Predicting Player Salary

By: Bartu, Lily, and Sutter

Objective

- Deciding how much to pay a player is an extremely complicated process
- Having a more accurate evaluation of a player's worth can provide a massive edge to any team
  - Trades
  - Structuring a roster
- To this end, we wanted to examine how the MLB values players
- Who is underpaid? Who is overpaid?

Predictions

- Age will be an important factor
  - Rookies will be underpaid
  - Veterans will be overpaid
- HR and WAR will be important
- We will find large discrepancies between how the MLB values a player and how much they are being paid

The Data

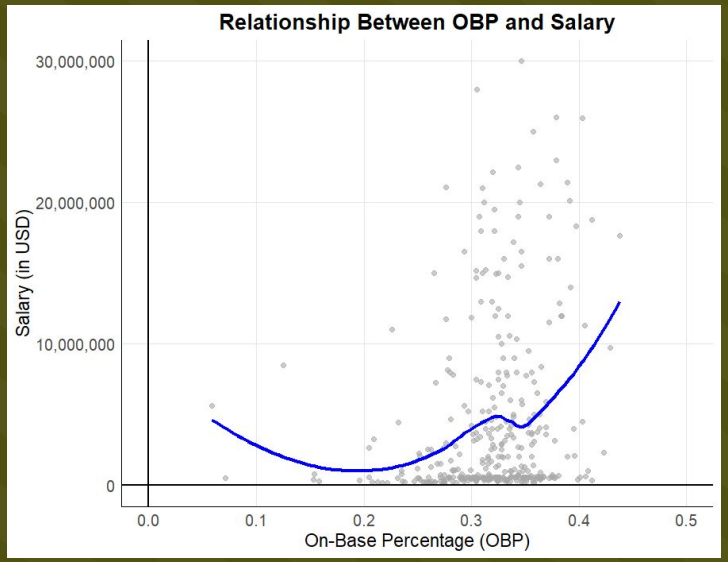
mlbSalaries

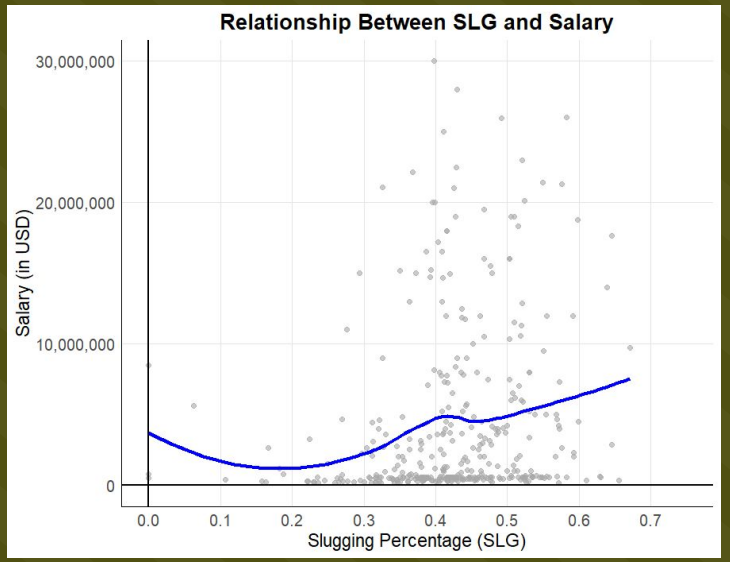
year	team	name	salary	playerID
2019	arizona-diamondbacks	Eduardo Escobar	6166666	500871
2019	arizona-diamondbacks	Robbie Ray	6050000	592662
2019	arizona-diamondbacks	Taijuan Walker	5025000	592836
2019	arizona-diamondbacks	Jake Lamb	4825000	571875
2019	arizona-diamondbacks	Adam Jones	4500000	430945
2019	arizona-diamondbacks	Alex Avila	4250000	488671
2019	arizona-diamondbacks	Jarrod Dyson	4000000	502481
2019	arizona-diamondbacks	Wilmer Flores	3750000	527038
2019	arizona-diamondbacks	Nick Ahmed	3662500	605113
2019	arizona-diamondbacks	Yoshihisa Hirano	3500000	673633
2019	arizona-diamondbacks	Mike Leake	3000000	502190
2019	arizona-diamondbacks	Merrill Kelly	2000000	518876
2019	arizona-diamondbacks	Ketel Marte	2000000	606466
2019	arizona-diamondbacks	Andrew Chafin	1945000	605177
2019	arizona-diamondbacks	Archie Bradley	1830000	605151

mlb-player-stats-Batters

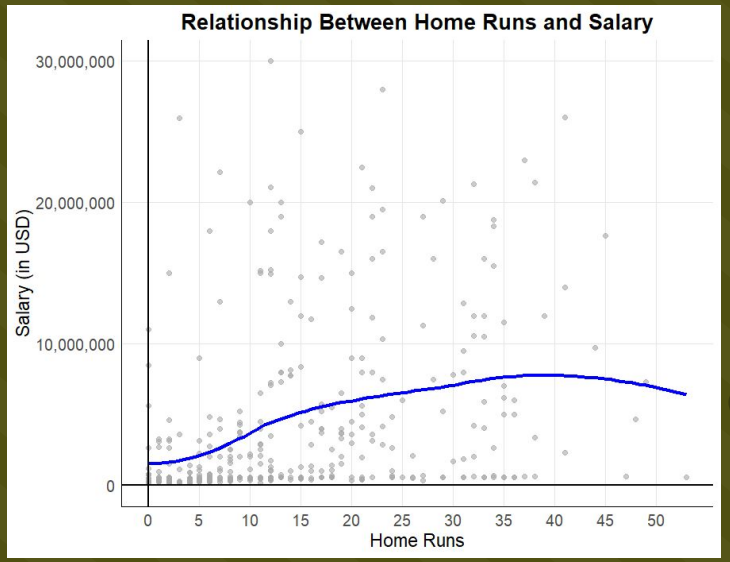
Player	Team	Pos	Age	G	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	SH	SF	HBP	AVG	OBP	SLG	OPS
Whit Merrifield	KC	2B	35	162	681	105	206	41	10	16	74	20	10	45	126	0	4	5	0.302	0.348	0.463	0.811
Marcus Semien	OAK	2B	34	162	657	123	187	43	7	33	92	10	8	87	102	0	1	2	0.285	0.369	0.522	0.891
Rafael Devers	BOS	3B	28	156	647	129	201	54	4	32	115	8	8	48	119	1	2	4	0.311	0.361	0.555	0.916
Jonathan Villar	BAL	3B	33	162	642	111	176	33	5	24	73	40	9	61	176	2	4	4	0.274	0.339	0.453	0.792
Ozzie Albies	ATL	2B	27	160	640	102	189	43	8	24	86	15	4	54	112	0	4	4	0.295	0.352	0.500	0.852
Eduardo Escobar	ARI	3B	35	158	636	94	171	29	10	35	118	5	1	50	130	0	10	3	0.269	0.320	0.511	0.831
Starlin Castro	MIA	3B	34	162	636	68	172	31	4	22	86	2	2	28	111	0	9	3	0.270	0.300	0.436	0.736
Jose Abreu	CWS	1B	37	159	634	85	180	38	1	33	123	2	2	36	152	0	10	13	0.284	0.330	0.503	0.833
Jorge Polanco	MIN	2B	31	153	631	107	186	40	7	22	79	4	3	60	116	2	7	4	0.295	0.356	0.485	0.841
Ronald Acuna	ATL	OF	26	156	626	127	175	22	2	41	101	37	9	76	188	0	1	9	0.280	0.365	0.518	0.883
Eric Hosmer	SD	1B	35	160	619	72	164	29	2	22	99	0	3	40	163	0	5	3	0.265	0.310	0.425	0.735
Amed Rosario	NYM	2B	28	157	616	75	177	30	7	15	72	19	10	31	124	2	3	3	0.287	0.323	0.432	0.755
Xander Bogaerts	BOS	SS	32	155	614	110	190	52	0	33	117	4	2	76	122	0	6	2	0.309	0.384	0.555	0.939
Cesar Hernandez	PHI	2B	34	161	612	77	171	31	3	14	71	9	2	45	100	0	4	6	0.279	0.333	0.408	0.741
DJ LeMahieu	NYY	3B	36	145	602	109	197	33	2	26	102	5	2	46	90	1	4	2	0.327	0.375	0.518	0.893
Trey Mancini	BAL	1B	32	154	602	106	175	38	2	35	97	1	0	63	143	0	5	9	0.291	0.364	0.535	0.899
Elvis Andrus	TEX	SS	36	147	600	81	165	27	4	12	72	31	8	34	96	0	10	4	0.275	0.313	0.393	0.706
Francisco Lindor	CLE	SS	30	143	598	101	170	40	2	32	74	22	5	46	98	1	6	3	0.284	0.335	0.518	0.853
Paul Goldschmidt	STL	1B	37	161	597	97	155	25	1	34	97	3	1	78	166	0	3	2	0.260	0.346	0.476	0.822
Pete Alonso	NYM	1B	29	161	597	103	155	30	2	53	120	1	0	72	183	0	3	21	0.260	0.358	0.583	0.941
Mookie Betts	BOS	OF	32	150	597	135	176	40	5	29	80	16	3	97	101	0	9	3	0.295	0.391	0.524	0.915
Freddie Freeman	ATL	1B	35	158	597	113	176	34	2	38	121	6	3	87	127	0	2	6	0.295	0.389	0.549	0.938
David Fletcher	LAA	2B	30	154	596	83	173	30	4	6	49	8	3	55	64	1	1	0	0.290	0.350	0.384	0.734
Kevin Pillar	SF	OF	35	156	595	82	157	37	3	21	87	14	5	18	86	0	6	9	0.264	0.293	0.442	0.735

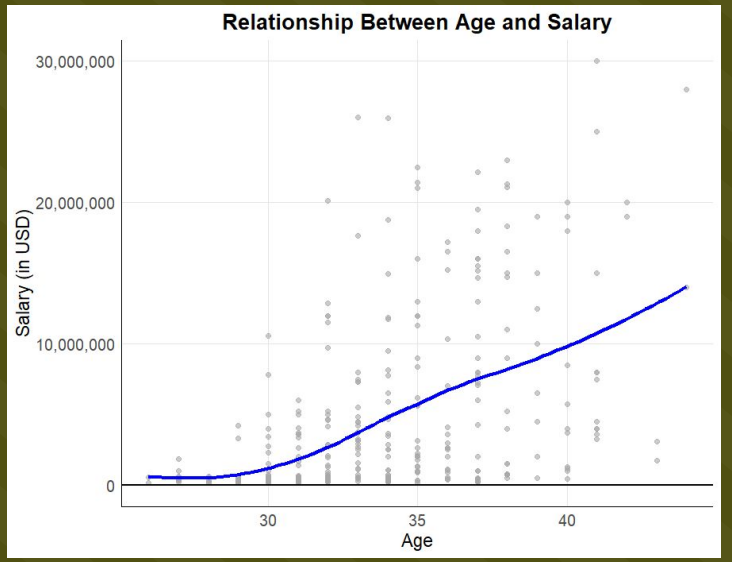
Examining Correlations



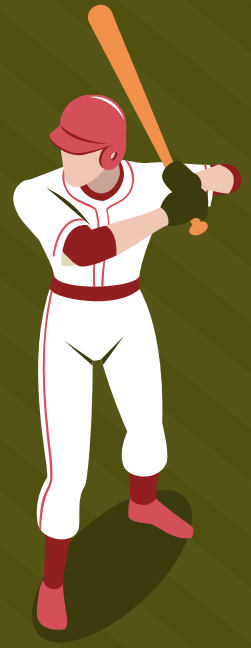


Examining Correlations





Our Process



- One data set with various player stats(age, hits, home runs, AVG, SLG, RBI…)
- One data set with player salaries
- Merged the data sets
- Pre-processed the data
  - Dropping unnecessary columns
  - Removing players with <50 at bats
- Created a Support Vector Regression model
- Created a Random Forest Model
- Created a Linear Regression model(control)

Refining Our Model

- We tuned hyperparameters
- Tried several different types of SVR
  - Linear
  - Polynomial
  - *Radial Basis Function(RBF)*
- Experimented with different combinations and number of features

Results

SVR

```
Root Mean Squared Error: 4508876.223824617
R-squared Score: 0.5269883826991046
```

Random Forest

```
Root Mean Squared Error: 4886104.713525955
R-squared Score: 0.4445298564353426
```

Linear Regression

```
Root Mean Squared Error: 4772336.273188315
R-squared Score: 0.4700959288963734
```

Using SVR Model to Predict a Player's Salary

- We withheld one player, Pete Alonso, who made his MLB debut in 2019, from training and testing
- Then, using our SVR model, we predicted what his salary would be
- It gave a result of 2 million even though he was being paid 500,000
- This demonstrates what we expected about rookies being underpaid

Permutation Importance

	Feature	Importance
3	Age	0.725375
1	H	0.282499
0	HR	0.135981
7	BB	0.102782
5	OBP	0.044643
6	OPS	0.024291
2	AVG	0.016007
4	SLG	0.012440

When we manually changed Pete Alonso's age to 35, his expected salary skyrocketed to 7.5 million

Some advanced metrics might have lower importance in our model due to overlap between features

Findings

- Although our SVR model outperforms Linear Regression, it still on average misses player salary by 4.5 million dollars
- Although this number is high, given salary in our data set ranges from 143,000 to 30 million dollars, this is understandable
- This shows the difficulty of predicting player's salary using player statistics
- It also demonstrates that some MLB player valuations might be tenuous

Future Improvements

- We could incorporate data from other years and normalize their salary data to increase the depth of our data set
- We could explore different combinations of features as well as engineering more features to see what improves performance the most
- We could find a more objective metric of player value that isn't tied to salary, and then compare that with their salary