Ben Kaufmann

11/20/2022

Stat 419

Project Progress Report

Introduction

During these first few weeks of my project, I have mostly worked on researching and creating my data set. I suspect many other students built their project around a data set in mind, but due to the nature of my project I could not do this. It took a lot of effort to create my data set since there were a lot of external factors I had to consider. I have not started applying the statistical methods I plan to use yet, but I believe the extra research and data preparation I have done will expedite the coding process.

Dataset Work

To create my data set, I used the Stathead Baseball advanced search feature. This feature allowed me to query from the Stathead database which houses an exorbitant number of different statistics for every player in professional baseball history. I then selected the variables I wanted to be included and pruned out tuples which could create noise and affect the findings of the project. Here is a snippet of the data set.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	**Player**	**PA**	**dWAR**	**R**	**H**	**1B**	**2B**	**3B**	**HR**	**RBI**	**SB**	**BB**	**BA**	**OBP**	**SLG**	**OPS**	**OPS+**	**HOF**
2	Josh Gibson	2511	1.6	610	806	448	135	58	165	730	40	331	0.374	0.458	0.72	1.178	214	y
3	Babe Ruth	10626	-2.3	2174	2873	1517	506	136	714	2214	123	2062	0.342	0.474	0.69	1.164	206	y
4	Ted Williams	9792	-13.3	1798	2654	1537	525	71	521	1839	24	2021	0.344	0.482	0.634	1.116	191	y
5	Oscar Charleston	3962	-1.2	862	1218	750	242	80	146	861	209	484	0.364	0.448	0.614	1.062	184	y
6	Buck Leonard	2556	-1.4	538	728	448	138	47	95	545	34	396	0.345	0.452	0.59	1.042	180	y
7	Lou Gehrig	9665	-9	1888	2721	1531	534	163	493	1995	102	1508	0.34	0.447	0.632	1.08	179	y
8	Turkey Stearnes	4291	0.2	910	1319	793	228	112	186	1001	129	416	0.349	0.417	0.616	1.034	177	y
9	Rogers Hornsby	9481	13.9	1579	2930	1919	541	169	301	1584	135	1038	0.358	0.434	0.577	1.01	175	y
10	Mule Suttles	3649	-3.1	717	1089	624	214	72	179	877	84	370	0.34	0.41	0.619	1.029	172	y

The complete data set is submitted alongside this report.

Variable Selection

When selecting the variables that would be included in the dataset, many things had to be considered. I wanted to include statistics that best encapsulate the value of the batter and are statistically relevant to the Hall of Fame classification. Fortunately, standard baseball statistics do an adequate job of fulfilling this task. BA, OBP, SLG, OPS, R, SB, RBI, H, 1B, 2B, 3B, HR, BB, and PA are all commonly used to determine the value of hitters and were subsequently included in the data set. Other common statistics such as SO and CS were removed from the data set since they appeared as null values in the Stathead database for many of the older players since they were not always tracked. CS, for instance, only started being tracked in 1951 (Edes, 2013).

Cumulative vs. Non-Cumulative Variables

Cumulative statistics such as R, SB, RBI, H, 1B, 2B, 3B, HR, BB, and PA represent the production a player has over their career. Meanwhile, Non-Cumulative statistics such as BA, OBP, SLG, and OPS represent the efficiency a player has over their career. In sports, a player must have both high overall

production and efficiency to be considered for the Hall of Fame (Stueve, 2017). For this reason, I decided to include a healthy mix of both Cumulative and Non-Cumulative statistics to represent both qualities in the data set.

Comparing Players of Different Eras

Professional baseball has changed a lot over the 100+ years it's been played. Comparing players from different eras can be tricky since there have been varying levels of difficulty and league tendencies over time. OPS is a statistic that encompasses a hitter's value into a single number. This is helpful for comparing players who played in similar time periods but does not account for different time periods. OPS+ however, takes a player's OPS and compares it with the performance of other players who played at the same time (MLB). For example, Pete Browning and Aaron Judge have extremely different OPS values, but the same OPS+ values since Browning played at a time when there was less offensive production. OPS+ is included in the data set and will likely be weighted when statistical methods are applied.

Player	From	To	OPS	OPS+
Pete Browning	1882	1894	.869	163
Aaron Judge	2016	2022	.977	163

Defensive Statistics

In baseball, defense is difficult to evaluate. There are many statistics used to represent defensive value, but each one has its own shortcomings and isn't perfect (Malinowski, 2020). Baseball Reference has a complicated formula to compute a defensive statistic called dWAR (Baseball Reference). dWAR compares the defensive efficiency of a player to the league average and outputs a single number. A dWAR of 0 indicates a perfectly average defender. I decided to include this statistic in the data set since it has high praise from baseball statisticians. It was necessary for me to include dWAR since quality of defense is heavily judged when a player is considered for the Hall of Fame. dWAR will almost certainly be weighted so it has a greater impact on the results since it is outnumbered by offensive statistics.

Data Pruning

This data set required a lot of pruning to get rid of noise that could affect the results of the statistical methods. The first pruning I had to do was eliminate players with too small of a sample size. I did this by pruning out any player who had under 2,500 plate appearances. This threshold was used because I noticed that the vast majority of Hall of Fame players were over this mark. The Hall of Fame players that were not over this mark were almost always inducted for their off-the-field achievements, which of course has no statistical representation so it will be ignored.

In the old days of professional baseball (before the 1950's), many pitchers would play other positions and hit much more frequently than they do in the modern game. As a result, the data set was filled with pitchers who racked up over 2,500 plate appearances but were more-so inducted in the Hall of Fame for their pitching prowess rather than their hitting. To combat this, I pruned out any player who spent over 40% of their games as a pitcher rather than a position player. This number was used because it kept "hitters who pitched" like Babe Ruth in and excluded "pitchers who hit" that would skew the data set.

Class Imbalance

Class Imbalance became a moderate concern as I was creating the data set. According to machine learning expert Jason Brownlee, class imbalance becomes a problem when the minority class is out represented 1:100 and over (Brownlee, 2020). Since the number of Hall of Fame players is incredibly

small compared to the number of players not in the Hall of Fame, this became a potential issue. To steer clear of this potential problem, I limited the number of players not in the Hall of Fame in the data set. I decided to use 2,000 of these players so my class imbalance ratio would be about 1:10. When choosing which of these players to include/exclude, I decided to use the top 2,000 OPS+ players. OPS+ is a general hitting statistic that is best for representing a hitter's value in a single number. As a result, the lowest OPS+ hitter not in the Hall of Fame included in the data set ended up being around the same as the lowest OPS+ Hall of Fame player.

References

Baseball Reference. *Position player war calculations and details*. Baseball. (n.d.). Retrieved November 20, 2022, from https://www.baseball-reference.com/about/war_explained_position.shtml

Brownlee, J. (2020, January 14). *A gentle introduction to imbalanced classification*. MachineLearningMastery.com. Retrieved November 20, 2022, from https://machinelearningmastery.com/what-is-imbalanced-classification/

Edes, G. (2013, October 2). *Jacoby Ellsbury: Sultan of swipe*. ESPN. Retrieved November 20, 2022, from https://www.espn.com/boston/mlb/story/_/id/9757214/jacoby-ellsbury-quick-learner-craft-stealing-bases

Malinowski, I. (2020, April 15). *What is the best statistic for measuring defense?* DRaysBay. Retrieved November 20, 2022, from https://www.draysbay.com/2020/4/15/21219514/mlb-defense-best-statistic-baseball-prospectus

MLB. On-base plus slugging plus (OPS+): Glossary. MLB.com. (n.d.). Retrieved November 20, 2022, from https://www.mlb.com/glossary/advanced-stats/on-base-plus-slugging-plus

Stueve, W. (2017, September 30). *What are the trademarks of a NFL hall of fame player?* Bleacher Report. Retrieved November 20, 2022, from https://bleacherreport.com/articles/1284262-what-are-the-trademarks-of-a-nfl-hall-of-fame-player#:~:text=There%20isn't%20just%20one,by%20position%20and%20total%20value.