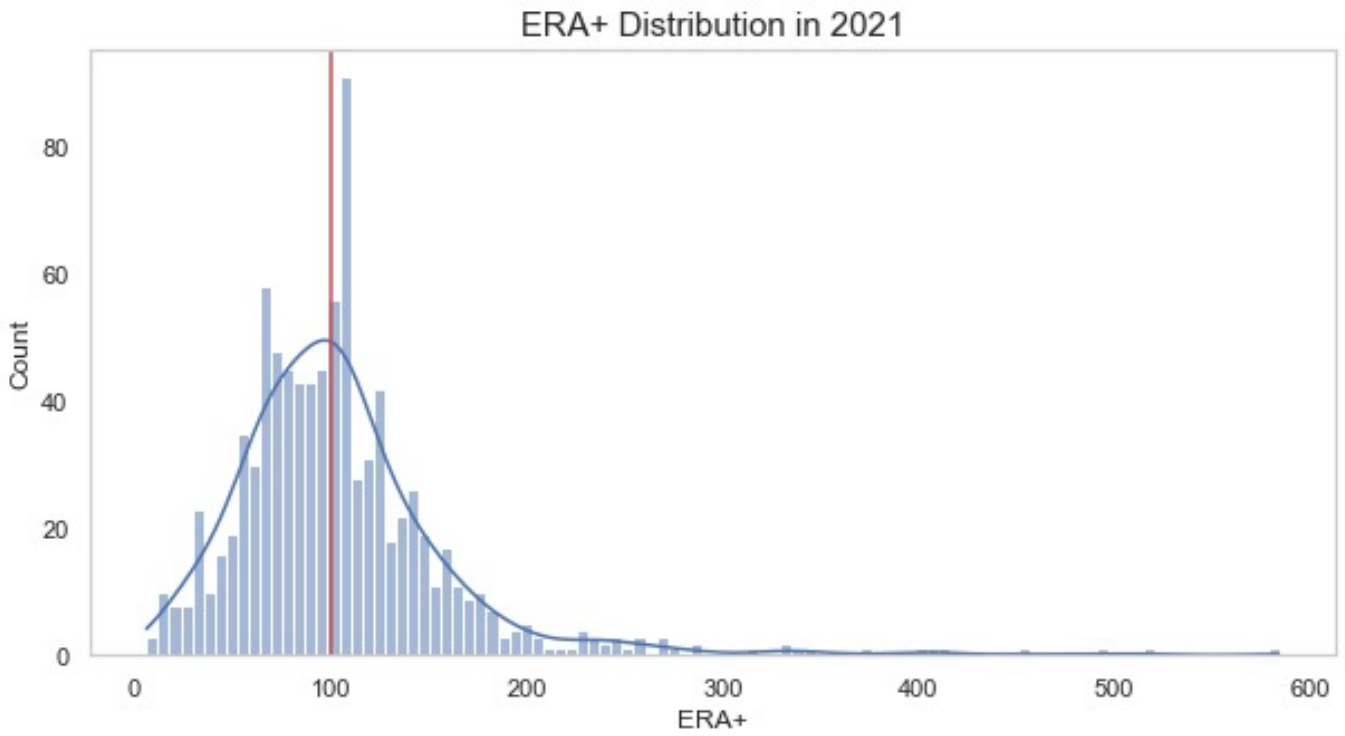
Methodology: Data

Roughly 900 unique rows representing all pitchers from this 2021 season.

				Batting Against					Pitching Ratios			Batted Ball						Win Probability		
Rk	Name	Age	Tm	BA	OBP	SLG	OPS	BAbip	HR%	SO%	BB%	EV	HardH%	LD%	GB%	FB%	GB/FB	WPA	cWPA	RE24
1	Fernando Abad*	35	BAL	.292	.354	.444	.799	.328	1.3%	12.7%	8.9%	86.6	35.5%	24.2%	50.0%	17.7%	2.82	0.0	0.0%	0.9
2	Cory Abbott	25	CHC	.308	.410	.673	1.083	.282	8.1%	12.9%	14.5%	91.4	48.9%	17.8%	40.0%	31.1%	1.29	0.0	0.0%	-6.2
3	Albert Abreu	25	NYY	.203	.314	.438	.751	.205	5.2%	22.9%	12.4%	83.4	28.4%	22.9%	44.8%	27.1%	1.65	0.8	0.6%	-2.0
4	Bryan Abreu	24	HOU	.254	.348	.406	.754	.310	2.5%	22.4%	11.2%	89.7	40.4%	19.2%	48.1%	25.0%	1.92	-0.8	-0.6%	-8.7
5	Domingo Acevedo	27	OAK	.188	.257	.406	.663	.190	5.7%	25.7%	8.6%	85.1	34.8%	13.0%	52.2%	30.4%	1.71	-0.1	-0.1%	1.6

Rk	Name	Age	Tm	Lg	W	L	W-L%	ERA	G	GS	GF	CG	SHO	SV	IP	H	R	ER	HR	BB	IBB	SO	HBP	BK	WP	BF	ERA+	FIP	WHIP	H9	HR9	BB9	SO9	SO/W
1	Corbin Burnes	26	MIL	NL	11	4	.733	2.29	27	27	0	0	0	0	165.0	121	44	42	6	33	0	230	6	0	5	648	185	1.55	0.933	6.6	0.3	1.8	12.5	6.97
2	Brandon Woodruff	28	MIL	NL	9	10	.474	2.56	30	30	0	0	0	0	179.1	130	54	51	18	43	0	211	7	0	2	708	166	2.95	0.965	6.5	0.9	2.2	10.6	4.91
3	Max Scherzer	36	TOT	NL	15	4	.789	2.46	30	30	0	1	0	0	179.1	119	53	49	23	36	0	236	10	0	2	693	164	2.97	0.864	6.0	1.2	1.8	11.8	6.56
4	Robbie Ray*	29	TOR	AL	13	6	.684	2.68	31	31	0	0	0	0	188.0	146	57	56	29	49	0	244	4	0	5	750	163	3.42	1.037	7.0	1.4	2.3	11.7	4.98
5	Walker Buehler	26	LAD	NL	15	4	.789	2.49	32	32	0	0	0	0	202.2	146	60	56	19	51	2	201	5	0	5	795	162	3.23	0.972	6.5	0.8	2.3	8.9	3.94

Methodology: ERA+

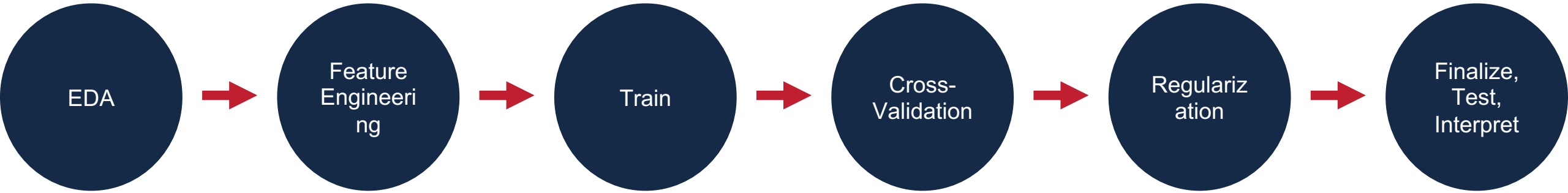


Initial Baseline Linear Regression poor

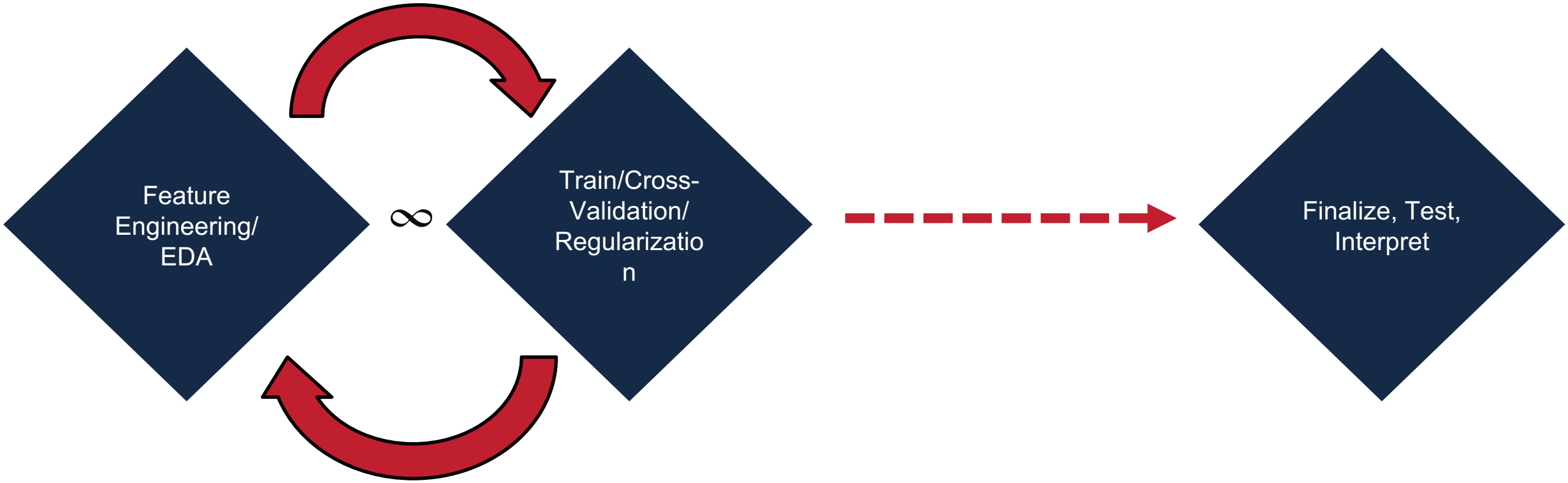
Model being influenced by outliers

Fix: Remove ERA+ over 300

Methodology: Expectation



Methodology: Reality

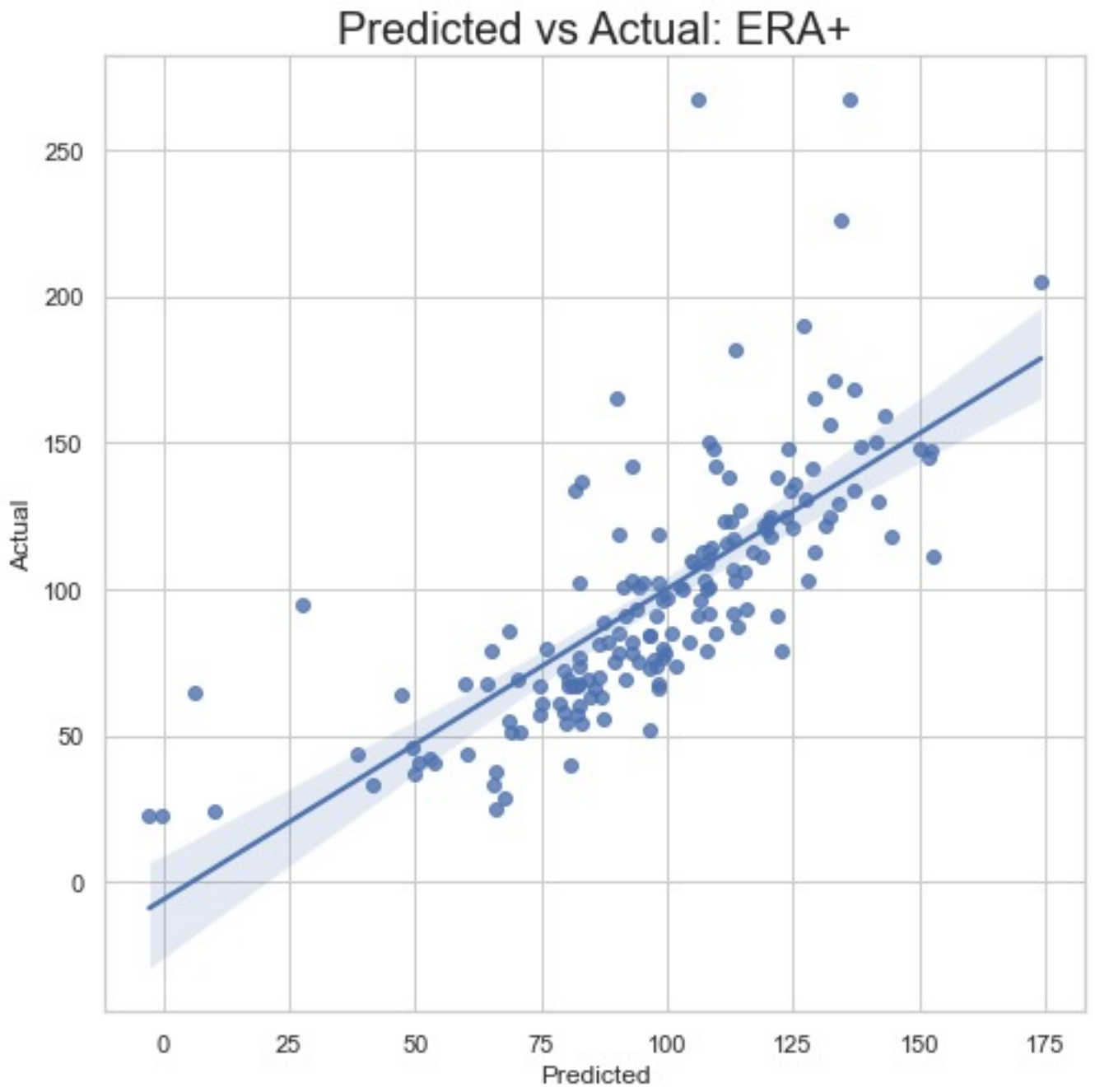


Results

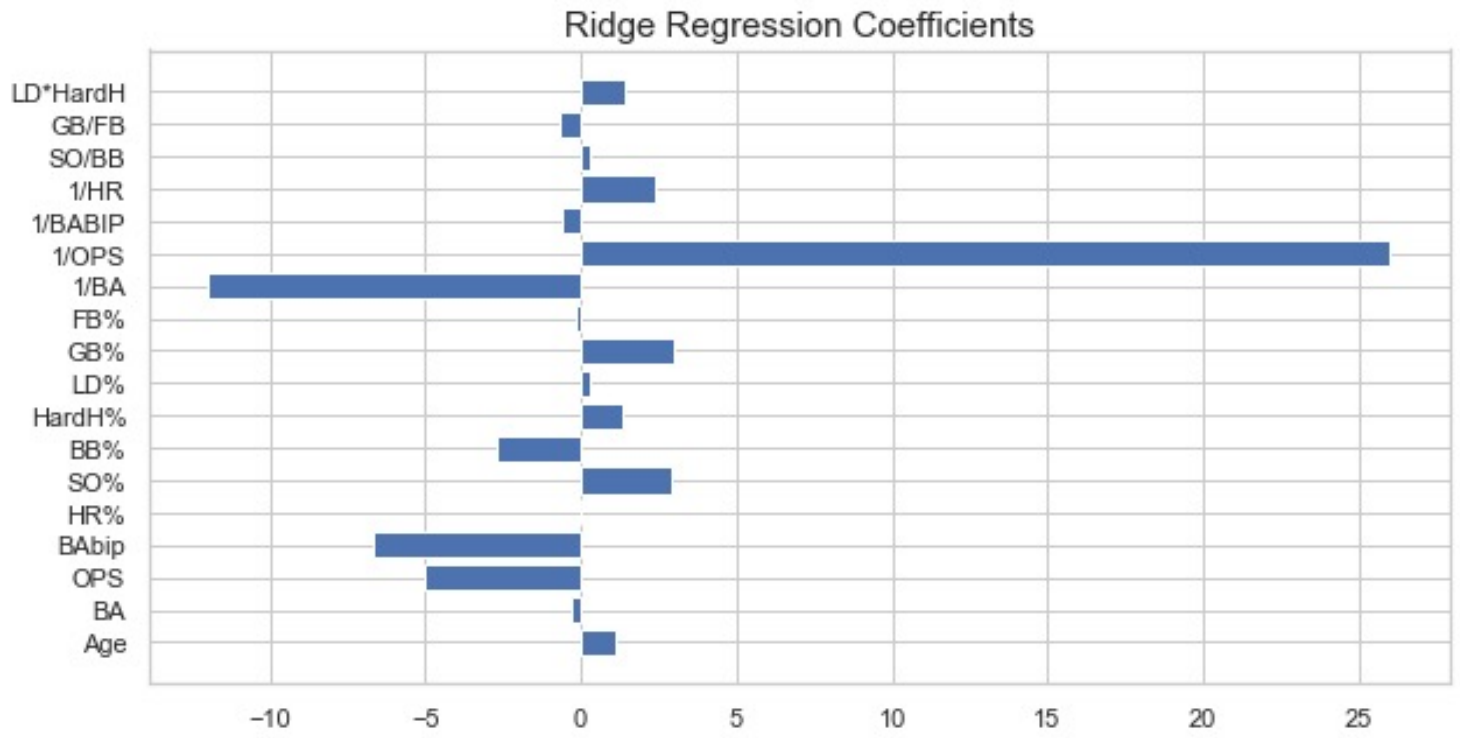
R^2: 0.48 → 0.54

MAE: 24.29→ 19.61

Simple Linear, Lasso, Ridge produced comparable results



Feature Deep Dive




```
Ridge Coefficients
Age : 1.14
BA : -0.32
OPS : -5.05
BAbip : -6.7
HR% : -0.03
SO% : 2.88
BB% : -2.73
HardH% : 1.34
LD% : 0.27
GB% : 3.0
FB% : -0.13
1/BA : -11.98
1/OPS : 26.01
1/BABIP : -0.58
1/HR : 2.42
SO/BB : 0.28
GB/FB : -0.69
LD*HardH : 1.42
```

Insights and Conclusions

Minimizing BABIP and OPS while Maximizing Strikeouts and Ground Balls contributes to Higher ERA+

Looking Ahead

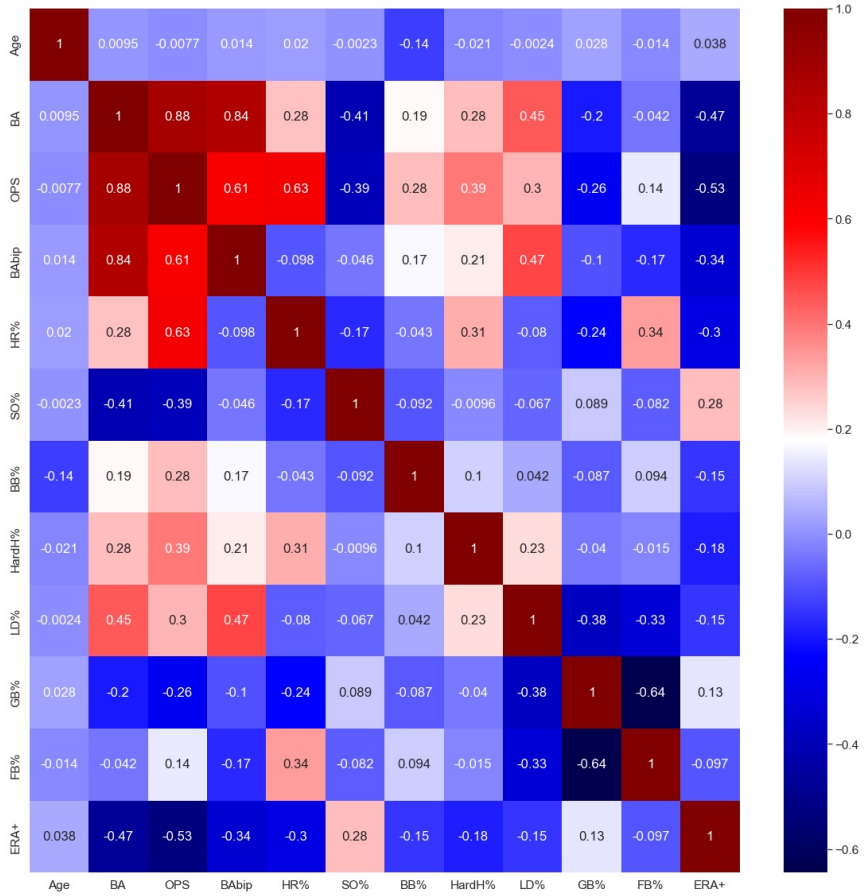
More features: pitch mix (FB%, Curve%), pitch location (In/Out of Zone)

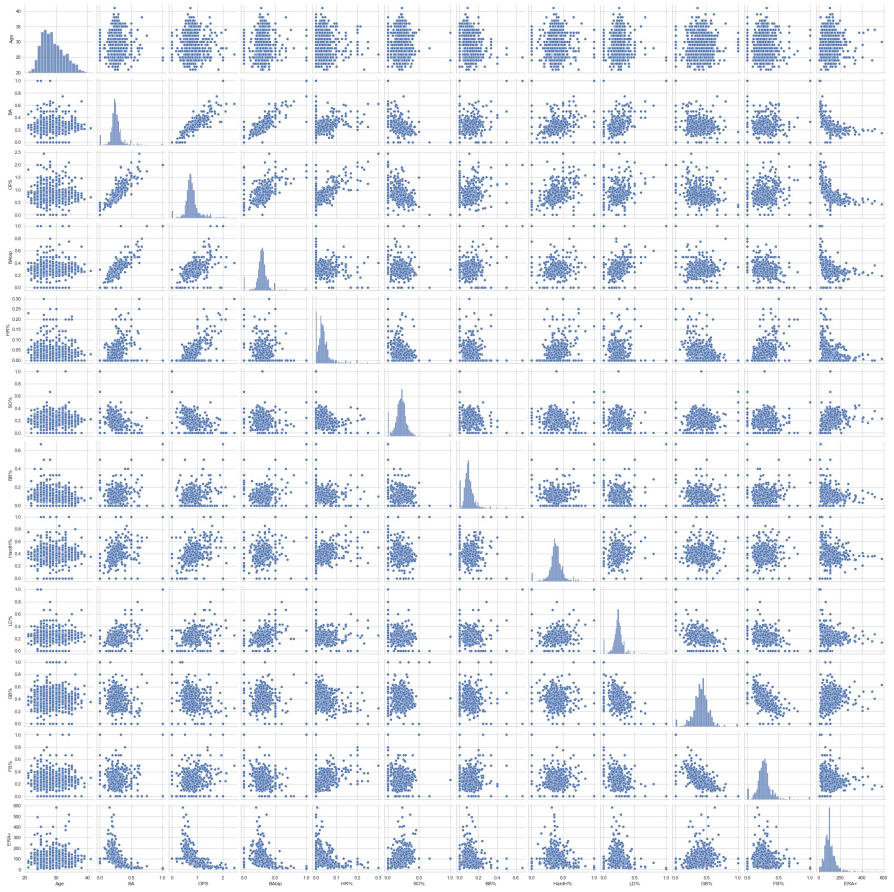
Greater dataset: ERA+ should be comparable season/season

Different target: FIP (fielding independent pitching)

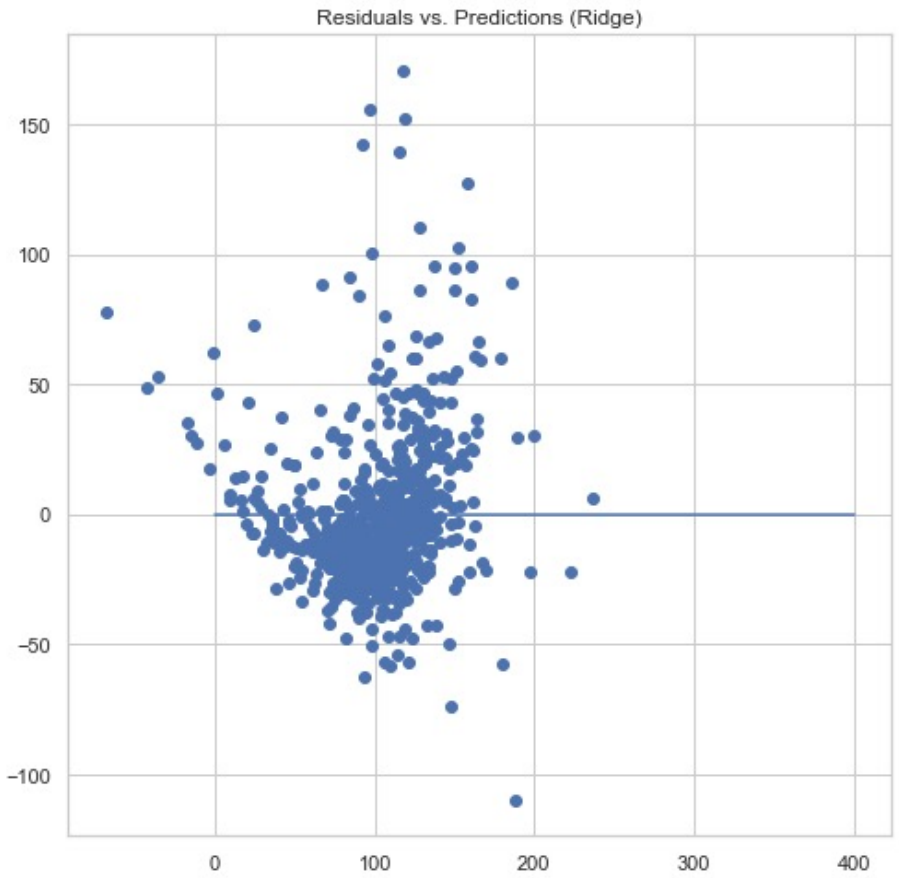
Questions?

Appendix

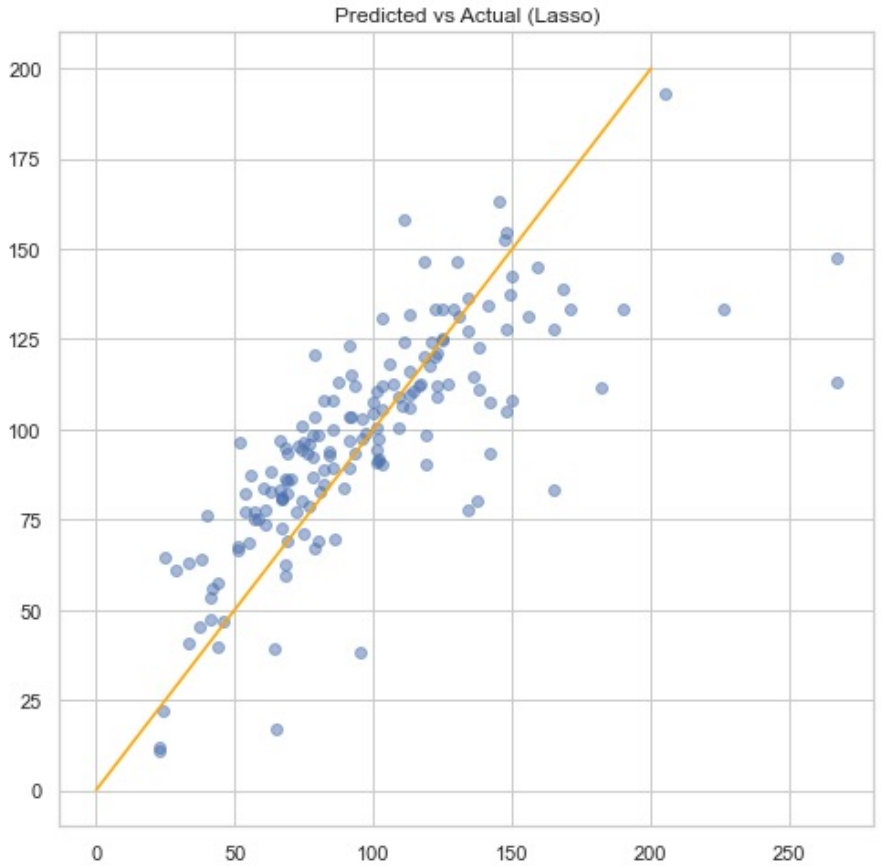

Age 1 0.0095 -0.0077 0.014 0.02 -0.0023 -0.14 -0.021 -0.0024 0.028 -0.014 0.038
BA 0.0095 1 0.88 0.84 0.28 -0.41 0.19 0.28 0.45 -0.2 -0.042 -0.47
OPS -0.0077 0.88 1 0.61 0.63 -0.39 0.28 0.39 0.3 -0.26 0.14 -0.53
BAbip 0.014 0.84 0.61 1 -0.098 -0.046 0.17 0.21 0.47 -0.1 -0.17 -0.34
HR% 0.02 0.28 0.63 -0.098 1 -0.17 -0.043 0.31 -0.08 -0.24 0.34 -0.3
SO% -0.0023 -0.41 -0.39 -0.046 -0.17 1 -0.092 -0.0096 -0.067 0.089 -0.082 0.28
BB% -0.14 0.19 0.28 0.17 -0.043 -0.092 1 0.1 0.042 -0.087 0.094 -0.15
HardH% -0.021 0.28 0.39 0.21 0.31 -0.0096 0.1 1 0.23 -0.04 -0.015 -0.18
LD% -0.0024 0.45 0.3 0.47 -0.08 -0.067 0.042 0.23 1 -0.38 -0.33 -0.15
GB% 0.028 -0.2 -0.26 -0.1 -0.24 0.089 -0.087 -0.04 -0.38 1 -0.64 0.13
FB% -0.014 -0.042 0.14 -0.17 0.34 -0.082 0.094 -0.015 -0.33 -0.64 1 -0.097
ERA+ 0.038 -0.47 -0.53 -0.34 -0.3 0.28 -0.15 -0.18 -0.15 0.13 -0.097 1
Age BA OPS BAbip HR% SO% BB% HardH% LD% GB% FB% ERA+
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6

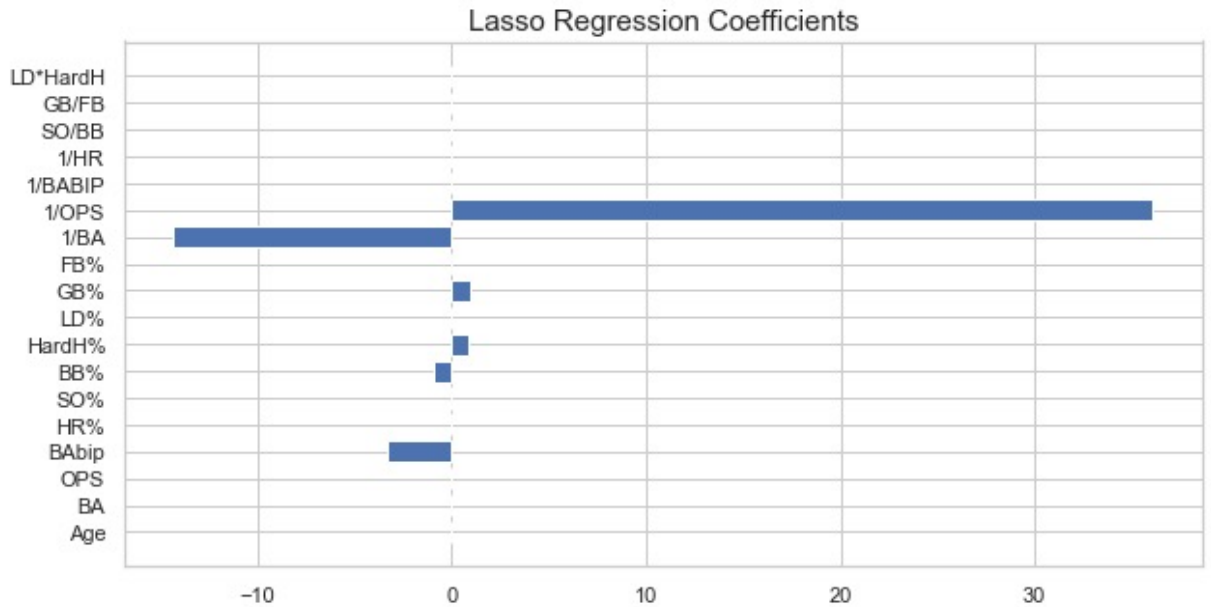


Appendix: Cont.



Appendix: Cont.





Glossary

- BA (Batting Average): Hits allowed/Batters faced
- On-Base %: % of batters allowed to reach base (besides errors)
- Slugging %: Total bases allowed/Batters faced
- OPS: On-Base % + Slugging %
- BAbip (Batting Average on Balls In Play): Batting average on all balls that are put into the field of play (no HRs, strikeouts)
- HR% (Home Run %): Home runs allowed/Batters faced
- SO% (Strikeout %): Strikeouts/Batters faced
- BB% (Walk %): Walks Issued/Batters faced
- HardH% (Hard Hit %): Percentage of balls in play with exit % of 95 mph or greater
- LD (Line Drive %): Percentage of balls put into play that are line drives
- GB (Ground Ball %): Percentage of balls put into play that are ground balls
- FB% (Fly Ball %): Percentage of balls put into play that are fly balls