WORLD SERIES CHAMPIONS

Chicago Tribune

THURSDAY, NOVEMBER 3, 2016 CHICAGOTRIBUNE.COM

BRIAN CASSELLA/CHICAGO TRIBUNE

Anthony Rizzo hugs Kris Bryant, while Mike Montgomery, Javier Baez and Addison Russell celebrate after winning the World Series in Cleveland on Wednesday night.

At last!

Cubs capture first title in 108 years in extra-inning Game 7 thriller

PAUL SULLIVAN
Chicago Tribune

CLEVELAND — Finally.

The most epic drought in sports history is over, and the Cubs are world champions.

After more than a century of waiting, the Cubs won the 2016 World Series with an 8-7 Game 7 victory over the Indians in 10 innings Wednesday night at Progressive Field. The triumph completed their climb back from a 3-1 Series deficit to claim their first championship since 1908.

This is not a dream. The Cubs did it.

It was real, and it was spectacular.

Tears flowed across Cubs Nation after the final out in the 10th, and fans responded with the world's biggest group hug, remembering all the loved ones who could only imagine what it would be like to experience this moment of pure bliss.

The 1969 Cubs, the team that defined the word "collapse" wore off the hook. So were their predecessors in '84 and 2003, who also came close only to suffer painful endings that scarred two generations of Cubs fans and kept the drought alive.

The billy goat is gone, and the black cat too. And what was the name of the fowl-ball chaser? No matter. It was never really his fault, and now he's just a footnote in Cubs history.

The catchphrase Cubs fans uttered over the last century and change has been "just one before I die," a plea that fell on deaf ears decade after decade.

Well, you can die in peace now, thanks to Joe Maddon's resilient club which was bloodied and on the mat after a Game 4 loss at Wrigley Field.

The Cubs picked themselves up when Jon Lester and Aroldis Chapman tag

Turn to **Cubs, Page 6**

$1.99 city and suburbs, $2.50 elsewhere 149th year No. 308 © Chicago Tribune

WAR and PCA

Predicting Player Value in Baseball

Presented by Team: The Matplotlibs

Addison Naylor, Eleni Bovalis, Sahil Sangani,
Tanvee Athavale, Annika Luthe

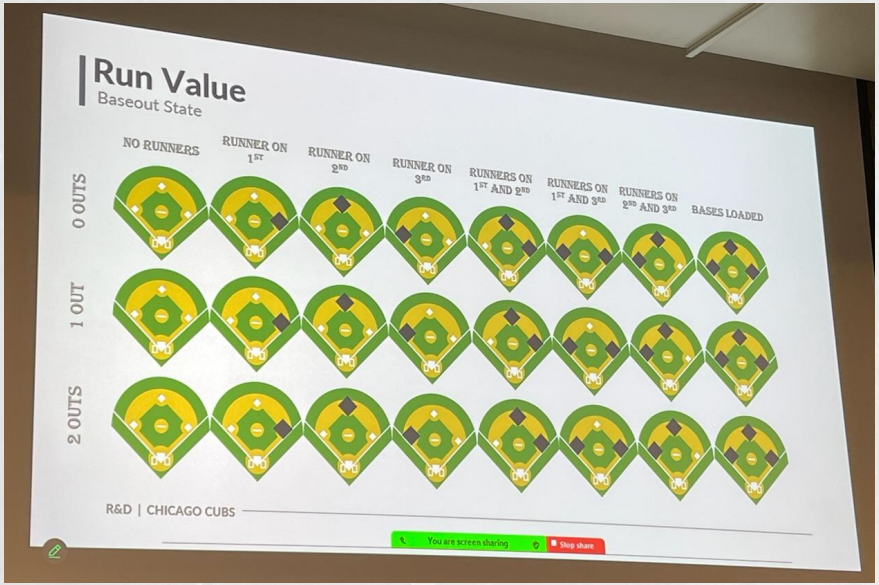
Table of contents

01

Project Overview

Initial Idea

Our first idea for this project was to generate a **lineup optimization order using Markov chains**. However, after attending a Quantitative Psychology baseball seminar with Cubs Assistant General Manager Dr. Ehsan Bokhari, we realized that we would have to complete **too many simulations** to create our model, which would require **a lot of computational power**.



15 usable players in
9 possible positions
would need
1,816,214,400+
simulations.

Our Project

Our **goal** was to build a model to predict how valuable a baseball player will be to their team in the upcoming season based on their current season's statistics.

How this project can be used:

- Help baseball managers and executives make trading decisions
- Decide which players to bring up from the minor leagues
- Fans can use it to predict which players will contribute the most to their favorite team in the upcoming season!

What is WAR?

In baseball, "WAR" stands for **Wins Above Replacement** which measures a player's overall value to a team based on how many wins they can earn compared to a "replacement-level" player.

Example: A player with a **WAR of 6.0** contributes **6 more wins** to their team compared to a different player who could be acquired at minimal cost (like a minor league player).

Fun fact: Babe Ruth has the highest career WAR of all time: 182.6

02
Exploratory Data Analysis

Data Cleaning

We collected our data from: **https://www.baseball-reference.com**, which has stats on a player's games played, on base percentage, runs scored, and more.

Standard Batting ✨Upgraded✨ Share & Export ▾ Glossary

Regular Season | Playoffs

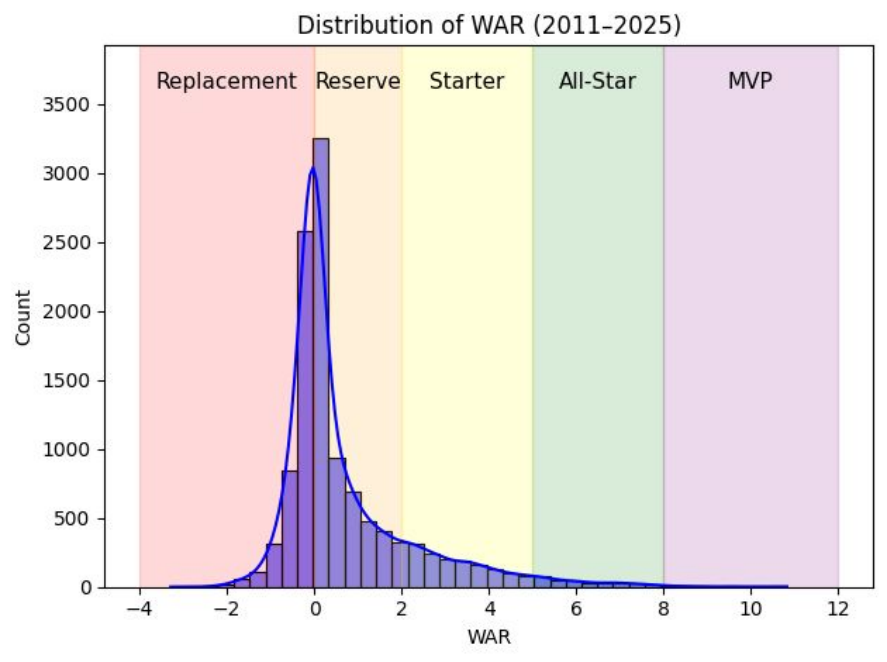
Rk	Player	Age	Pos	WAR	G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	OPS	OPS+	rOBA	Rbat+	TB	GIDP	HBP	SH	SF	IBB	Pos	Awards
1	Carson Kelly	30	C	3.6	111	421	369	48	92	13	1	17	50	2	0	45	80	.249	.333	.428	.761	119	.337	121	158	11	3	0	4	0	2/HD	
2	Michael Busch*	27	1B	4.6	155	592	524	78	137	25	5	34	90	4	0	56	139	.261	.343	.523	.866	147	.376	148	274	2	10	0	2	2	*3H/D	MVP-16
3	Nico Hoerner	28	2B	6.2	156	649	599	89	178	29	4	7	61	29	6	39	49	.297	.345	.394	.739	114	.338	118	236	4	7	0	4	1	*4/6	MVP-19,GG
4	Dansby Swanson	31	SS	4.5	159	645	590	84	144	24	3	24	77	20	3	47	168	.244	.300	.417	.717	105	.323	108	246	9	2	1	5	1	*6	
5	Matt Shaw	23	3B	3.1	126	437	393	57	89	21	3	13	44	17	5	38	94	.226	.295	.394	.690	98	.309	99	155	2	2	0	4	0	*5/H4	ROY-9
6	Ian Happ#	30	LF	4.0	150	663	569	87	138	32	0	23	79	6	3	87	151	.243	.342	.420	.762	120	.338	122	239	8	2	0	5	2	*7/DH	GG
7	Pete Crow-Armstrong*	23	CF	6.0	157	647	591	91	146	37	4	31	95	35	8	29	155	.247	.287	.481	.768	118	.337	117	284	1	9	6	***12***	1	*8/HD	AS,MVP-9,GG
8	Kyle Tucker*	28	RF	4.6	136	597	500	91	133	25	4	22	73	25	3	87	88	.266	.377	.464	.841	143	.375	143	232	8	4	1	3	7	*9D/H	AS,SS
9	Seiya Suzuki	30	DH	2.6	151	651	571	75	140	31	3	32	103	5	2	71	164	.245	.326	.478	.804	130	.348	129	273	13	1	0	8	1	D97/H8	MVP-20

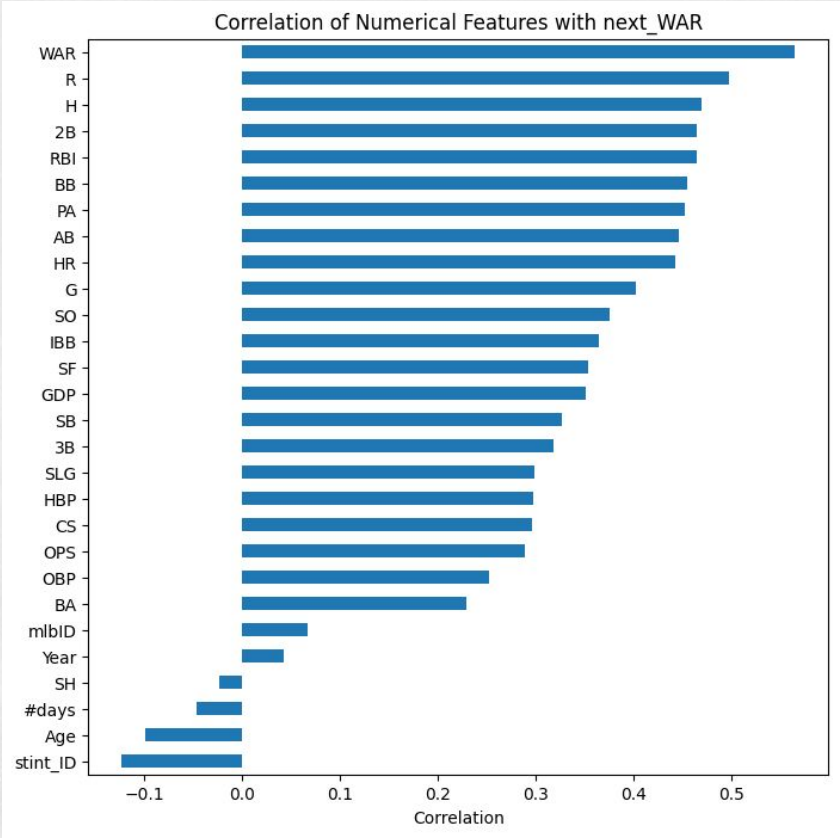
Data Cleaning

We used data from all MLB teams from **2010-2023 to train** our model and data from **2024-2025 to test** our model. We started with about **13000 rows of data**.

After filtering out unneeded variables, we found **1206 missing values**, and given our large dataset, decided to drop them.

EDA


Distribution of WAR (2011–2025)
Replacement
Reserve
Starter
All-Star
MVP
Count
WAR
3500
3000
2500
2000
1500
1000
500
0
−4
−2
0
2
4
6
8
10
12

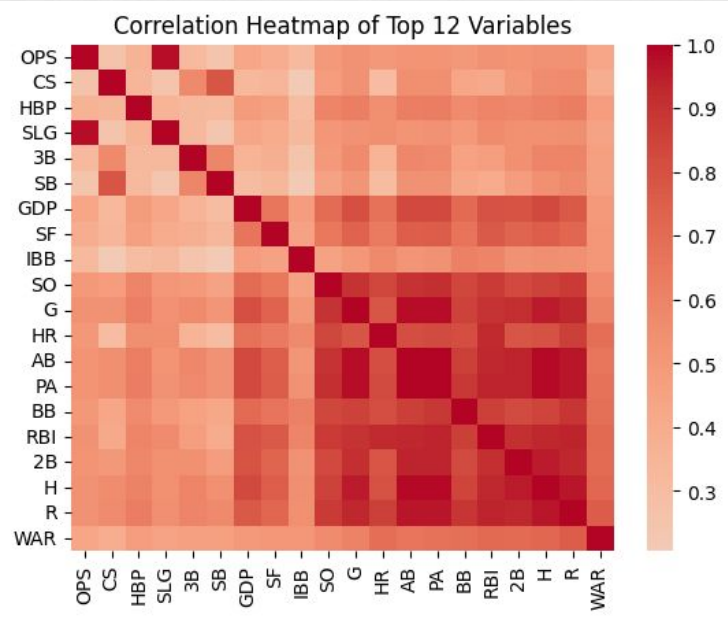

Correlation of Numerical Features with next_WAR
WAR
R
H
2B
RBI
BB
PA
AB
HR
G
SO
IBB
SF
GDP
SB
3B
SLG
HBP
CS
OPS
OBP
BA
mlbID
Year
SH
#days
Age
stint_ID
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
Correlation

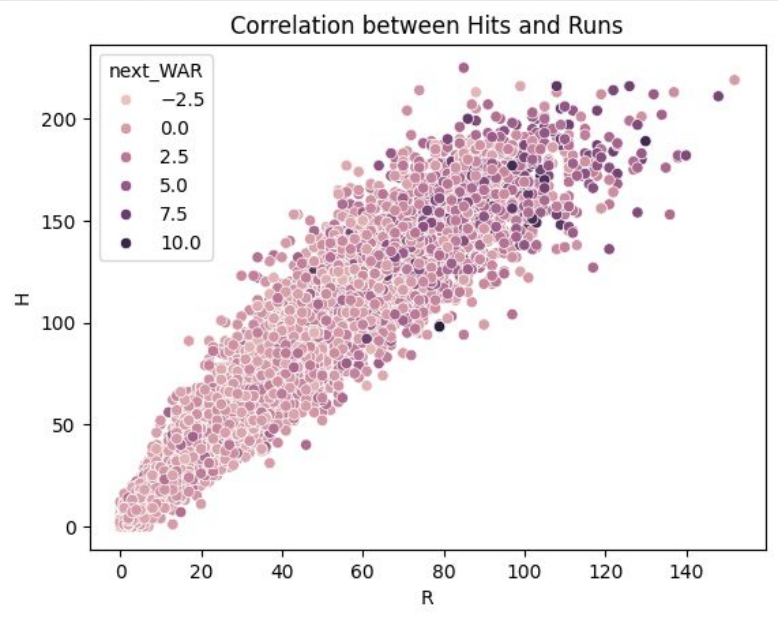
PCA

n_components = **0.90**

From 34 variables to just **14 components**

- Mitigation of multicollinearity
- Faster training times
- Reduced overfitting

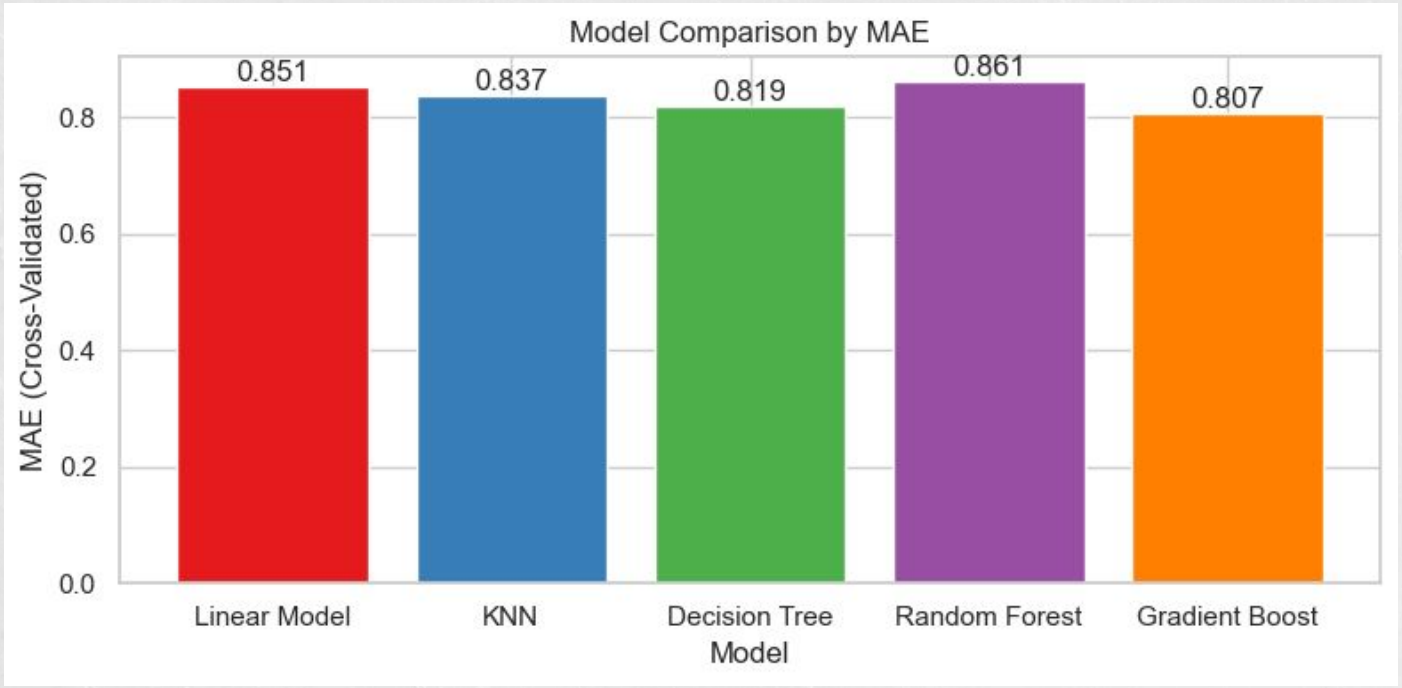




03

Modeling

Models


Model Comparison by MAE
MAE (Cross-Validated)
0.0
0.2
0.4
0.6
0.8
0.851
0.837
0.819
0.861
0.807
Linear Model
KNN
Decision Tree
Random Forest
Gradient Boost
Model

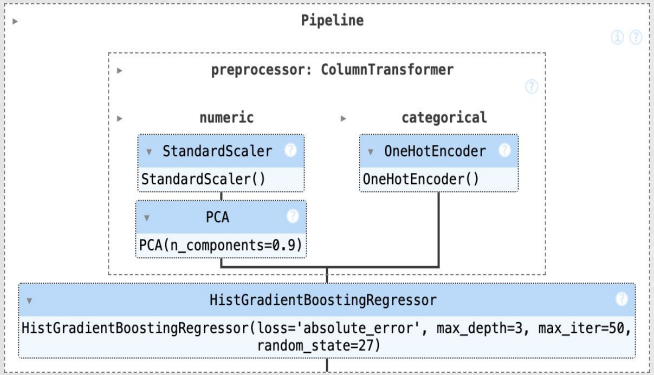
Best Model

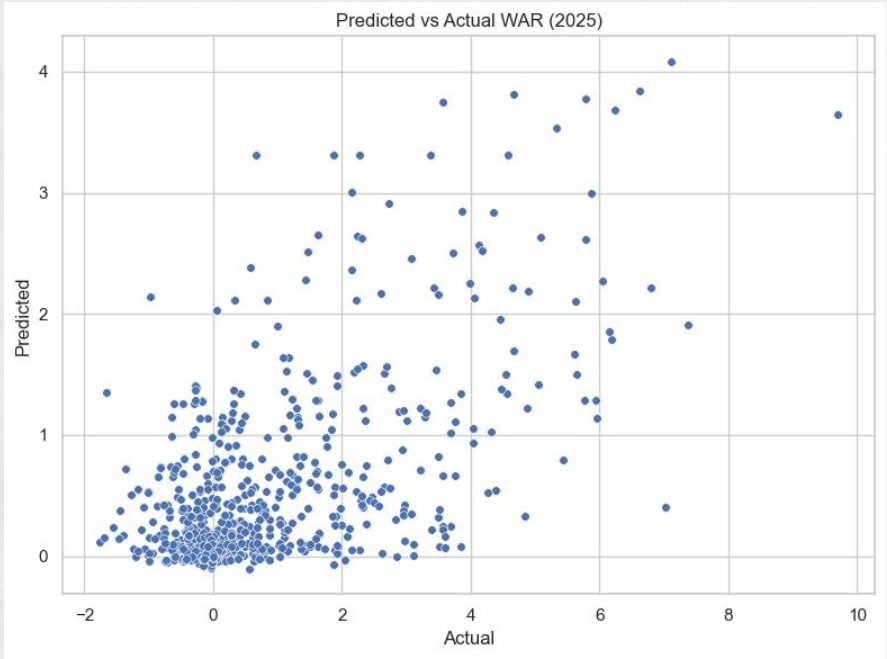
Histogram-Based Gradient Boosting Regressor from Scikit-Learn

5 Fold CV MAE: 0.807

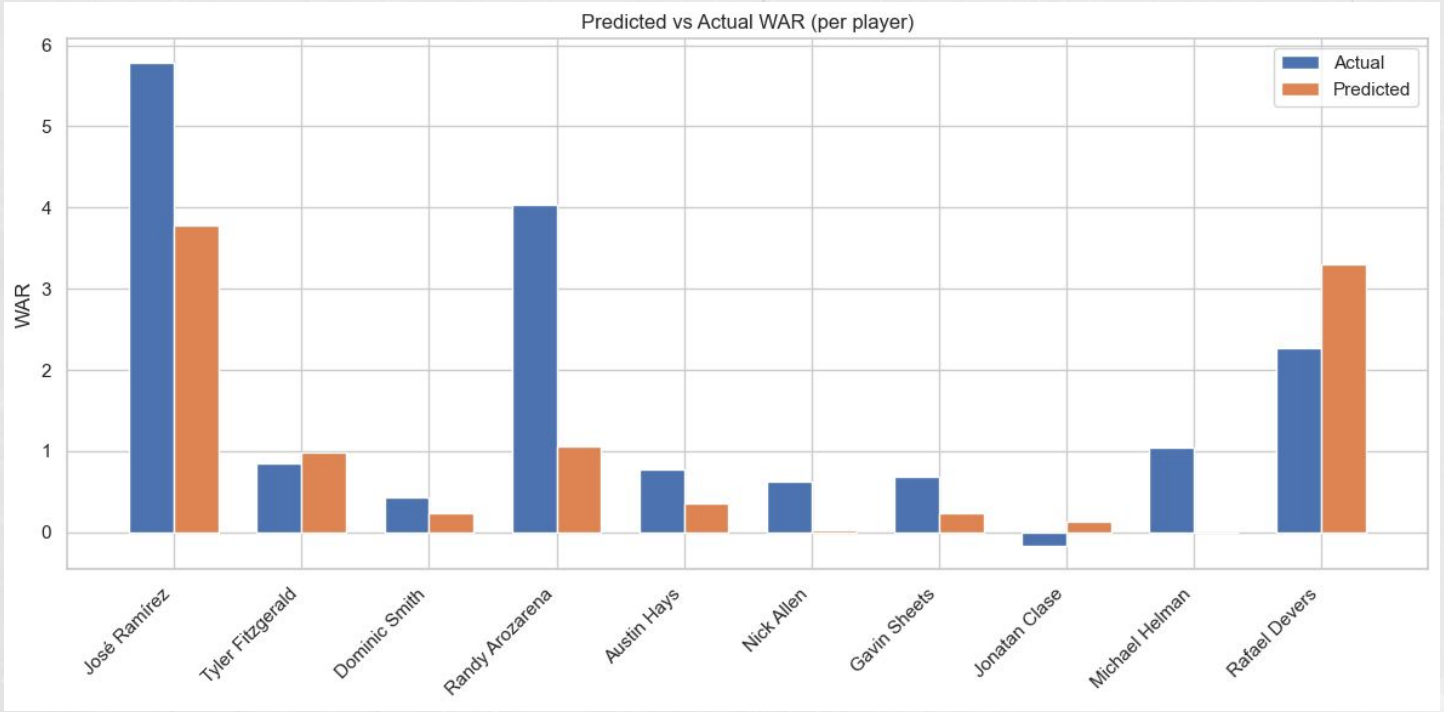
Test RMSE: 1.368

Test MAE: 0.931





Prediction Accuracy



04

Challenges

Roadblocks & Breakthroughs

- Figuring out our project (initial Markov chain idea)
- Looking for an efficient way to import our data
- Finding fixes for variable selection and multicollinearity

- We asked for advice (Dr. Bokhari's presentation)
- Got help from our Project Lead who introduced us to pybaseball
- Implemented PCA for dimensionality reduction

05

Conclusion

Model Reliability

Our best model has a **MAE** of **0.807**, indicating that it is **reasonably trustworthy**.

It can be used for predicting whether a player is valuable, but shouldn't be used to determine if a player needs to be let go.

Example: If a player's WAR for next season is predicted to be 3.0, a baseball manager can be reasonably certain (within about a 2.2 - 3.8 range) that this player will make valuable addition to their team.

Future

Further analysis could include...

- Predicting the value of pitchers (different than batters)
- Incorporating minor league and college data
- Predicting WAR further into the future
- Time series analysis to examine trends across a player's whole career

Thank You!

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Sources

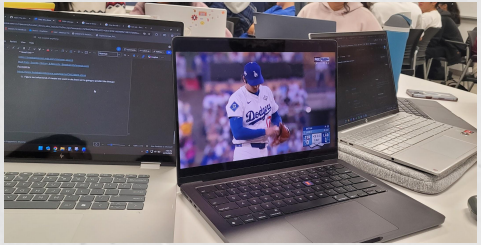


Image 1 (upper left): https://www.businessinsider.com/kris-bryant-smiling-chicago-cubs-world-series-champions-2016-11

Image 2 (bottom left):
https://cloudfront-us-east-1.images.arcpublishing.com/gray/ZVM2P6PZVBMLHDC2O4Z3LLWOJE.jpg

Image 3 (upper middle):
https://ca-times.brightspotcdn.com/dims4/default/d4ed453/2147483647/strip/true/crop/2048x1152+0+0/resize/1200x675!/quality/75/?url=https%3A%2F%2Fcalifornia-times-brightspot.s3.amazonaws.com%2Fb7%2F27%2F7d3a7e9705d8dcb949891fb6e519%2Fla-xcxcschilken-1478188604-snap-photo

Image 4 (bottom middle):
https://www.usatoday.com/gcdn/-mm-/cfddfdec04a3541d7c1e4182939fcdd316de13cc/c=0-239-3933-2461/local/-/media/2016/11/03/USATODAY/USATODAY/636137488792285604-USP-MLB--World-Series-Chicago-Cubs-at-Cleveland-In.7.jpg

Image 5 (right side):
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.chicagotribune.com%2F2016%2F11%2F03%2Fphotos-world-series-front-pages%2F&psig=AOvVaw2O4-H9GMPaEv-JriGYw3FB&ust=1763849512988000&source=images&cd=vfe&opi=89978449&ved=0CBYQjRxqFwoTCND3ic-hhJEDFQAAAAAdAAAAABAL

Image 6 (right side)
https://www.chicagotribune.com/wp-content/uploads/2024/04/CTC-L-Chicago-Cubs-03_187925179.jpg?w=1541

WAR meaning Google search:
https://www.google.com/search?q=what+does+wins+above+replacement+mean&oq=what+does+wins+above+re&gs_lcrp=EgZjaHJvbWUqBwgAEAAYgAQyBwgAEAAYgAQyBggBEEUYOTIHCAIQABiABDIHCAMQABiABDIICAQQABgWGB4yCAgFEAAYFhgeMggIBhAAGBYYHjIICAcQABgWGB4yCAgIEAAYFhgeMggICRAAGBYYHtIBCDQ0MzVqMGo3qAIAsAIA&sourceid=chrome&ie=UTF-8

Replacement-level player meaning Google search:
https://www.google.com/search?q=replacement-level+player+meaning&oq=replacement-level+player+meaning&gs_lcrp=EgZjaHJvbWUyBggAEEUYOTIICAEQABgWGB4yDQgCEAAYhgMYgAQYigUyDQgDEAAYhgMYgAQYigUyCggEEAAYgAQYogQyCggFEAAYgAQYogQyBwgGEAAY7wUyBwgHEAAY7wXSAQgxOTMyajBqN6gCALACAA&sourceid=chrome&ie=UTF-8

All-time WAR ranking website: https://www.baseball-reference.com/leaders/WAR_career.shtml

Baseball Data + screenshot: https://www.baseball-reference.com

Wii sports image: https://i.ytimg.com/vi/hp3De8lflWQ/hqdefault.jpg