Sahil Patel

August 12, 2020

Predicting the 2018-2019 NBA MVP

Introduction

This project was inspired by a project made similarly by Daniel Bratulić. Our goal in the project was to alter Bratulić's project in order to find key differences in our results, from the data that was scrapped, to even how the results are analyzed and interpreted. The data was collected from BasketBallReference.com. In Bratulić's experiment, he only used some of the MVP voting data from a few seasons. In this project we want to collectively use all the data on MVP votings from 1985 to 2017 to train and validate our model

Data Preprocessing/Data Interpretation

The data was scraped from BasketballReference.com. First we obtained the data regarding MVP statistics. The dataset MVPstats in the python notebook is formatted as such:

Figure 1: Header for the MVP statistics category

	Rank	Player	Age	Tm	First	Pts Won	Pts Max	\
0	1.0	Larry Bird\birdla01	29	BOS	73.0	765.0	780	
1	2.0	Dominique Wilkins\wilkido01	26	ATL	5.0	407.0	780	
2	3.0	Magic Johnson\johnsma02	26	LAL	0.0	205.0	780	
3	4.0	Hakeem Olajuwon\olajuha01	23	HOU	0.0	193.0	780	
4	5.0	Kareem Abdul-Jabbar\abdulka01	38	LAL	0.0	135.0	780	

	Share	G	MP	PTS	TRB	AST	STL	BLK	FG%	3P%	FT%	WS	\
0	0.981	82	38.0	25.8	9.8	6.8	2.0	0.6	0.496	0.423	0.896	15.8	
1	0.522	78	39.1	30.3	7.9	2.6	1.8	0.6	0.468	0.186	0.818	10.8	
2	0.263	72	35.8	18.8	5.9	12.6	1.6	0.2	0.526	0.233	0.871	12.1	
3	0.247	68	36.3	23.5	11.5	2.0	2.0	3.4	0.526	NaN	0.645	9.5	
4	0.173	79	33.3	23.4	6.1	3.5	0.8	1.6	0.564	0.000	0.765	10.8	

	WS/48
0	0.244
1	0.170
2	0.226
3	0.186
4	0.197

As you can see we have the rank in terms of MVP placement for that player's respective season, Age, team(Tm), voting points won (Pts Won), award share (Share), games played (G), minutes played (MP), points per game (PTS), total rebounds a game (TRB), assists per game (AST), steals per game (STL), blocks per game (BLK), field goal percentage (FG%), three point percentage (3P%), free throw percentage (FT%), win share (WS) and win share per 48 min (WS/48). Then we grabbed the player stats for all players in the 2018 - 2019 season along with advanced statistics like win share in order to create testing data to predict the 2018 - 2019 MVP. Because this is a ranking task, it would be sensible to first determine our target and features. The target chosen was award share, which represents the percentage of votings earned for each player in the MVP running for the season. Our features run into some complications. With the statistics we have, mathematically, can be considered redundant. For example win share is a statistics that is used by this formula:

Figure 2: Win Share equation given by BasketBallReference.com

$$(points\ made) - .92 * (points\ per\ possesion) * (offensive\ possesions) = marginaloffense$$

$$(points\ per\ game) * ((teampace)/(leaguepace)) = marginalpointsperwin$$

$$(marginaloffense)/(marginalpointsperwin) = Winshare$$

This equation essentially takes the points made by a team and calculates the offensive contribution a player makes to a win for a team. Marginal offense takes into account points per game, points per possession and offensive possession which takes into account turnovers and other measurements of efficiency on possession such as assists. Because of this, some of the data in the MVPStats category may be redundant. In order to define which parameters actually have a

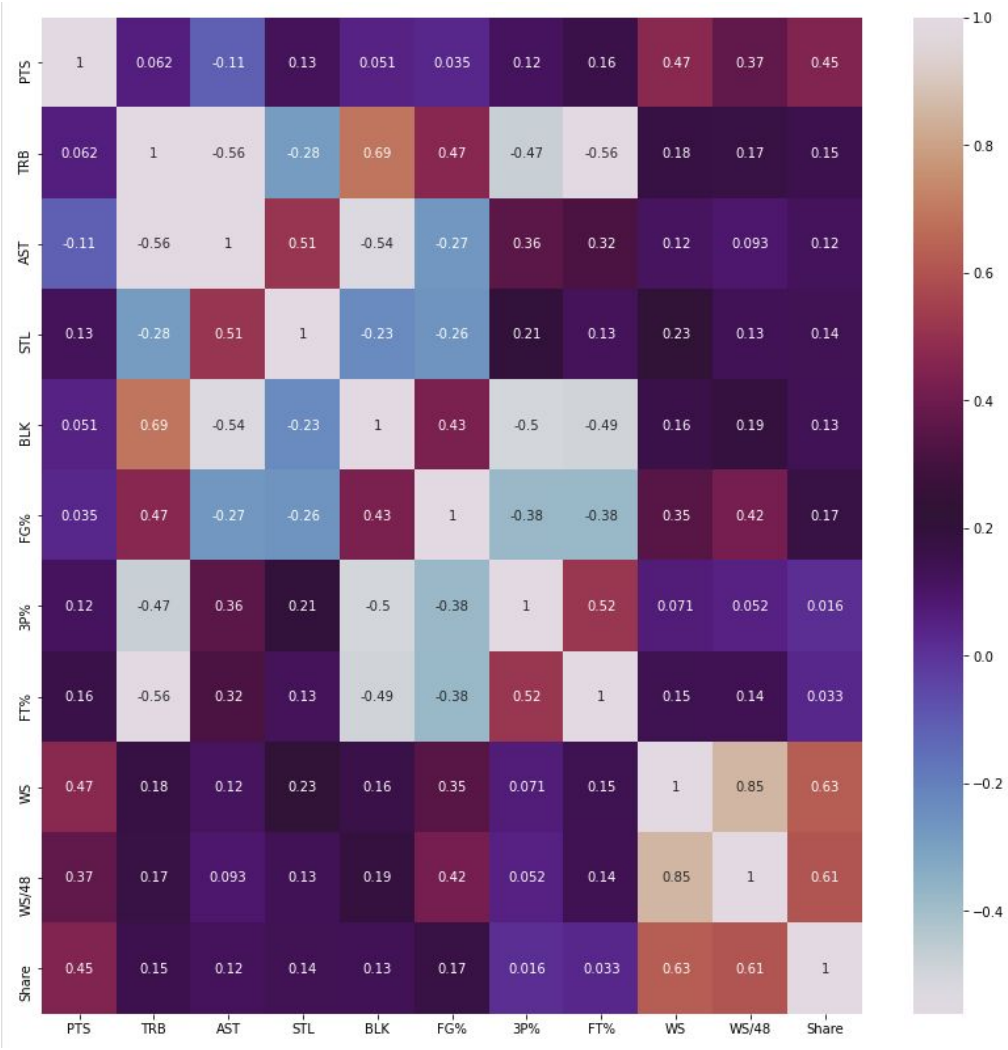
high relevance to the ranking task, we can find the parameters that are highly correlated with award share. The way this was done was by using SelectKBest, a method provided by SciKit. This method takes each parameter in question and compares it to a target, in this case awards shares. The reason SelectKBest was chosen was because it uses the ANOVA F-test. The F-test uses a scaled sum of squares in order to analyze the variance of the parameters to analyze if a parameter is independent or highly correlated with the target variable. Essentially, if a parameter's SelectKBest value is low, it means it is more independent from the target parameter and has less underlying influence on said target parameter. This means that we are looking for the parameters that are correlated with award share. When we select the 5 best k classes to find the most correlated parameters we get these results.

Figure 3: Scores from the F-test with respect to award share

```
WS: 0.3096
WS/48: 0.2835
PTS: 0.1898
TRB: 0.0756
AST: 0.0456
FT%: 0.0409
STL: 0.0369
3P%: 0.0029
FG%: 0.0016
BLK: 0.0000
```

And when we put the values on a heatmap we get a visual representation of the data above

Figure 4: Heatmap of the parameters based on their scores



By looking at the bottom row, we can see again how each parameter correlates with win share. With this the data can be reduced to less parameters to increase efficiency. In this case the features chosen to predict award share values are points per game, total rebounds per game, assists per and both winshare statistics. This is especially beneficial for training the model because it reduces the training data size by half, reducing the time while still maintaining close to the same amount of parameters.

Model Training

Now that the data has been prepared and the parameters that have significant importance are considered, a model must be selected for the task. Since this is a ranking problem, a regression would be an appropriate method. The reason is that with each MVP candidate, the award score is determined by their average stats for the season, and with the data considered, a regression would ideally learn the underlying relations of the data to the award share. There are multiple regressions that would work for this task, however, a Ridge Regression would be ideal for this task. The main reason for this is that Ridge Regression uses a lambda value in order to create some bias on the data. As shown with the heatmap, besides stats like win share and points per game there is not heavy correlation for the other parameters. In this case, we have high amounts of variance on those values. Take for example, Steve Nash. Steve Nash averaged around 19 points in his NBA career but averaged 10 assists as well. In fact for the majority of MVP winners, they averaged beyond 20 points per game but assists varied widely based on things like position and playstyle of each player. Ridge Regression would be beneficial because the high amount of variance on the other parameters like assists can be mitigated by implementing some bias by using the lambda value. In this case the lambda value was set to 0.01. When comparing the accuracies of a logistic regression and ridge regression, ridge regression scored a higher accuracy than logistic regression by around 9 percent on the validation set. To train the Ridge Regression model, the data was split into a 60/40 split as that had resulted in the highest accuracy and recall.

Obtaining the Predictions

Once the model was trained and validated, the test set needed to be created. By creating a dataframe using regular and advanced statistics of every player in the NBA for the 2018-2019

season, the top 40 players with the highest winshare scores were selected for the test set. This was done in order to mimic how the NBA MVP voting process is done because only the top 40 players in the league are considered for voting in the race for league MVP. The test data was then given to the model and a predictive award share score was given to the top 40 players. The results are shown below:

Figure 5: MVP predictions, along with the probability they will win the award

```
1. ('James Harden\\hardeja01', 0.5430978605511866)
2. ('Giannis Antetokounmpo\\antetgi01', 0.5305509950258177)
3. ('Anthony Davis\\davisan02', 0.2954994217120529)
4. ('Rudy Gobert\\goberru01', 0.2886506521905766)
5. ('Nikola JokiÄ‡\\jokicni01', 0.28283137803222835)
6. ('Paul George\\georgpa01', 0.24864435381610128)
7. ('Damian Lillard\\lillada01', 0.24780113916127455)
8. ('Kevin Durant\\duranke01', 0.23713910012792472)
9. ('Joel Embiid\\embiijo01', 0.21536457511438634)
10. ('Kawhi Leonard\\leonaka01', 0.20532973695965795)
```

These results are pleasing to say the least. All top MVP finalists that were announced (James Harden, Giannis Antetokounmpo, Paul George, Anthony Davis and Nikola Jokic) all appeared in the top five. The main difference is that the MVP for the 2018-2019 season was Giannis Antetokounmpo not James Harden. However, in the actual votes for MVP the race between Giannis and James were very close, as shown above. This can be interpreted that maybe there are some features not taken into consideration. One potential parameter could be popularity as the media plays a big role in influencing and deciding the NBA MVP.

Interpreting the Results

The results from the test data shows that potential MVPs can be predicted based on statistics. In MVP votings and the proliferation of advanced statistics, it has been seen that in the last 20 years and more, offensive contribution has played a key role in defining an MVPs success. In this

project, this was highlighted when win share was so highly correlated with award shares. This project also highlights that the trend of how votes are distributed have not changed much in the last 40 years. With Bratulić's project, his training and validation sets were much more focused based on a few seasons rather than 40 like the MVP dataset scraped for this project. This has its advantages in terms of efficiency but also our predictions ended up identical to his. As highlighted, it seems the MVP standard has not changed much at all.

Conclusion/Future Work

There can be some improvements made in continuation for this project. For example with the data, team win percentage can be added to show if a team is winning due to the contribution the player is making or we can scrap social media platforms to track the frequency of reports made around an individual to represent popularity. There are many steps forward that can make the results more accurate. Also this process could transfer well to predicting Defensive Player of the Year or even Rookie of the Year because the problem is similar with just different statistics. In fact, in terms of many voting processes in sports, because much of the data is now readily available, underlying statistical representations can be discovered which may reflect on how individuals are selected for certain achievements.