Moneyball

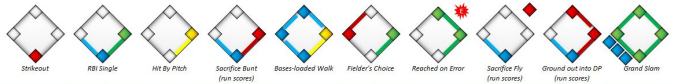
An investigation of MLB players' batting efficiency

Background

Baseball Batting Statistics Illustrated

SPORTCHART.WORDPRESS.COM

Calculating batter statistics using 10 hypothetical plate appearances



Basic Stats												
Plate Appearances	10	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
At Bats	6	✓	✓	✗	✗	✗	✓	✓	✗	✓	✓	
Hits	2	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	
Extra-Base Hits	1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	
Runs	1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	Batter my score subsequent runs as base runner
RBIs	8	✗	1	✗	1	1	✗	✗	1	✗	4	
Home Runs	1	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	
Times on Base	4	✗	✓	✓	✗	✓	✗	✗	✗	✗	✓	
Total Bases	5	✗	1	✗	✗	✗	✗	✗	✗	✗	4	

Ratios												
Batting Average	.333	0/1	1/1				0/1	0/1		0/1	1/1	2/6	Hits / AB
On Base %	.444	0/1	1/1	1/1		1/1	0/1	0/1	0/1	0/1	1/1	4/9	(H+BB+HBP) / (AB+BB+HBP+SF)
Slugging %	.833	0/1	1/1				0/1	0/1		0/1	4/1	5/6	Total Bases / AB
Runs Created	2.143	0x5/1	1x5/1			1x5/1	0x5/1	0x5/1		0x5/1	1x5/1	15/7	((Hits+BB) x Total Bases) / (AB+Walks)
BA on Balls in Play	.200		1/1				0/1	0/1	0/1	0/1		1/5	(Hits-HR) / (AB-K-HR+SF)
AB Per Home Run	6.0	1/0	1/0				1/0	1/0		1/0	1/1	6/1	AB / HR
PA Per Strikeout	10.0	1/1	1/0	1/0	1/0	1/0	1/0	1/0	1/0	1/0	1/0	10/1	PA / K
BB / K Ratio	1.0	0/1				1/0						1/1	Walks / K

Source: MLB.com

Player Standard Batting

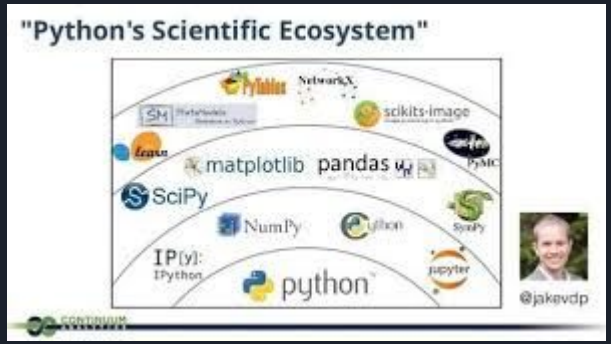
Share & more ▾ | Hide non-qualifiers for rate stats (min. 3.1 PA/G(lgAvg)) | Glossary | Hide Partial Rows

Rk	Name	Age	Tm	Lg	G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB	Pos Summary
1	Fernando Abad*	33	SFG	NL	18	0	0	0	0	0	0	0	0	0	0	0	0						0	0	0	0	0	0	1
2	José Abreu	32	CHW	AL	159	693	634	85	180	38	1	33	123	2	2	36	152	.284	.330	.503	.834	118	319	24	13	0	10	4	*3D
3	Ronald Acuna Jr.	21	ATL	NL	156	715	626	127	175	22	2	41	101	37	9	76	188	.280	.365	.518	.883	121	324	8	9	0	1	4	*879/H
4	Jason Adam	27	TOR	AL	3	0	0	0	0	0	0	0	0	0	0	0	0						0	0	0	0	0	0	1
5	Cristhian Adames#	27	SFG	NL	10	24	22	1	7	1	0	0	2	0	0	2	8	.318	.375	.364	.739	99	8	0	0	0	0	0	/H45
6	Willy Adames	23	TBR	AL	152	584	531	69	135	25	1	20	52	4	2	46	153	.254	.317	.418	.735	96	222	9	3	3	1	1	*6/H
7	Austin Adams	32	TOT	AL	1	0	0	0	0	0	0	0	0	0	0	0	0						0	0	0	0	0	0	1
8	Austin Adams	32	DET	AL	1	0	0	0	0	0	0	0	0	0	0	0	0						0	0	0	0	0	0	1
9	Austin Adams	28	TOT	MLB	4	1	1	0	0	0	0	0	0	0	0	0	1	.000	.000	.000	.000	-100	0	0	0	0	0	0	1
10	Austin Adams	28	WSN	NL	1	0	0	0	0	0	0	0	0	0	0	0	0						0	0	0	0	0	0	/1
11	Austin Adams	28	SEA	AL	3	1	1	0	0	0	0	0	0	0	0	0	1	.000	.000	.000	.000	-100	0	0	0	0	0	0	1
12	Matt Adams*	30	WSN	NL	111	333	310	42	70	14	0	20	56	0	0	20	115	.226	.276	.465	.741	86	144	7	2	0	1	1	3H
13	Jim Adduci*	34	CHC	NL	2	5	5	0	0	0	0	0	0	0	0	0	3	.000	.000	.000	.000	-100	0	0	0	0	0	0	/H9
14	Ehire Adrianza#	29	MIN	AL	83	236	202	34	55	8	3	5	22	0	2	20	40	.272	.349	.416	.765	103	84	2	6	2	4	1	563H/4971
15	Dario Agrazal	24	PIT	NL	14	24	22	0	2	0	0	0	2	0	0	0	9	.091	.087	.091	.178	-53	2	0	0	1	1	0	1
16	Jesus Aguilar	29	TOT	MLB	131	369	314	39	74	12	0	12	50	0	0	43	81	.236	.325	.389	.714	87	122	12	2	0	7	0	3HD/5
17	Jesus Aguilar	29	MIL	NL	94	262	222	26	50	9	0	8	34	0	0	31	59	.225	.320	.374	.694	80	83	11	2	0	4	0	3H/5
18	Jesus Aguilar	29	TBR	AL	37	107	92	13	24	3	0	4	16	0	0	12	22	.261	.336	.424	.760	104	39	1	0	0	3	0	3D/H
19	Nick Ahmed	29	ARI	NL	158	625	556	79	141	33	6	19	82	8	2	52	113	.254	.316	.437	.753	92	243	15	4	1	12	2	*6/H
20	R.J. Alaniz	28	TOT	MLB	8	1	1	0	1	0	0	0	1	0	0	0	0	1.000	1.000	1.000	2.000	417	1	0	0	0	0	0	1
21	R.J. Alaniz	28	CIN	NL	8	1	1	0	1	0	0	0	1	0	0	0	0	1.000	1.000	1.000	2.000	417	1	0	0	0	0	0	/1
22	Matt Albers*	36	MIL	NL	63	0	0	0	0	0	0	0	0	0	0	0	0						0	0	0	0	0	0	1
23	Hanser Alberto	26	BAL	AL	139	550	524	62	160	21	2	12	51	4	4	16	50	.305	.329	.422	.751	98	221	9	4	3	3	1	45/H7D19
24	Ozzie Albies#	22	ATL	NL	160	702	640	102	189	43	8	24	86	15	4	54	112	.295	.352	.500	.852	113	320	2	4	0	4	6	*4/H

Workflow and Tools

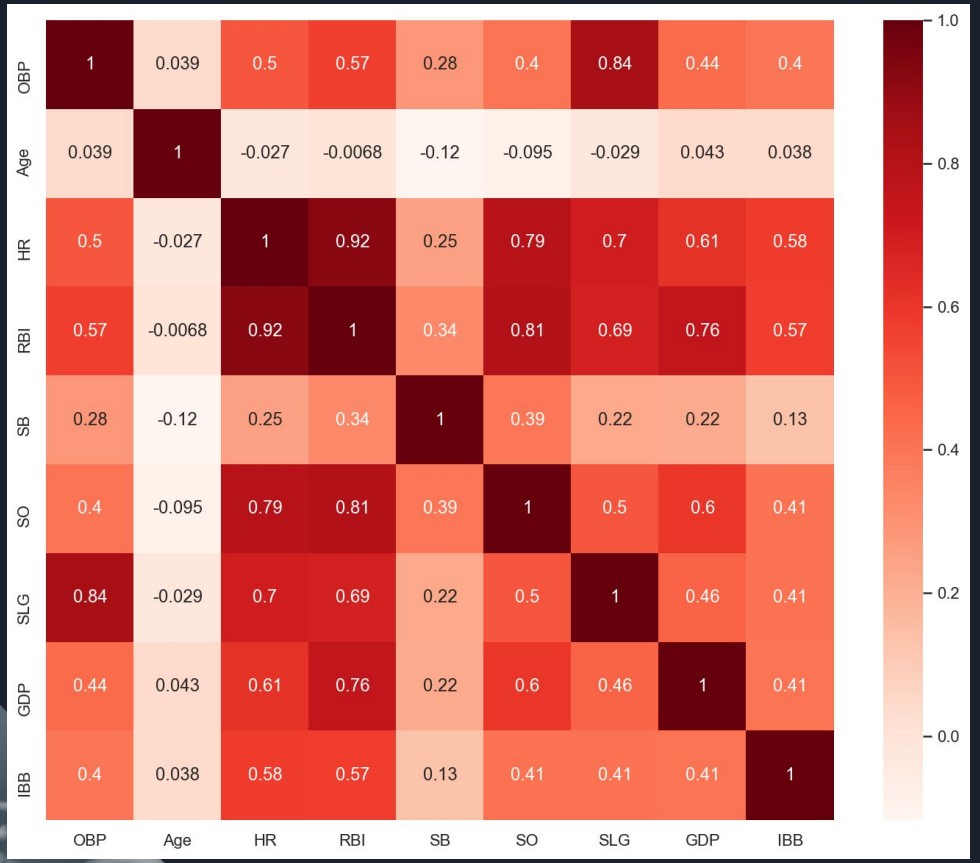
Webscrape & Clean Data → **Feature Engineering & Selection** → **Train & Evaluate Model (Model Selection)**

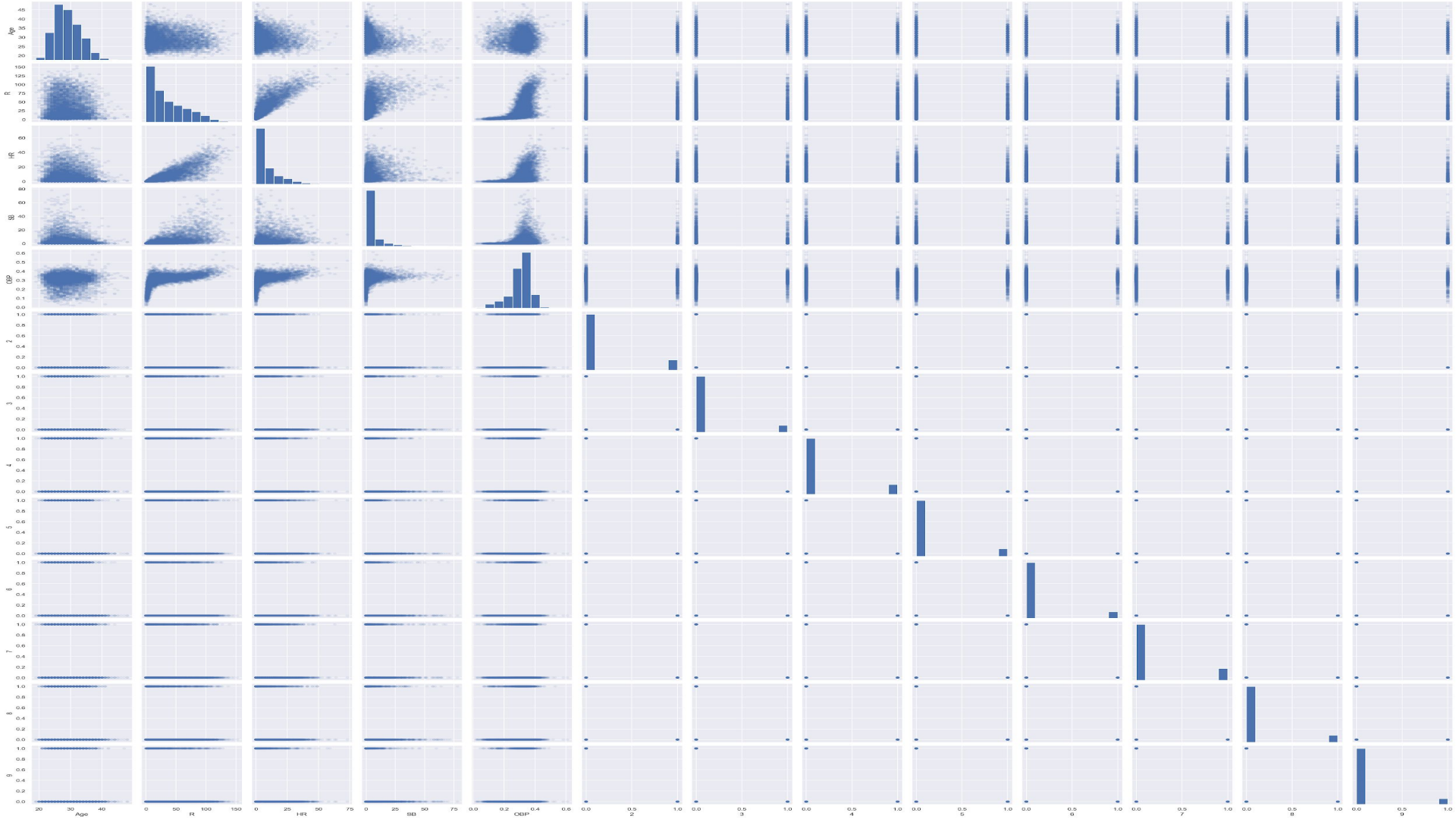
- Beautifulsoup : Web-scraping
- Pandas : Dataframes
- Numpy : Transforming data
- Sklearn : Ridge/LASSO/ElasticNet
- Statsmodels : OLS for stats summary
- Matplotlib & Seaborn : Plot graphs

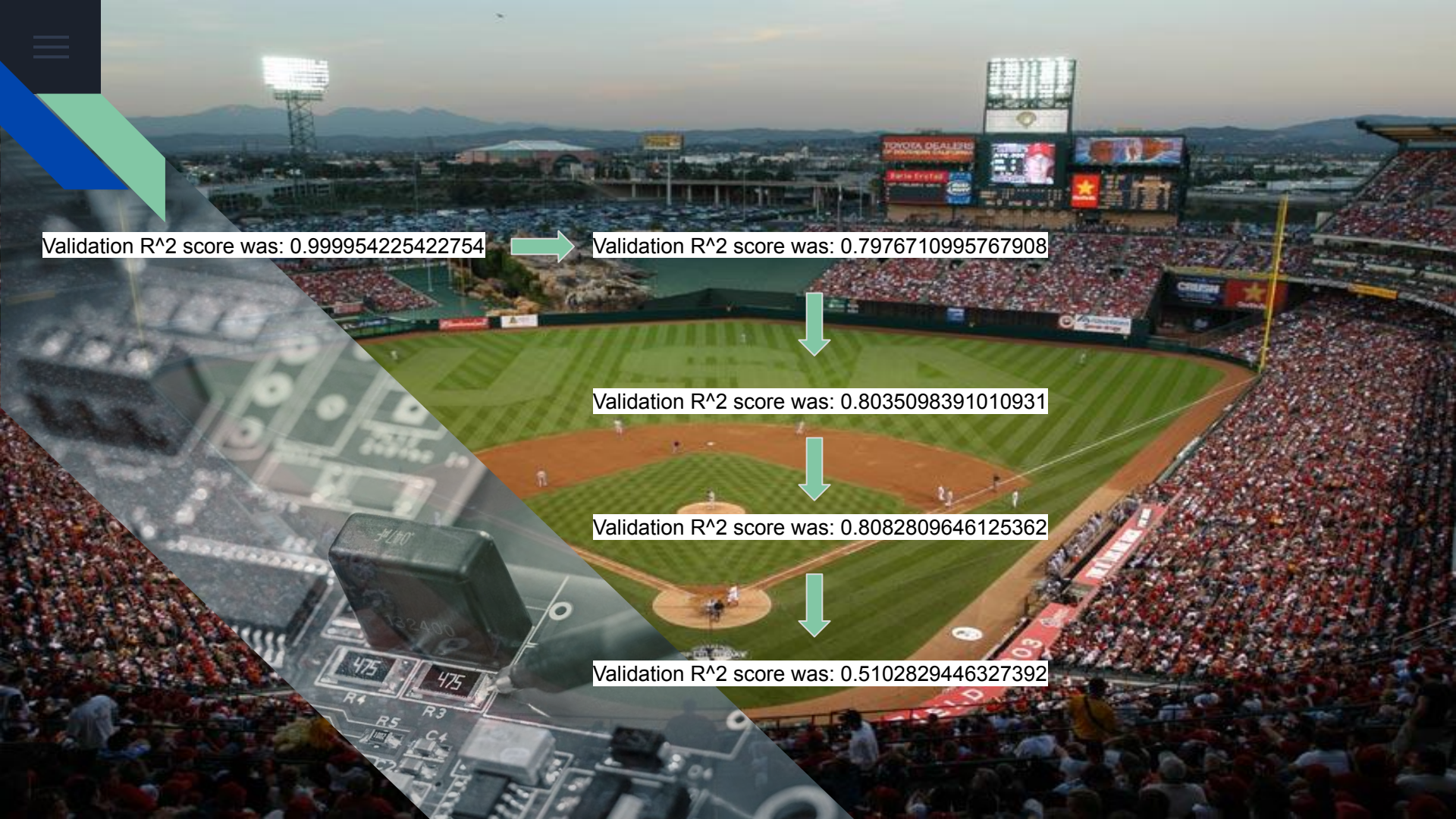


Goals:

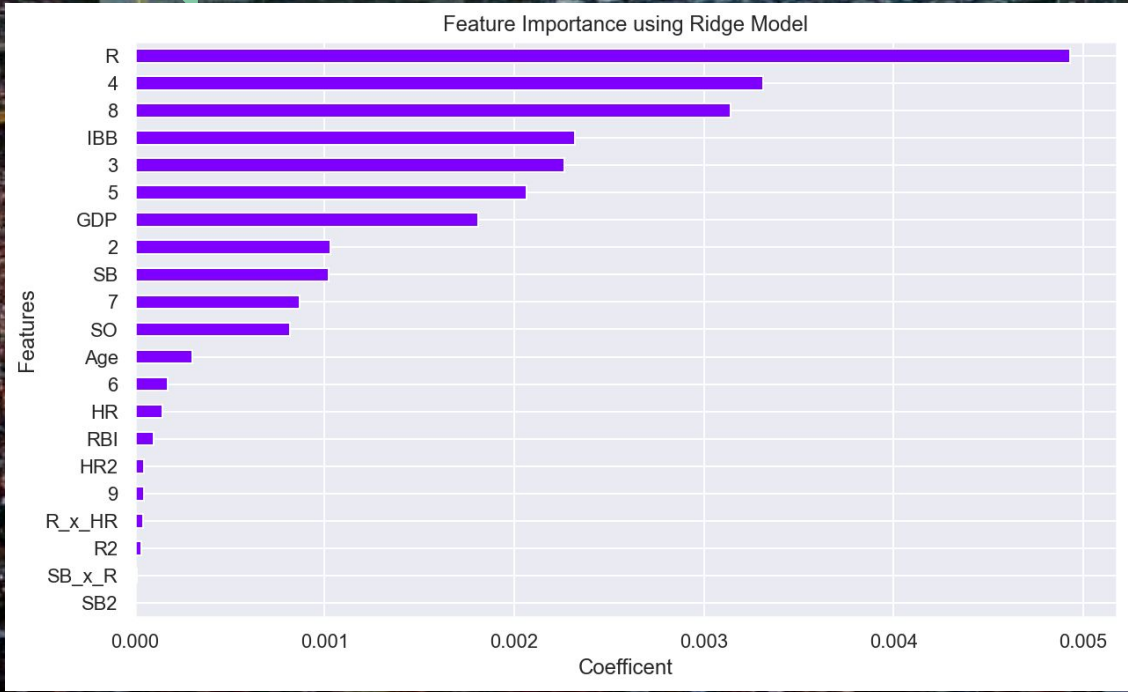
- Establish connection between features
- Use log expressions to eliminate impeding features

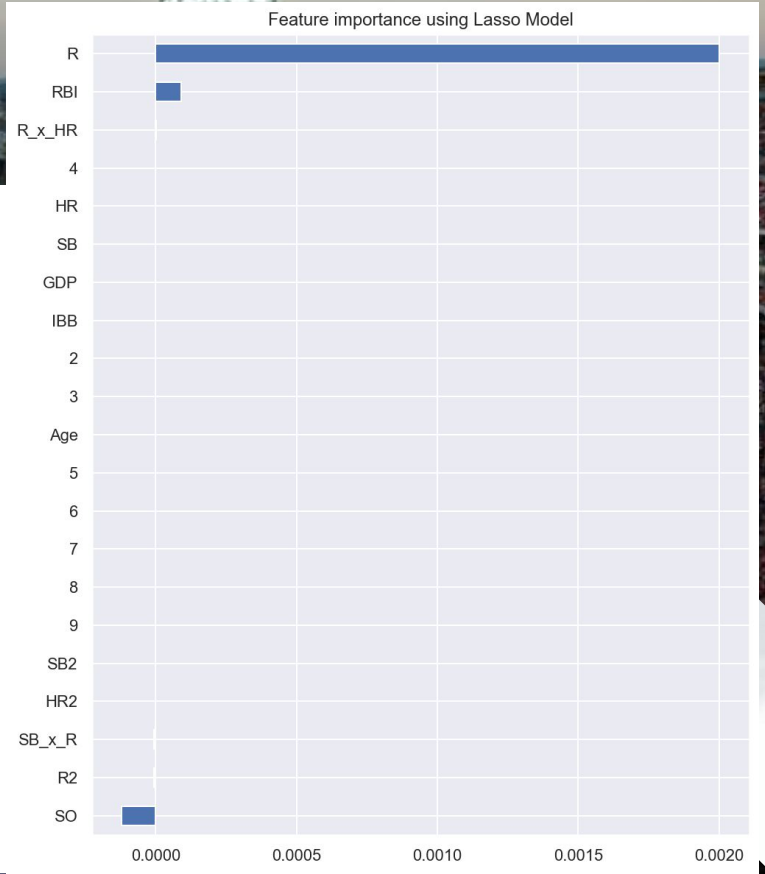



Age
R
HR
SB
OBP
2
3
4
5
6
7
8
9


Validation R^2 score was: 0.999954225422754
Validation R^2 score was: 0.7976710995767908
Validation R^2 score was: 0.8035098391010931
Validation R^2 score was: 0.8082809646125362
Validation R^2 score was: 0.5102829446327392

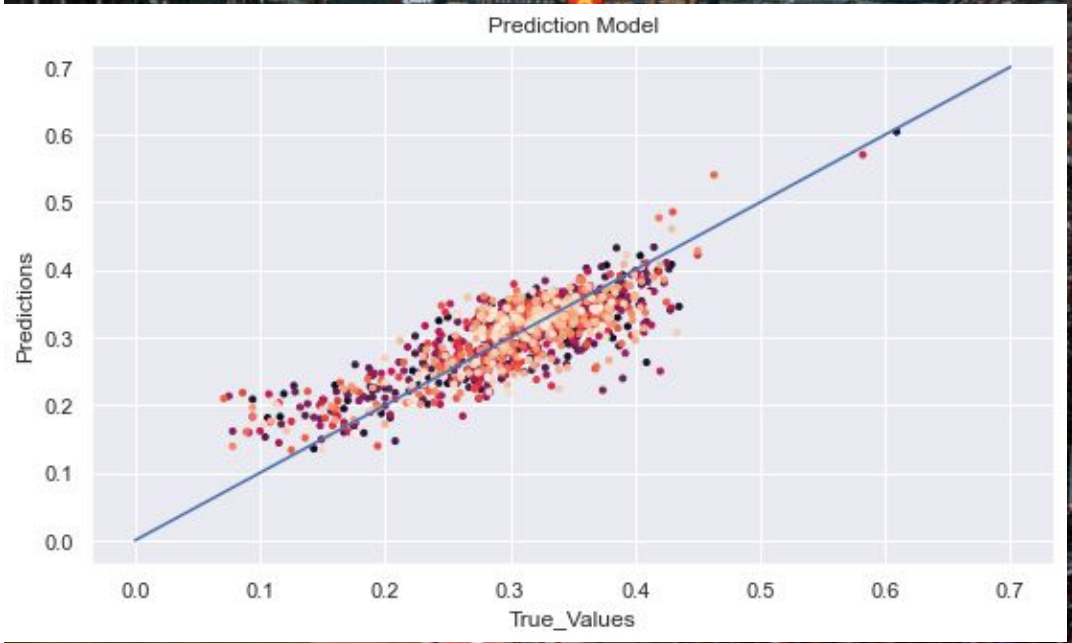
Ridge vs Lasso: initial tests





OLS Regression and Model Performance

Dep. Variable:	OBP	**R-squared:**	0.510
Model:	OLS	**Adj. R-squared:**	0.507
Method:	Least Squares	**F-statistic:**	191.9
Date:	Fri, 22 Jan 2021	**Prob (F-statistic):**	0.00
Time:	05:23:26	**Log-Likelihood:**	6031.4
No. Observations:	3710	**AIC:**	-1.202e+04
Df Residuals:	3689	**BIC:**	-1.189e+04
Df Model:	20		
Covariance Type:	nonrobust		



Model Comparison

TEST RESULT (POLYNOMIAL)

R^2: 0.6714972351229689

MAE: 0.028992589730067703

RSME: 0.03790966798090626

TEST RESULT (RIDGE)

R^2: 0.5063249302696317

MAE: 0.03568400027756962

RSME: 0.04647303805789564

TEST RESULT (LASSO)

R^2: -0.0023379222137960287

MAE: 0.04888737069925673

RSME: 0.06621975155999843

Persona 02

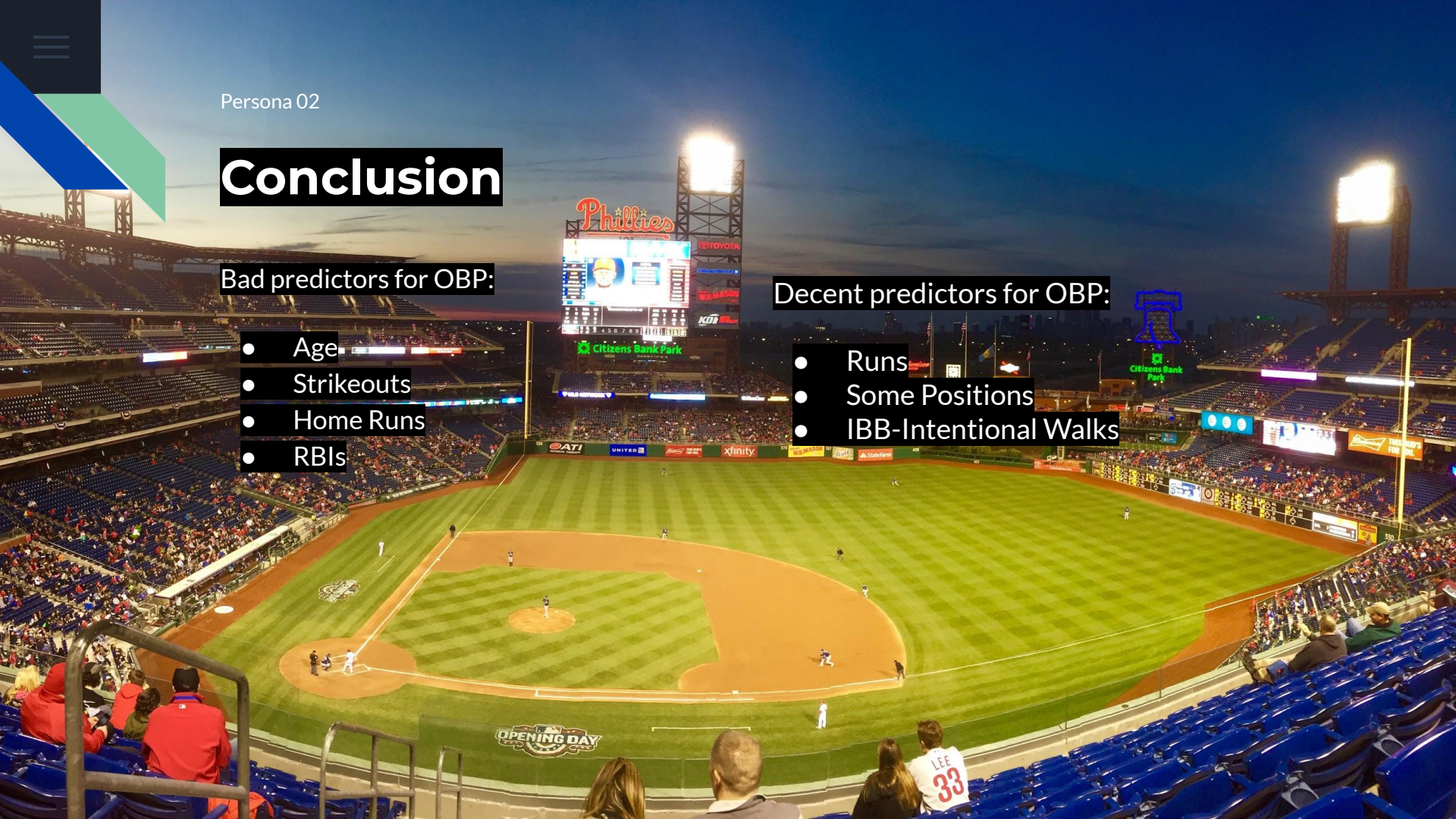
Conclusion

Bad predictors for OBP:

- Age
- Strikeouts
- Home Runs
- RBIs

Decent predictors for OBP:

- Runs
- Some Positions
- IBB-Intentional Walks



Insight and Future Exploration

-not the best predictive model

- Data for minor leagues (-->success in MLB)
- Does OBP differ based on players origin
- Physical traits (height, weight, ~speed)