2021 MLB season

Jungwon Yoon
ojyoon@ucdavis.edu

Hyunjin Chang
hjchang@ucdavis.edu

Sung Won Lee
sgwlee@udavis.edu

Jong Wook Choe
wchoe@ucdavis.edu

Abstract

The All-Star Game is a game between teams of outstanding players and many baseball fans are interested in players to be chosen for next coming All-Star game. We made a prediction by applying effective machine learning techniques to MLB 2021 data, which was crawled from the web page. Through visualizing the data, we could understand the data better and found some meaningful insights that might improve our prediction. Since our target variable is a binary categorical variable, we used Logistic regression to train the model and we were able to get the probability of the players to be involved in the next All-Star game.

1 Introduction

Major League Baseball (MLB) is a professional baseball organization and the oldest major professional sports league in the world. It is hard to find those who don't know about it. As of 2022, a total of 30 teams play in Major League Baseball. 15 teams in the National League (NL) and 15 in the American League (AL).

Throughout our analysis, we will be using the Data set for the 2021 MLB season in the US. Rather than downloading straight from websites, we used the scrapping data technique to get data sets. First, from the player data, we will find which factors might affect players to be nominated as All-Star. Using linear regression we will see the interrelation between variables.

The MLB All-star game occurs in mid-July every year. All-star rosters consist of 32 players on each side, made up of twenty position players and twelve pitchers, which each players are determined by a fan vote and manger reference. Here, we are going to create a model that predicts who will be the MLB 2022 season All-star.

2 MLB DATASET

The data is about 2021 Major League Baseball Team Statistics.

Acknowledgements Data-set:
All the data is from the (https://www.baseball-reference.com/leagues/majors/2021.shtml).

2.1 Data Scraping

There are tables that contains MLB teams and players data on 'https://www.baseball-reference.com/'. Each team has their players data in a separate page so we needed to access 30 pages to get all players data. In order to scrape players data for all 30 teams, we had to scrape the data dynamically using Python libraries called Selenium and BeautifulSoup. Selenium is a Web browser Automation Tool and we chose to use Selenium to automate the crawling process. First, Selenium was used to render

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

web pages to fetch dynamic contents, then BeaufitulSoup was used to scrape MLB data from the HTML table elements of all teams.

2.2 Data Description

After removing insignificant columns, the data contains 1706 rows and 23 columns. The target column is "All-Star", which is False if the player is not a All-Star and True if the player is a All-Star. The factors that might be useful to predict all-star players are

1. *Age*: Age of the MLB player at Midnight of June 30th of 2021.
2. *Ht*: Height of the player in ft(inches).
3. *Wt*: Weight of the player in lbs.
4. *Yrs*: Years the player has been in the major leagues.
5. *G*: Games played. This includes all times that the player apperared on the lineup card.
6. *GS*: Games started.
7. *Batting*: Games appeard in the batting order, but may not have batted.
8. *Defense*: Games in lineup at a defensive position.
9. *P*: Games in lineup or announced as a pitcher
10. *C*: Games in lineup as a catcher.
11. *1B*: Games in lineup as a first baseman.
12. *2B*: Games in lineup as a second baseman.
13. *3B*: Games in lineup as a third baseman.
14. *SS*: Games in lineup as a short stop.
15. *LF*: Games in lineup as a left fielder.
16. *CF*: Games in lineup as a center fielder.
17. *RF*: Games in lineup as a right fielder.
18. *OF*: Games in lineup as a outfielder.
19. *DH*: Games in line up as a designated hitter.
20. *PH*: Games in lineup as a pinch hitter. May have played an0other position as well.
21. *PR*: Games in lineup as a pinch runner. May have played an0other position as well.
22. *WAR*: Wins Above Replacement. A single number that represents the number of wins the player added to the team above what a replacement player would add.
    Scale for a single season: 8+ MVP quality, 5+ all star quality, 2+ Starter, 0-2 Reserve, < 0 Replacement level.

2.3 Data Exploration

Data visualization is a key skill of any data expert and applied in a wide variety of areas from scientific research to industrial application. Following pie plot using 'Plotly' library we can visualize categorical data.

Plotly is a Python graphing library which makes interactive, publication-quality graphs.

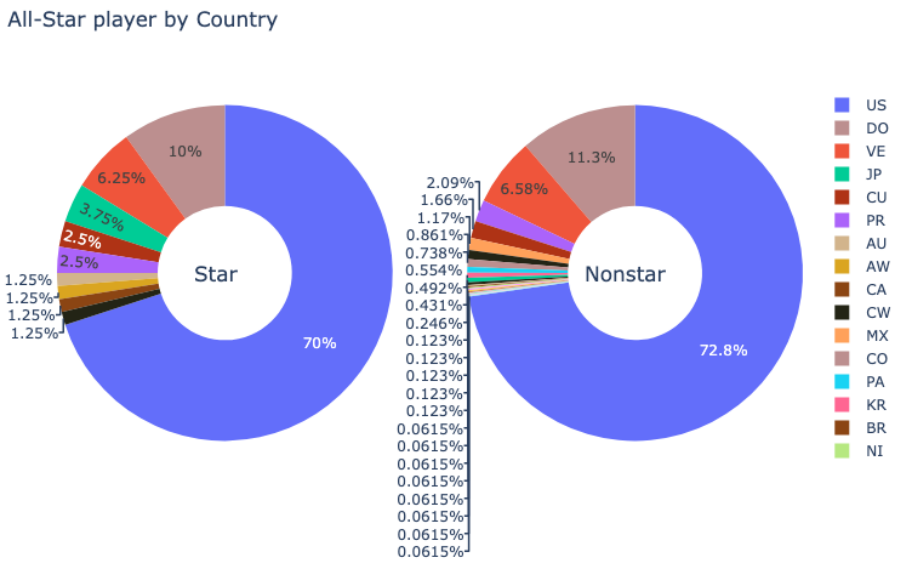

Figure 1: Distribution of athletes by Country

Athletes are classified by their nationalities with initials of different countries. The US for the United States, DO for the Dominican Republic, VE for Venezuela, JP for Japan, CU for Cuba, PR for Puerto Rico, AU for Australia, AW for Aruba, CA for Canada, CW for, MX for Mexico, CO for Colombia, PA for Panama, KR for South Korea, BR for Brazil, NI for Nicaragua.

This example of a pie chart represents the proportion of All-Star players and Non-Star players nominated in 2021. Most of the nominated players are from the United States. The Dominican Republic going next and then Venezuela. On the other hand, nonstar players also show a similar trend from the United States, the Dominican Republic going next, and then Venezuela. The reason is that the total number of international players in the MLB league is 28.2 % of a total pool of 975 players.

Team

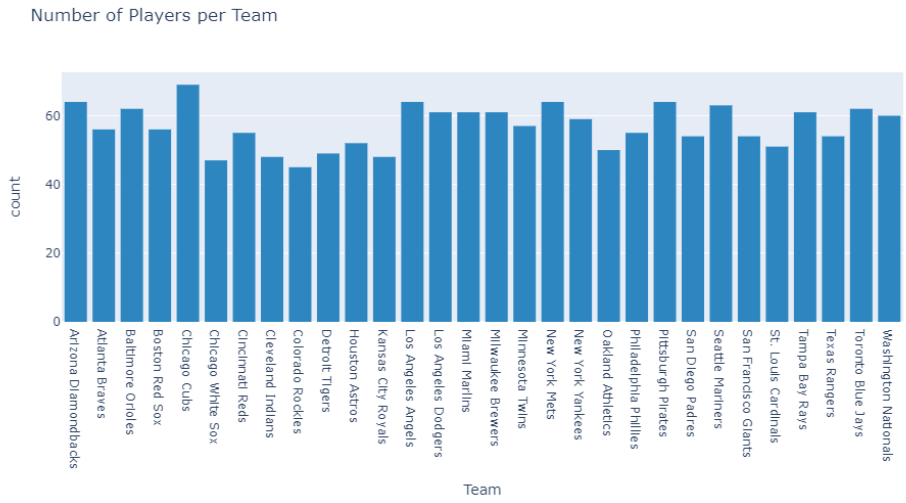

Figure 2: Number of Players

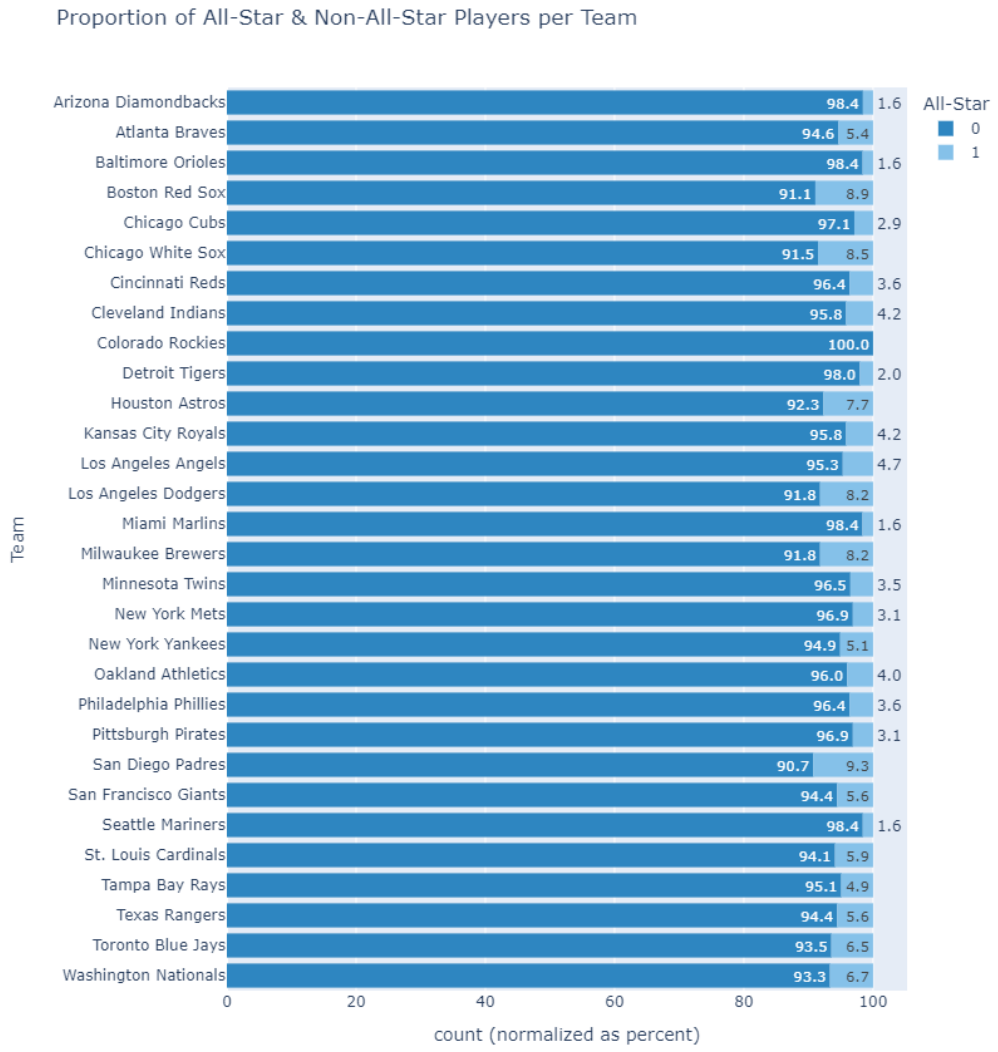

Figure 3: Proportion of All-Star Players

From these plots, we wanted to explore the team's composition.

The first plot describes how many players each team consists of. Among the 30 teams, Chicago Cubs has the largest number of players in their team with 69, with smallest team of 45 in Colorado Rockies.

The second plot is to see the proportion of players who were selected as all star or not. Surprisingly, Colorado Rocikes did not have any all star players. San Diego Padres has the highest all star players rate with 9.3 %. And following, Boston Red Sox has the next highest number of players selected as all start with a rate of 8.9%.

2.4 Data Visualization

Distribution of Players' Biometrics

Figure 4: Players' Biometrics by Age, Height, and Weights

Using the violin plot, we were able to take a look at the distribution of data of each feature against all-star players and non all-star players.

Age, Height, and Weights of players seem to be normally distributed, and there is not too much difference between the all-star and non all star. This can indicate that age, height, and weights (players of biometrics and physical parameters) are not the most significant factors deciding if the player will be selected as an all-star or not.

Distribution of Players' Information in the Leagues

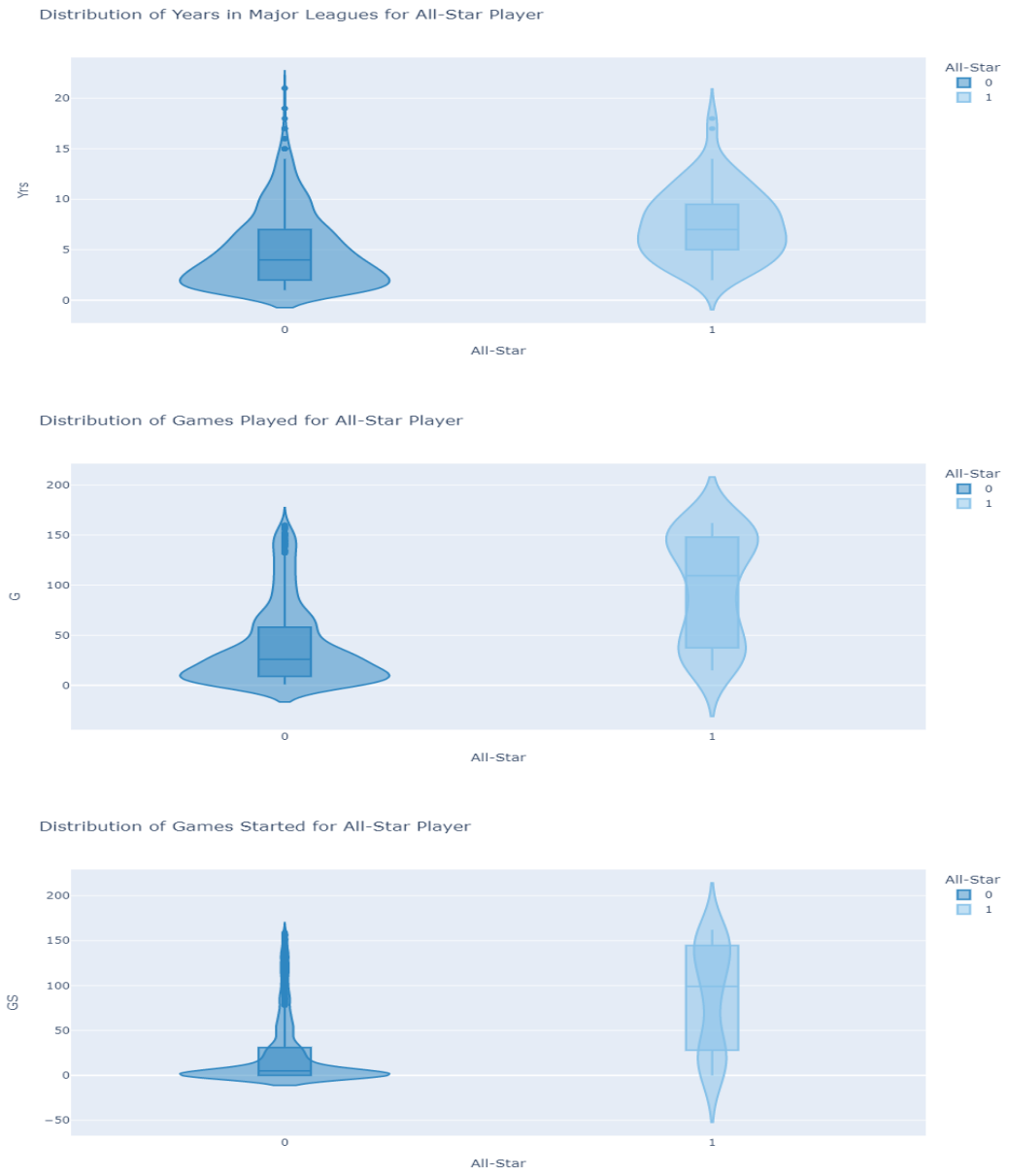

Figure 5: Players' Information by Years, Games Played and Started

Non all-star players' years of playing in the major leagues, and the number of games played, and number of times/games being the first pitcher to throw a pitch for their team seem to be very right skewed. This gives us an idea that a lot of non all star players are pretty new to the major league, have not played many games, have not served as the first pitcher many times.

On the other hand, for the all-star, it appears to be that years of being in the major leagues is normally distributed (pretty vary) and number of games played and games started as the first pitcher have bimodal distribution. This can indicate that some all star has played more games and there are also some all star participated in less games and similar observations to the number of games started as the first pitcher.

Comparing the non all-star and all-star players, all star players have served in the major league a bit longer (about 3 more years) than the non all-star and all-star players participated in 83 more games and served 84 more games as the first pitcher considering the median values.

To sum up, number of games players played and started as the first serve might be crucial features in predicting if a player will be participating in the all-star game.

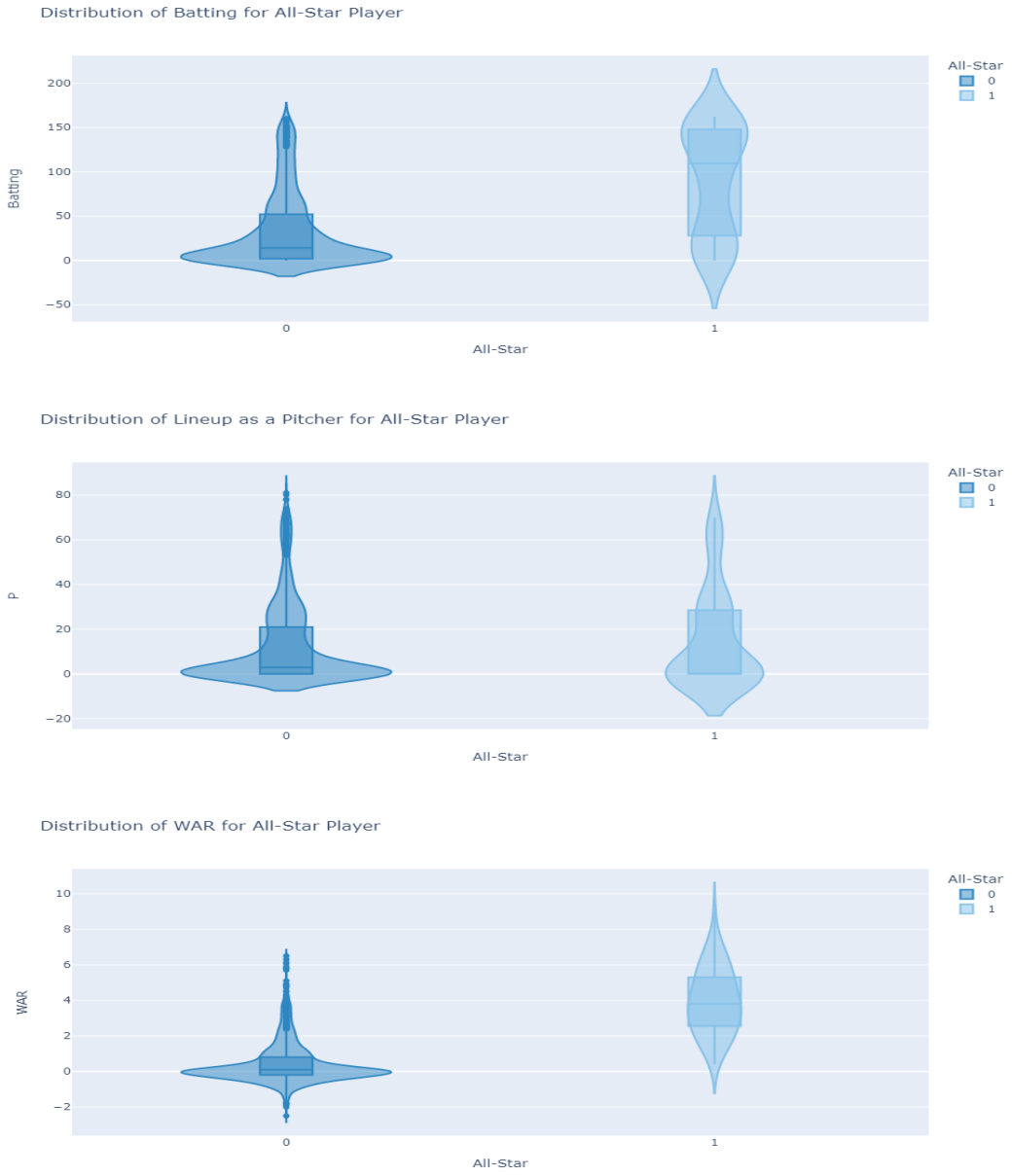

Figure 6: Players' Information by Batting, Defense, Pitcher Lineup, WAR

Similar observations in the second part of players information, all-star players seemed to have more batting. Yet, one interesting observation is that number of games in lineup or announced as a pitcher (not the first pitcher) does not really seem to be impactful. And definitely, the number of wins the player added to the team seems to be higher in the all-star group. This makes sense as it is a huge indication of player's performance in games, which is important when giving a player a chance to play in the all-star.

The following bar plots show the rank of the correlation with All-Star variable.

High correlation with **All-Star** status

- War, GS, Defense, G, Batting, Yrs, DH, RF, SS, and 2B

Low correlation with **All-Star** status

- Ht, Wt, P, LF, CF, Age, C, and 1B

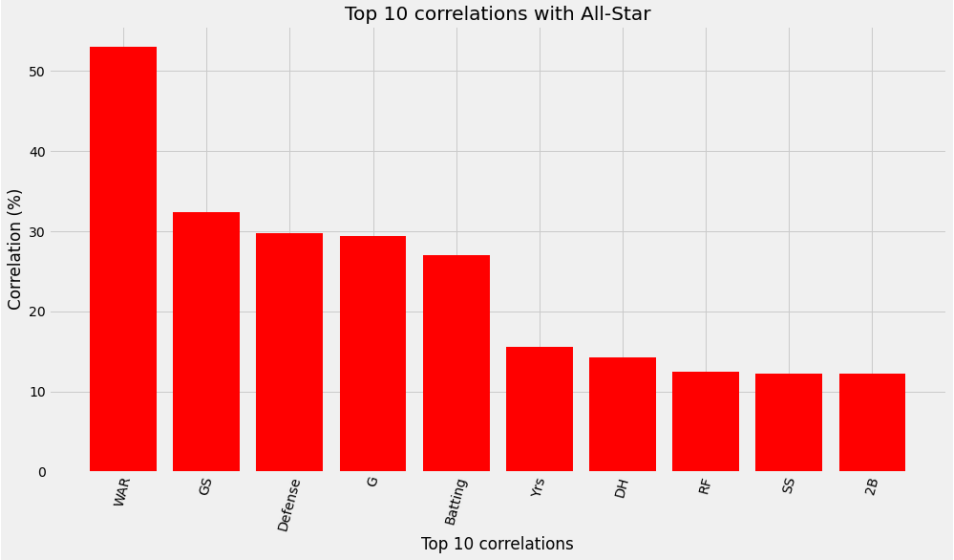

Figure 7: Correlation between variables

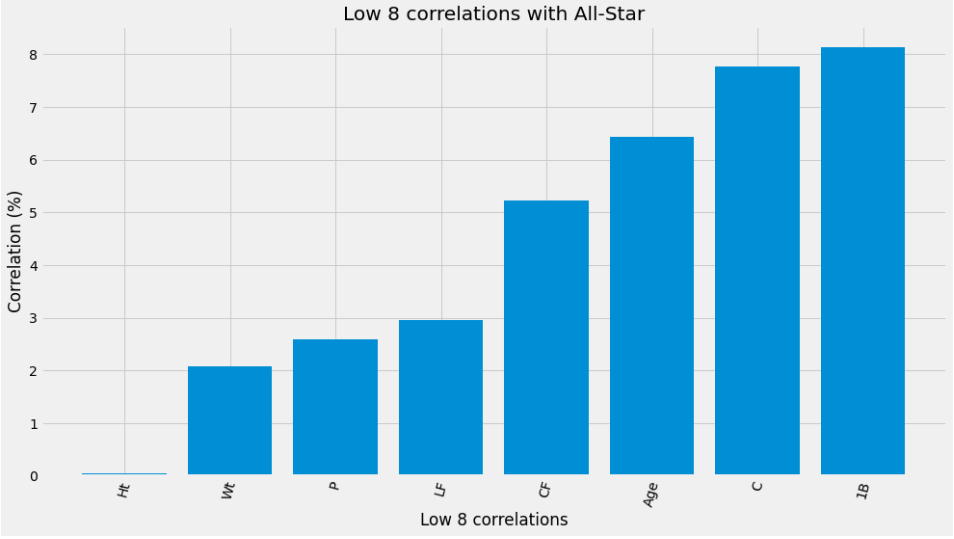

Figure 8: Correlation between variables

3 Methodology

Here, we are going to use each player's stats in MLB 2021 and 2022 season to forecast who will be playing in MLB 2022 All-star game.

3.1 Task

The target variable is the ”All_star” column which is a categorical variable in MLB 2021 season player statistics table. Since it is a categorical variable composed of True / False, we converted the target variable to 1 and 0 by one hot encoding.

Then, we selected explanatory variables in MLB 2021 seaon player table that had high correlation with the election to All-Star game. The explanatory variables we used for machine learning are {’Age’, ’Wt’, ’Yrs’, ’G’, ’GS’, ’Batting’, ’Defense’, ’P’, ’C’, ’1B’, ’2B’, ’3B’, ’SS’, ’LF’, ’CF’, ’RF’, ’OF’, ’DH’, ’PH’, ’PR’, ’WAR’}.

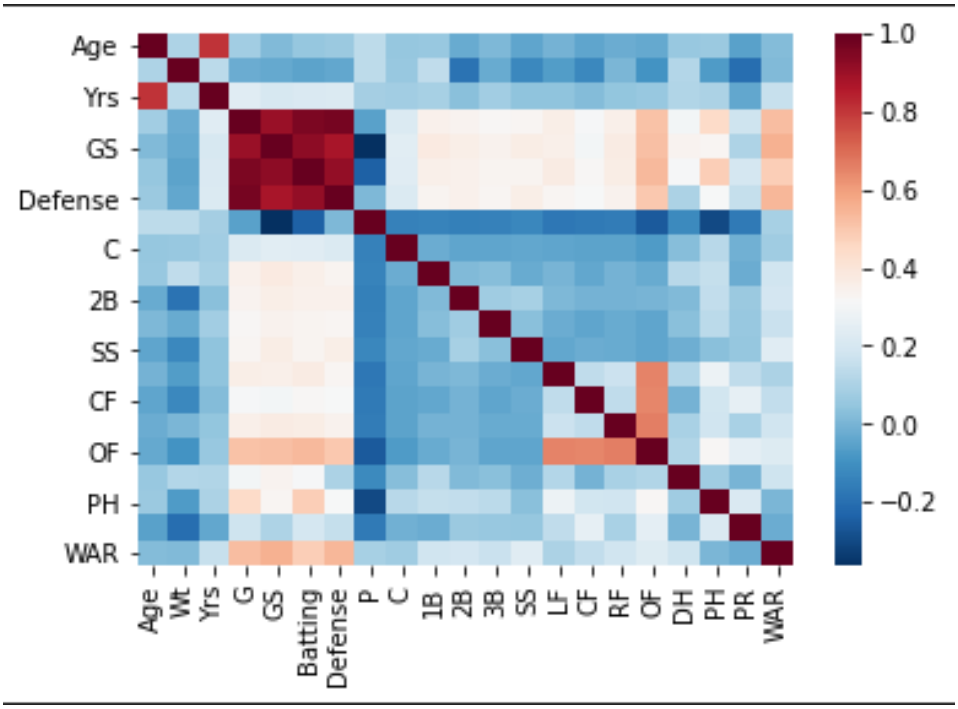

Figure 9: Correlation between each variables

The above graph shows the correlation between each explanatory variables in a heatmap.

To train the model for MLB 2021 season All-Star, we used Logistic regression because the target variable we are looking for is a binary categorical variable.

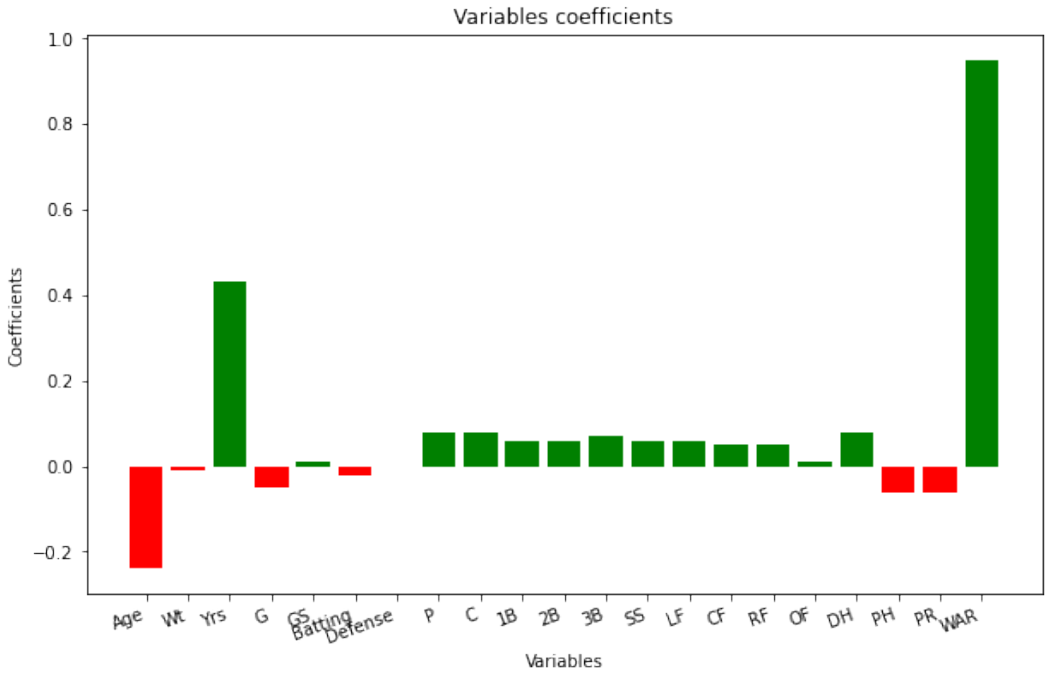

Figure 10: Feature Coefficients

The bar plot above shows the coefficients of each variable. As we could expected, WAR variable had the highest coefficient because since WAR measures a player's value in all facets of the game by deciphering how many wins one has contributed to the team, it refers one of the most important value in modern baseball stats. Therefore, a player with high WAR stat has higher probability to be elected as a All-Star player.

4 Analysis

After training the Logistic regression model, we were able to get the probability of the player to be involved in All-Star game.

	Name	Age	Wt	Yrs	G	GS	Batting	Defense	WAR	All_star	pred_lr
682	Shohei Ohtani	26	210	4	158	149	155	26	9.0	1	0.990009
1635	Marcus Semien	30	195	9	162	161	162	161	7.3	1	0.982946
636	Salvador Perez	31	255	10	161	160	161	124	5.3	1	0.980106
444	Jose Ramirez	28	190	9	152	151	152	133	6.7	1	0.975164
1148	Bryce Harper	28	210	10	141	140	141	139	5.9	0	0.945488
559	Carlos Correa	26	220	7	148	147	148	148	7.2	1	0.945458
1629	Robbie Ray	29	225	8	32	32	2	32	6.5	0	0.906612
1696	Juan Soto	22	224	4	151	146	151	144	7.1	1	0.892069
1181	Zack Wheeler	31	195	7	33	32	32	32	7.6	1	0.881271
938	Jorge Polanco	27	208	8	152	146	152	143	4.9	0	0.878303

Figure 11: Top 10 probability of All-Star

The table above shows Top 10 players who had the **highest** probability to be elected as All-Stars in MLB 2021 season. We can clearly see that most of the players in the table were actually played in All-Star game of 2021 season.

	Name	Age	Wt	Yrs	G	GS	Batting	Defense	WAR	All_star	pred_lr
1323	Yusei Kikuchi	30	205	3	29	29	1	29	1.7	1	0.009614
942	Taylor Rogers	30	190	6	40	0	2	39	0.4	1	0.014487
867	Omar Narvaez	29	220	6	123	101	123	111	1.4	1	0.018132
1259	Yu Darvish	34	220	9	30	30	29	30	1.4	1	0.020515
1457	Alex Reyes	26	220	5	70	0	67	69	0.6	1	0.024062
1015	Taijuan Walker	28	235	9	31	29	28	30	0.5	1	0.029601
709	Jared Walsh	27	210	3	144	136	144	144	2.8	1	0.030413
547	Gregory Soto	26	234	3	62	0	6	62	1.4	1	0.041108
1695	Kyle Schwarber	28	229	7	72	72	72	72	1.8	1	0.046121
411	Shane Bieber	26	200	4	16	16	2	16	2.7	1	0.055080

Figure 12: Lowest 10 probability of All-Star

The table above shows the Top 10 players who had the **lowest** probability to be elected as All-Star game in 2021 season. Most of the players in this table did not have the league top statistics compared to the other players but still they were elected as All-Star as the back up or replacement players of the actual All-Star players who refused to play the All-Star game for reasons such as injuries or tiredness.

Based on the data above, we applied our model to predict MLB 2022 season All-Stars based on the players' statistics of the first half 2022 season.

	Name	Age	Wt	Yrs	G	GS	Batting	Defense	WAR	prob_lr
894	Manny Machado	29	218	11	50	49	50	46	3.5	0.876715
316	Jose Ramirez	29	190	10	47	47	47	40	2.7	0.507890
519	Mookie Betts	29	180	9	50	49	50	49	3.2	0.466723
1103	Martin Perez	31	200	11	10	10	0	10	2.8	0.460149
504	Mike Trout	30	235	12	48	46	48	45	2.5	0.440299
805	Bryce Harper	29	210	11	46	46	46	8	1.8	0.380156
1009	Paul Goldschmidt	34	220	12	50	50	50	44	3.1	0.367002
998	Nolan Arenado	31	215	10	50	49	50	46	2.9	0.331921
429	Justin Verlander	39	235	17	10	10	0	10	1.9	0.315348
526	Freddie Freeman	32	220	13	52	52	52	52	1.6	0.223155
554	Sandy Alcantara	26	200	6	11	11	0	11	2.8	0.217820
1020	Yadier Molina	39	225	19	32	29	31	32	0.0	0.217663
442	Zack Greinke	38	200	19	10	10	0	10	0.0	0.215126
398	Jose Altuve	32	166	12	37	37	37	37	1.4	0.206260
706	Max Scherzer	37	208	15	8	8	0	8	1.6	0.196238
534	Clayton Kershaw	34	225	15	5	5	0	5	1.1	0.193754
1031	Adam Wainwright	40	230	17	10	10	0	10	1.4	0.173999
129	Xander Bogaerts	29	218	10	51	49	51	48	1.9	0.173201
363	Miguel Cabrera	39	267	20	44	42	44	0	0.0	0.168546
733	DJ LeMahieu	33	220	12	45	42	45	40	1.6	0.161645

Figure 13: Top 20 probability of All-Star in 2022 season

5 Conclusion

According to the recent news, only 11% adults listed baseball as their favorite sport to watch in a 2021 Washington Post poll (CNN Sports). Compares to 34% football it is far behind people's interests. But still Major League Baseball (MLB) is the highest level of professional baseball in the world and accounts for some of the most popular international sporting events. That is why many scholars have researched predicting the outcome of MLB matches.

Unlike that, we had observed which players are most likely to be in the next season's MLB ALL-Star game. All-Stars are selected by fans for starting fielders, by managers for pitchers, and by managers and players for reserves. Being in the All-Star game doesn't always mean they are the best players. But of course they are evaluated by their performance.

We figured out most of the All-Star players were from the United State but also there were many international players take an active part and were nominated for the All-Star game. The highest correlation of Wins Above Replacement (WAR) was a statistical summary of a player's total contributions to their team. Next was the year of experience. Also, from feature coefficients, we found a similar trend.

Lastly, we made a logistic regression modeling to find which players are most likely to be involved in the All-Star game. Also based on the 2021 season data set we applied our model to predict 2022 season All-Stars.

- Manny Machado, Jose Ramirez, Mookie Betts, Martin Perez, Mike Trout, Bryce Harper, Paul Goldschmidt, Justin Verlander, Nolan Arenado, Freddie Freeman, Yadier Molina, Zack Greinke, Sandy Alcantara, Max Scherzer, Clayton Kershaw, DJ LeMahieu, Jose Altuve, Xander Bogaerts, Rafael Devers, Miguel Cabrera

They are the twenty players we expect to be on the 2022 Major League Baseball All-Star Game Tuesday, July 19, 2022. And we will compare our results on the day that players get selected to participate in the All-Star Game.

6 Reference

Enten, Analysis by Harry. "Why Is Baseball No Longer America's Game?" CNN, Cable News Network, 7 Apr. 2022, https://www.cnn.com/2022/04/07/sport/mlb-opening-day-baseball-popularity-spt-intl/index.html: :text=Just

"2021 Major League Baseball Season Summary." Baseball, https://www.baseball-reference.com/leagues/majors/2021.shtml.