The Statistics of Baseball
Pond Skimmers

December 7, 2022

Thomas Chant, Teague Dilbeck, Ryan Fidel, Anderson Salisbury

Contents

1 Introduction

Major League Baseball does not have a salary cap on teams. This can create markets for the best players that are out of the price range of lower budget teams. This can make an unfair advantage for teams which is mitigated through competitive balance funding for small market teams and luxury taxation for larger market teams that have high payrolls.

Each season, there is an all-star team selected around halfway through the season. They are voted on by fans and not by experts. This means that the very best players that season are not guaranteed to be selected as all-stars. However, for the sake of our project, we are assuming that all-stars are a representation of the best and most valued players in the MLB. Therefore, we set out to determine what statistical measures are most influential to being selected as an all-star, if all-stars help teams win games, and if they are worth the high price that they come with.

2 Cleaning Process

Cleaning the data started with Narrowed to 1985 to 2015 to ensure accurate data within timeframe. Then, teams changed names over the years, so each team name had to be combined to a singular name, the current one. This was done over many data sets. Since each data set was an individual excel file, Python had to be used to combine the files even further. Each player and team were combined so that all the data on each sheet could be analyzed together. This ensured that the predictive models had as much data as possible while operating.

3 Statistical Analysis

The beginning statistical analysis took place looking at the all-stars by team and comparing these to the salary by team. The data was relatively normal, with only one outlier in both data sets. The outlier in each was the New York Yankees, as they had the greatest salary and all-stars out of every year. This meant the data was in a good place to biggin hypothesis testing.

For the hypothesis tests, first I had to break down various files of data into tables that I would be able to extract information out of for the tests. This part begins by making a table to categorize the total salaries for each team listed. This started in the year 1985 and concluded in 2015, making quite a large table. The 1994 season was excluded because there was a league-wide strike that year that caused the playoffs and part of the season to be missed. Next, a table of the same format (columns for each year and rows for each team) was created for counting the number of all-stars that were on each team each year. After this, I ran a function through a new table that served as a separator for teams that made the playoffs and did not. This was done by counting the amount of time a team populated in playoff matchups each year. If the team did not show up in a year, the cell prints “0”. This is important for the next step which used the same table as before to show the amount of wins a team had in the playoffs each year. Here, if a team did not make the playoffs, the equivalent cell in the previous table populated as “0”. This then showed up as “-“in the new cell, rather than “0” to differentiate teams that didn’t make the playoffs and those that did but didn’t win a game. Although I didn’t use this table for any of the tests, it was still interesting to visualize playoffs wins and how few teams make it to that stage of the season. The next table I made showed the number of all-stars on each playoff team through the studied time period with a “-“

populating in the cells of teams that missed the playoffs. The final table created was the opposite of this, showing the number of all-stars on teams that were left out of the playoffs.

After all that data crunching, it's time to get into hypothesis testing. This first test was a proportion test that had aa alternate hypothesis stating that 33% of all-stars in a given year will qualify for the playoffs. Inversely, the null hypothesis was that less than or equal to 33% of all-stars would make the playoffs. This threshold was chosen because between 1985 and 2015, the percentage of teams that qualified for the playoffs varied from 18% to 33% (as playoffs formats changed and expansion teams were added). If more than 33% of all-stars made the playoffs, then that is an indicator that all-stars do help teams be successful. The mean percentage in the time period was 38% and as expected, this yielded a low P-value. This does suggest that a higher portion of all-stars make the playoffs than other players, as we can reject the null hypothesis.

The second hypothesis test was a difference test that investigated the average number of all-stars on playoff teams versus non-playoff teams. The mean of the acquired data was 2.0 more all-stars on playoff teams with a high of 3.4 and a low of 0.7 (1987 was the only season that had a difference lower than 1.0). For this test, the alternate hypothesis was that the average number of all-stars on playoff teams in 1.75 more than non-playoff teams. This made the null hypothesis – the average number of all-stars on a playoff team is equal to or less than 1.75 more than non-playoff teams. The result of this test was a very small P-value under the 0.05 threshold. Therefore, the null hypothesis was rejected. This means that there is enough evidence to suggest that there is on average more than a 1.75 all-star difference between playoff and non-playoff teams.

An additional difference hypothesis test was provided. It was observed that the mean difference in total team salary between playoff teams and those that didn't make the playoffs was $19,020,852. That led to the null hypothesis – the difference in salary between playoff teams and non-playoff teams is $19 million or less. This alternate hypothesis put to test here was – the difference in salary between playoff teams and non-playoff teams is greater than $19 million. The P-value resulting from this t-test was significantly greater than 0.05, indicating that we do not have enough evidence to reject the null hypothesis.

4 Tableau Analysis

Tableau is an excellent tool for organizing data and developing clear and concise visual data representations. For our project, we utilized Tableau to create three key models. The first model represents our All-Star data geographically across the United States and Canada. Our second model emphasizes the number of All-Stars coming from each baseball team. Our final model identifies players that made multiple All-Star appearances. All three models are laid out in a concise Tableau dashboard. The data cleaning, methods, and filters used to create our three models and dashboard are discussed below. Are models, dashboards, and filter layout can be found in the appendix.

The main Tableau data file we used includes All-Star data between 1933 and 2015. The data set provides players associated with All-Star teams for each year. To clean our data, we used a conversion table to determine the team associated with each player. A second conversion table was utilized to associate baseball teams with their respective cities, states, league, and region.

After cleaning our data in Tableau, we developed our geographical model. Our geographical model is useful for people who want a bird's eye view of our data. Our geographical model identifies each team on a map. The identified geographical locations are represented by varying sized and colored circles. The circle sizes directly correlate with the number of All-Stars from each team location. The circle color is measured on a heat map scale where warmer colors represent teams with a higher number of all stars.

Our second model takes a closer look at the team data represented in our geographical model. The All-Star teams model shows the All-Stars produced by each team over a specific time. Building off our All-Star team model, our third model, All-Star players, shows which players were selected as All-Stars. This model sums each player's All-Star selection status over a specified time frame.

Our Dashboard pulls together all three models for an all-in-one interface. Several key filters are utilized to make our model seamless and connected. A yearly time frame filter is shared between all models so that each model on the dashboard represents the same time frame. We also included a shared league filter for all models which functions the same as the time frame filter. Finally, to clean our All-Star players model, we included an All-Star cutoff slider. The slider lets the dashboard user set a minimum limit for All-Star appearances. Be setting the slider to two, players with less than two appearances over the specified time frame will be left out of the All-Star player bar chart.

To summarize, we utilized Tableau to create visually appealing models from our data. Our models efficiently highlight a different aspect of our data for any user. Our dashboard combines our models and links them through shared filters like timeline and league.

5 Python Analysis

Four total models were constructed in Python. Two random forest classifier models were developed for predicting whether a player was selected for that year's All-Star Team: one for pitchers, and one for batters. Two random forest regression models were developed for predicting a player's salary for a given year: one for pitchers, and one for batters. These models were selected due to their superior performance, and for their being generally robust and less prone to overfitting when compared to other models. A train / test split of 80 / 20% was used to train the models and evaluate their performance. Four imaginary players were created with feature information that would simulate a dominant pitcher and a weak pitcher, as well as a dominant batter and a weak batter.

The batter regression model featured a training R^2 value of 0.910, and a test R^2 of 0.407, the pitcher regression model featured a training R^2 value of 0.915, and a test R^2 of 0.374. Both models seem to have severe issues with overfitting. For pitchers, the most important feature for predicting salary was year, followed by games finished and strikeouts. For batters, the most important feature for predicting salary was again year, followed by games, and then whether the player was on the all-star team. The great imaginary batter was predicted to have an annual salary of $119.8 million, while the weak batter was predicted to have an annual salary of $7.2 million, in large part because the batter was allowed to play in most of the games for a season. The dominant pitcher had a predicted salary of $95.6 million, and the weak pitcher had a predicted salary of $1.1 million, which is fairly comparable to other weak starting pitchers in the MLB.

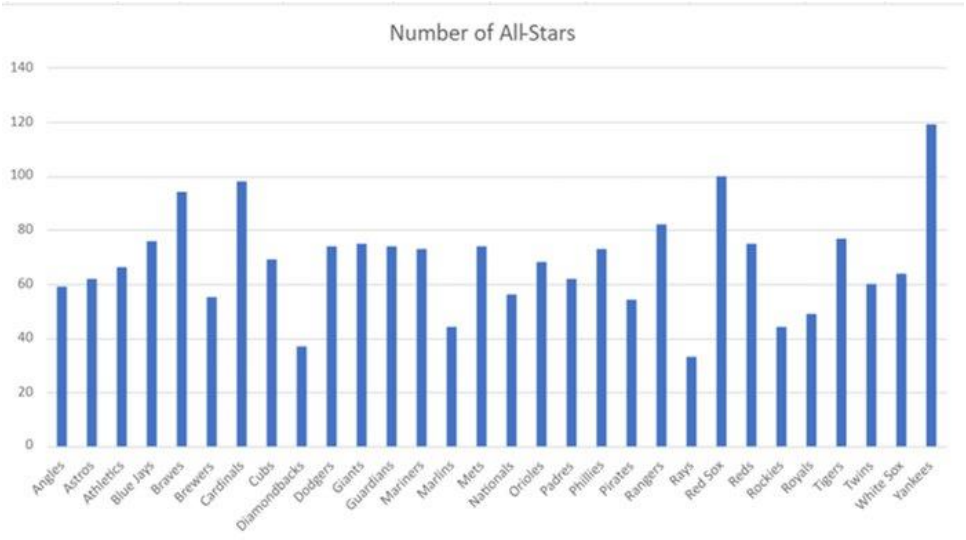
The classifier batter model featured a training mean accuracy score of 0.935, and a test mean accuracy score of 0.933. The classifier pitcher model featured a training mean accuracy score of 0.956, and a testing mean accuracy score of 0.957. For pitchers, the most important features for classifying all stars were saves, wins, and strikeouts. For batters, the most important features were at bats, runs, and home runs. Our dominant simulated pitcher was predicted to be an all-star, and our weak pitcher was not, so our model predicted perfectly there. Our dominant batter was predicted to be an all-star, and our weak batter was not, so our model was correct there as well

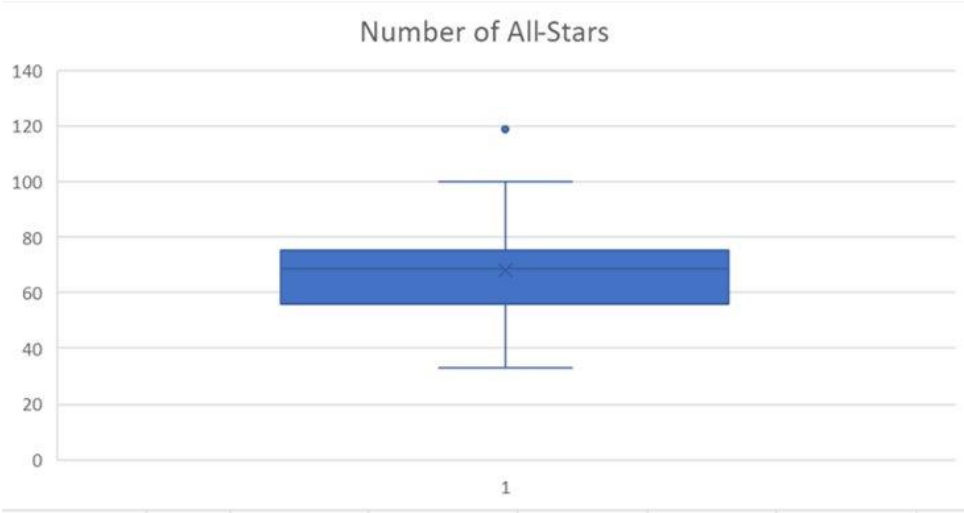
6 Conclusion

The data for MLB baseball teams was broken down into the date range of 1985 to 2015. There was a strong correlation between the number of all-stars on a team to the success of a team. The same was true for salary and success. The average salary of postseason baseball teams was almost 20 million dollars more than those who did not make the postseason. There were also predictive models made that could consider pitcher and hitter data and determine if they would be an all-star, and how much they would be paid. Overall, the models worked very well to predict these statistics.

Appendix

Team Name	# of All-Stars	Rank
Yankees	119	1
Red Sox	100	2
Cardinals	98	3
Braves	94	4
Rangers	82	5
Tigers	77	6
Blue Jays	76	7
Giants	75	8
Reds	75	8
Dodgers	74	10
Guardians	74	10
Mets	74	10
Mariners	73	13
Phillies	73	13
Cubs	69	15
Orioles	68	16
Athletics	66	17
White Sox	64	18
Astros	62	19
Padres	62	19
Twins	60	21
Angles	59	22
Nationals	56	23
Brewers	55	24
Pirates	54	25
Royals	49	26
Marlins	44	27
Rockies	44	27
Diamondbacks	37	29
Rays	33	30





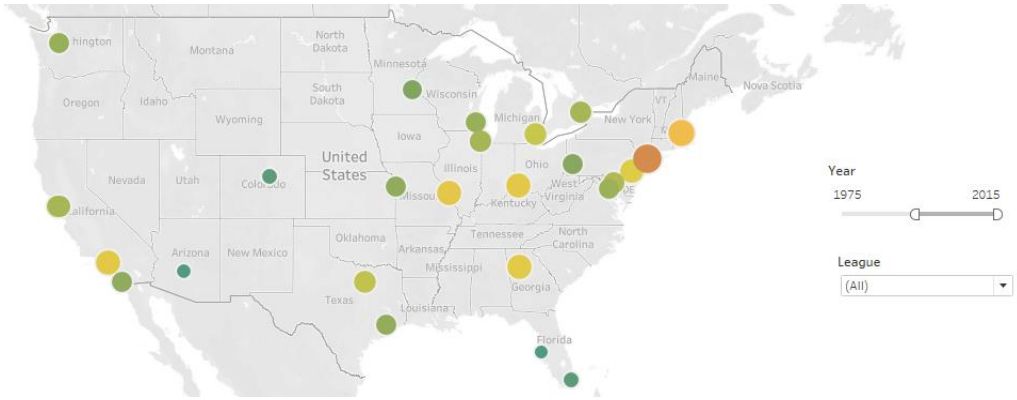
Team Name	Team Salary	Rank
Yankees	$ 3,495,871,291	1
Red Sox	$ 2,613,804,335	2
Dodgers	$ 2,453,558,703	3
Mets	$ 2,117,310,904	4
Phillies	$ 2,094,048,800	5
Giants	$ 2,004,454,588	6
Cubs	$ 1,975,712,625	7
Braves	$ 1,954,728,034	8
Tigers	$ 1,943,482,437	9
Cardinals	$ 1,894,872,832	10
Rangers	$ 1,813,833,287	11
White Sox	$ 1,812,238,843	12
Mariners	$ 1,795,352,131	13
Orioles	$ 1,745,244,871	14
Blue Jays	$ 1,683,867,753	15
Angles	$ 1,576,656,031	16
Reds	$ 1,547,847,634	17
Guardians	$ 1,492,679,894	18
Astros	$ 1,480,349,357	19
Rockies	$ 1,389,599,239	20
Brewers	$ 1,382,656,857	21
Twins	$ 1,377,992,124	22
Padres	$ 1,295,941,262	23
Royals	$ 1,289,573,993	24
Nationals	$ 1,279,151,254	25
Diamondbacks	$ 1,271,809,228	26
Athletics	$ 1,265,607,954	27
Pirates	$ 1,083,439,717	28
Marlins	$ 936,589,305	29
Rays	$ 832,632,108	30

Hypothesis test for proportion			
Type of alternative hypothesis	One-tailed - greater than	Test:	
Hypothesized proportion	0.33	H0: p > 0.33	Equal to or less than 33% of All-Stars don't make the Playoffs
Significance Level	0.05	Ha: p <= 0.33	More than 33% of All-Stars make the Playoffs
Sample size	1982		
Number with property of interest	744	**The P-value is smaller than 0.05 so we can reject the null hypothesis and accept the alternate hypothesis.**	
Sample proportion	0.375	**Meaning that there is enough evidence to suggests that more than 33% of All-Stars make the Playoffs.**	
Standard error of proportion	0.011		
Test statistic (z value)	4.296		
p-value	0.00000868		
Result	Reject H0		

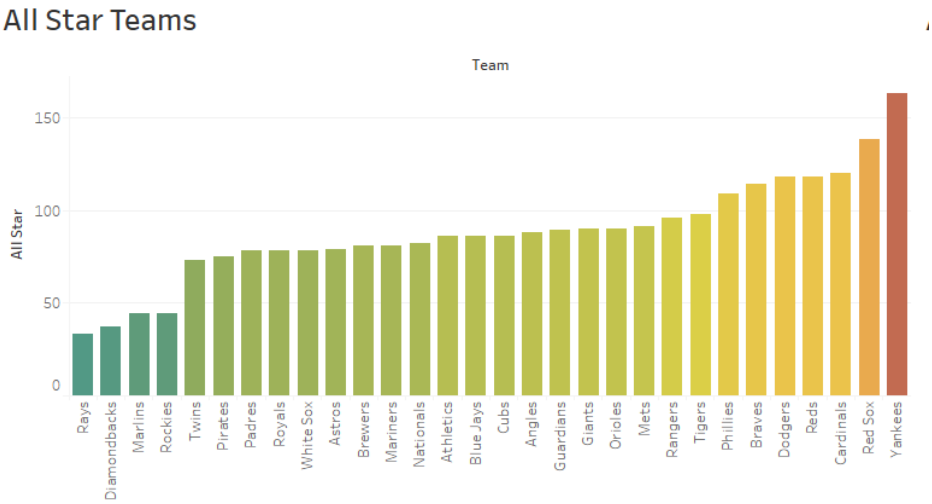
Hypothesis test for difference		Test:	
Type of alternative hypothesis	One-tailed - greater than	H0: Difference > 1.75	The average number of All-Stars on playoff teams is equal to 1.75 or lower than non-playoff teams.
Hypothesized difference	1.75	Ha: Difference <= 1.75	The average number of All-Stars on playoff teams is 1.75 higher than non-playoff teams.
Signifficance level	0.05		
Sample size	29	**The P-value is smaller than 0.05 so we can reject the null hypothesis and accept the alternate hypothesis.**	
Sample mean diff	2	**Meaning that there is enough evidence to suggest that the average number of All-Stars on playoff teams is 1.9 higher than non-playoff teams.**	
Sample std dev of diff	0.62		
Standard error	0.11		
Test statistic (t-value)	2.15		
Degrees of freedom	28		
p-value	0.020299181		
Result	Reject H0		

Hypothesis test for difference		Test:	
Type of alternative hypothesis	One-tailed - less than	H0: Difference > $19 M	The difference between Playoff teams and non-Playoff teams salary is equal to or less than $19 M.
Hypothesized difference	19000000	Ha: Difference <= $19 M	The difference between Playoff teams and non-Playoff teams salary is more than $19 M.
Signifficance level	0.05		
Sample size	29	**The P-vale is greater than 0.05 so we fail to reject the null hypothesis.**	
Sample mean diff	19020852	**Meaning that there is not enough evidence to suggest that the average distance between playofff teams and non-playoff teams is more than $19 M.**	
Sample std dev of diff	27126265.82		
Standard error	5037221.1		
Test statistic (t-value)	0		
Degrees of freedom	28		
p-value	0.498363242		
Result	Fail to Reject H0		

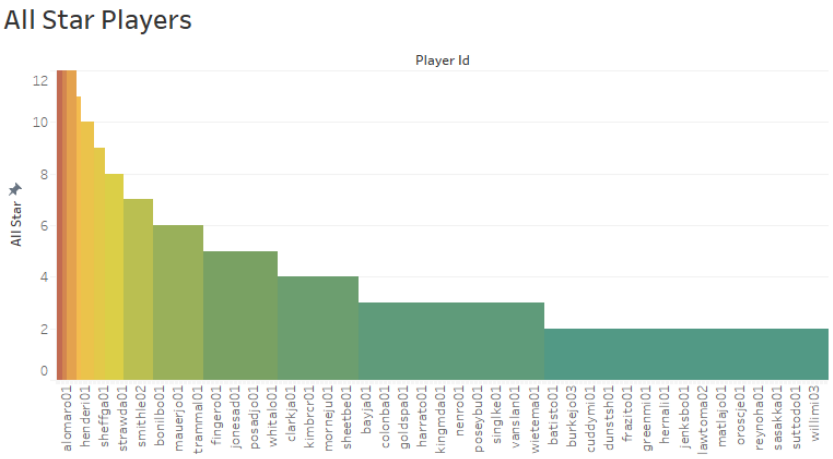
ALL-Star Geographic Model:



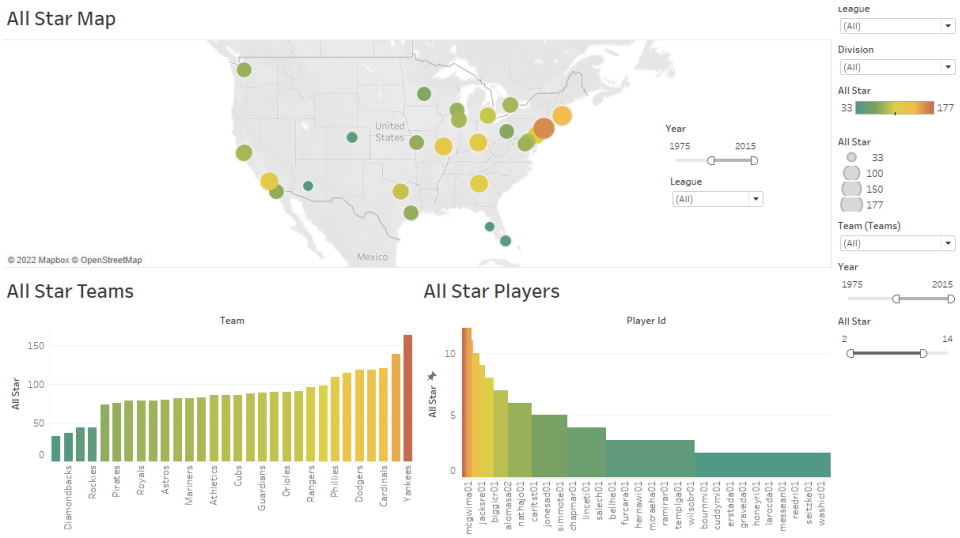
All-Star Teams Model:



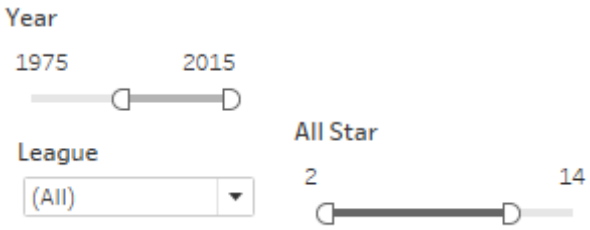
All-Star Players Model:

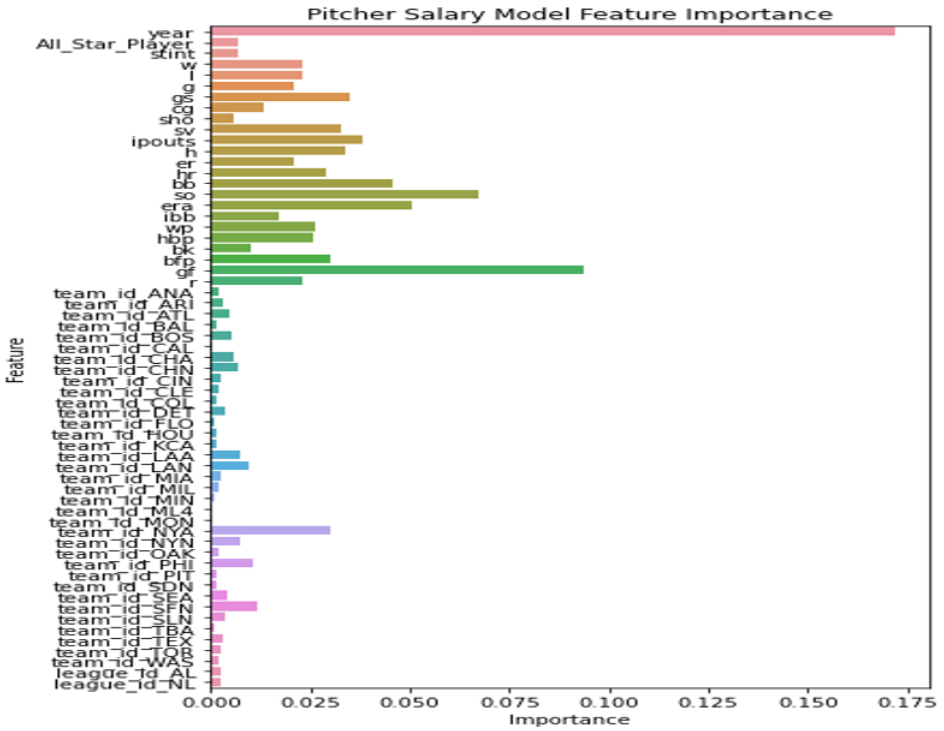


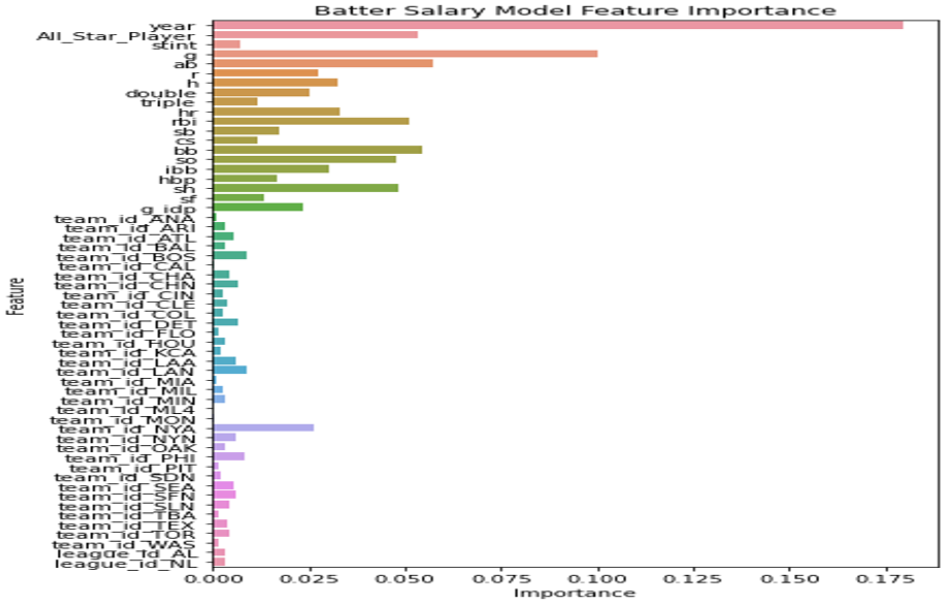
All-Star Dashboard:

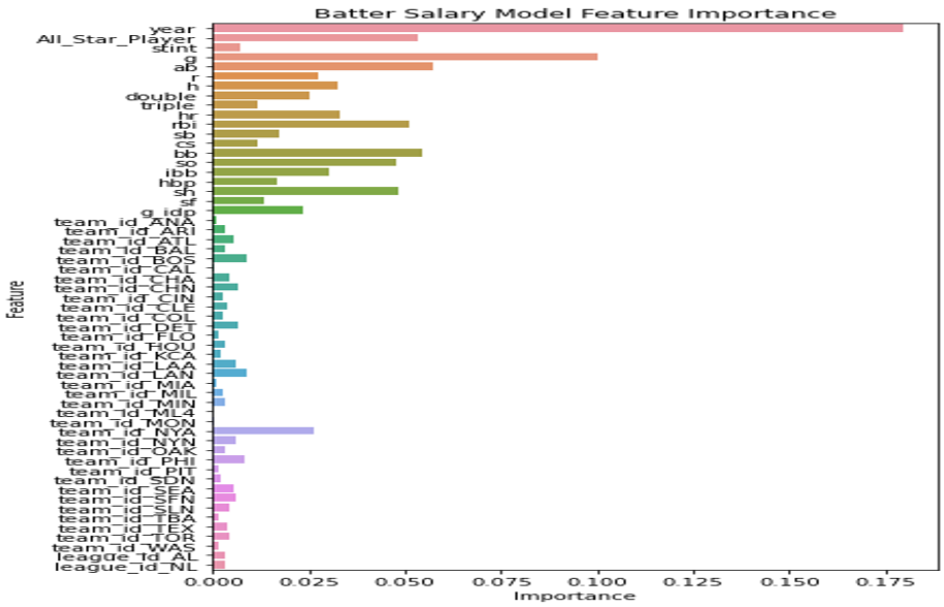


Dashboard Key Filters:




Pitcher Salary Model Feature Importance
Feature
Importance
year
All_Star_Player
stint
w
l
g
gs
cg
sho
sv
ipouts
h
er
hr
bb
so
era
ibb
wp
hbp
bk
bfp
gf
r
team_id_ANA
team_id_ARI
team_id_ATL
team_id_BAL
team_id_BOS
team_id_CAL
team_id_CHA
team_id_CHN
team_id_CIN
team_id_CLE
team_id_COL
team_id_DET
team_id_FLO
team_id_HOU
team_id_KCA
team_id_LAA
team_id_LAN
team_id_MIA
team_id_MIL
team_id_MIN
team_id_ML4
team_id_MON
team_id_NYA
team_id_NYN
team_id_OAK
team_id_PHI
team_id_PIT
team_id_SDN
team_id_SEA
team_id_SFN
team_id_SLN
team_id_TBA
team_id_TEX
team_id_TOR
team_id_WAS
league_id_AL
league_id_NL
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175


Batter Salary Model Feature Importance
Feature
Importance
year
All_Star_Player
stint
g
ab
r
h
double
triple
hr
rbi
sb
cs
bb
so
ibb
hbp
sh
sf
g_idp
team_id_ANA
team_id_ARI
team_id_ATL
team_id_BAL
team_id_BOS
team_id_CAL
team_id_CHA
team_id_CHN
team_id_CIN
team_id_CLE
team_id_COL
team_id_DET
team_id_FLO
team_id_HOU
team_id_KCA
team_id_LAA
team_id_LAN
team_id_MIA
team_id_MIL
team_id_MIN
team_id_ML4
team_id_MON
team_id_NYA
team_id_NYN
team_id_OAK
team_id_PHI
team_id_PIT
team_id_SDN
team_id_SEA
team_id_SFN
team_id_SLN
team_id_TBA
team_id_TEX
team_id_TOR
team_id_WAS
league_id_AL
league_id_NL
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175



Our model predicts that our great batter will be paid $11980968.18
Our model predicts that our bad batter will be paid $7236464.79

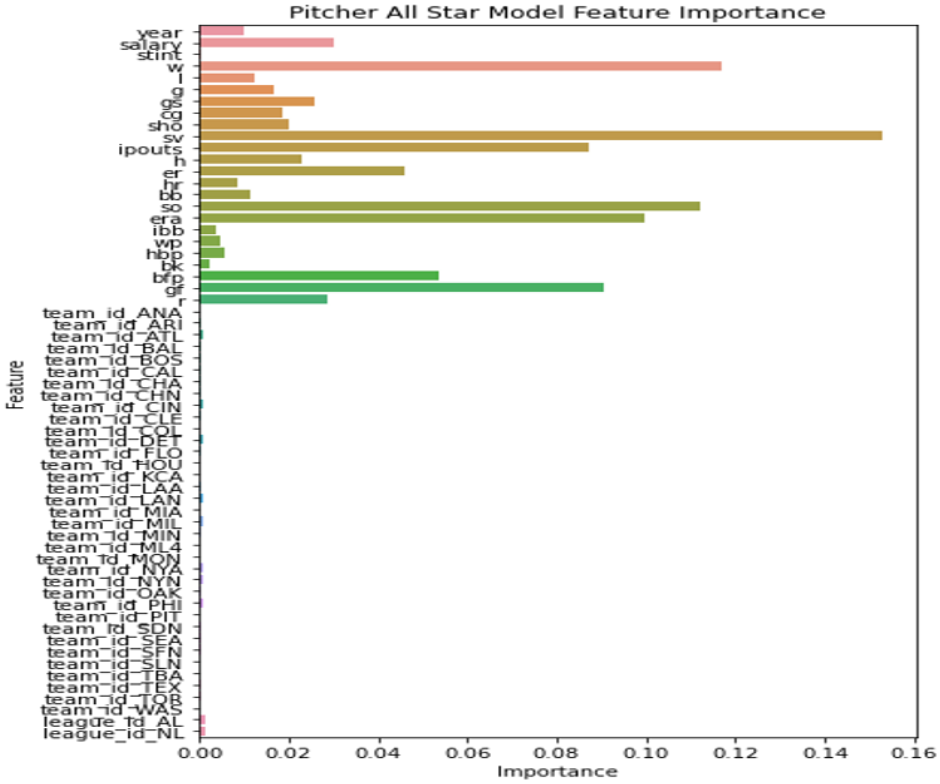
	year	All_Star_Player	stint	g	ab	r	h	double	triple	hr	...
0	2015.0	1.0	1.0	163.0	650.0	120.0	220.0	50.0	15.0	70.0	...
1	2015.0	1.0	1.0	100.0	400.0	5.0	10.0	5.0	1.0	10.0	...

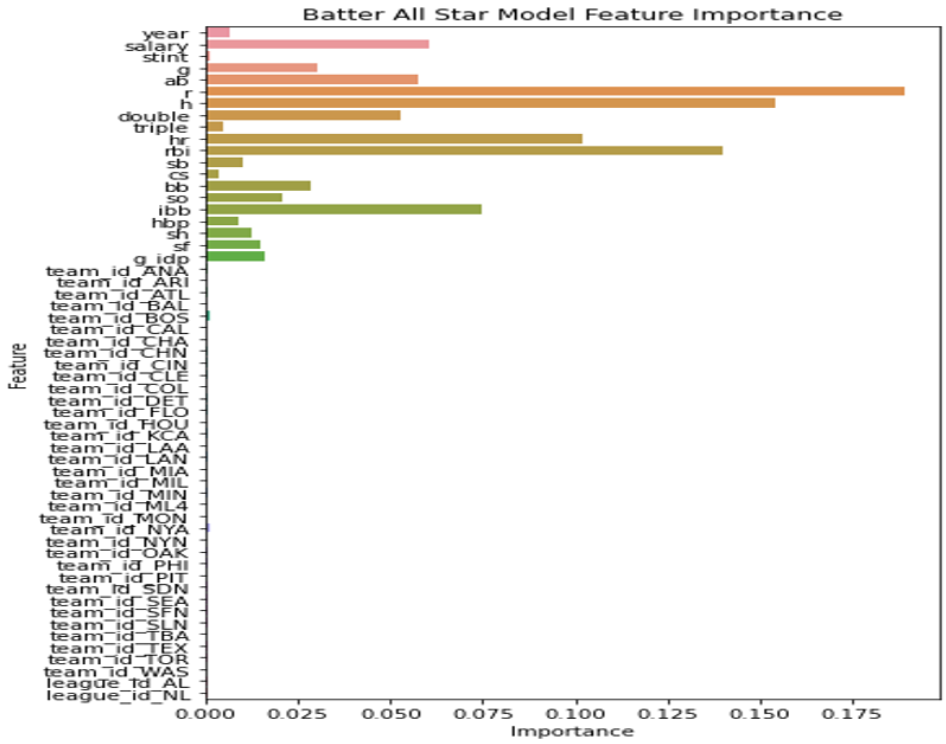
Our model predicts that our great pitcher will be paid $9555275.66
Our model predicts that our bad pitcher will be paid $1131121.74

	year	All_Star_Player	stint	w	l	g	gs	cg	sho	sv	...	team_id_SDN
0	2015.0	1.0	1.0	32.0	0.0	32.0	32.0	32.0	32.0	10.0	...	0.0
1	1985.0	0.0	1.0	2.0	18.0	20.0	20.0	2.0	0.0	0.0	...	0.0

	precision	recall	f1-score	support
0.0	0.94	0.99	0.97	4675
1.0	0.74	0.19	0.30	373
accuracy			0.94	5048
macro avg	0.84	0.59	0.63	5048
weighted avg	0.92	0.94	0.92	5048

	precision	recall	f1-score	support
0.0	0.96	1.00	0.98	2435
1.0	0.91	0.22	0.35	142
accuracy			0.96	2577
macro avg	0.93	0.61	0.66	2577
weighted avg	0.95	0.96	0.94	2577






```
Our model predicts that our great batter will be an all star!
Our model predicts that our bad batter will not be an all star.
```

	year	salary	stint	g	ab	r	h	double	triple	hr	...
0	2015.0	11000000.0	1.0	163.0	650.0	120.0	220.0	50.0	15.0	70.0	...
1	2015.0	5000000.0	1.0	100.0	400.0	5.0	10.0	5.0	1.0	10.0	...

```
Our model predicts that our great pitcher will be paid $9555275.66
Our model predicts that our bad pitcher will be paid $1131121.74
```

	year	All_Star_Player	stint	w	l	g	gs	cg	sho	sv	...	team_id_SDN
0	2015.0	1.0	1.0	32.0	0.0	32.0	32.0	32.0	32.0	10.0	...	0.0
1	1985.0	0.0	1.0	2.0	18.0	20.0	20.0	2.0	0.0	0.0	...	0.0