Predicting the MVP and DPOY of NBA in 19-20 season

Daoheng Liu

May 3, 2020

1. Introduction

1.1 Background

Several years ago, in my econometrics class, I wrote a regression analysis of NBA players' efficiency values and their various data, which was the first time I analyzed my favorite NBA. Now, in the online class "Applied Data Science Capstone", I need to do a data analysis through what I have learned from the class. At first, I thought that I may use Foursquare and other tools to continue the work of Week 3. However, when I saw the example named *Predicting the Improvement of NBA players*, my love of NBA pushed me to write a report of the prediction of the MVP and DPOY in this season.

1.2 Problem

Due to the influence of corona virus, the 19-20 season of NBA was only 60+ long and then ended. We do not know which team will win the champion and which players will win the awards. This project aims to predict two of these awards-MVP and DPOY-according to the data of this season.

1.3 Interest

In these days, many sports fans and also many professors of ESPN have made their prediction of the rewards. In the days when there is no game, all fans and players are enthusiastic about the award prediction, and they are eager to know the answer.

2. Data

2.1 Sources

All data are scraped from the website: http://www.stat-nba.com/. I screened the players who played 25 times or more games and averaged 25 minutes or more per game, because the selection requirements of NBA awards require players to play more than 40% games. Besides, each player's data contains player, season, team, point, rebound, assist and that sort of things.

2.2 cleaning

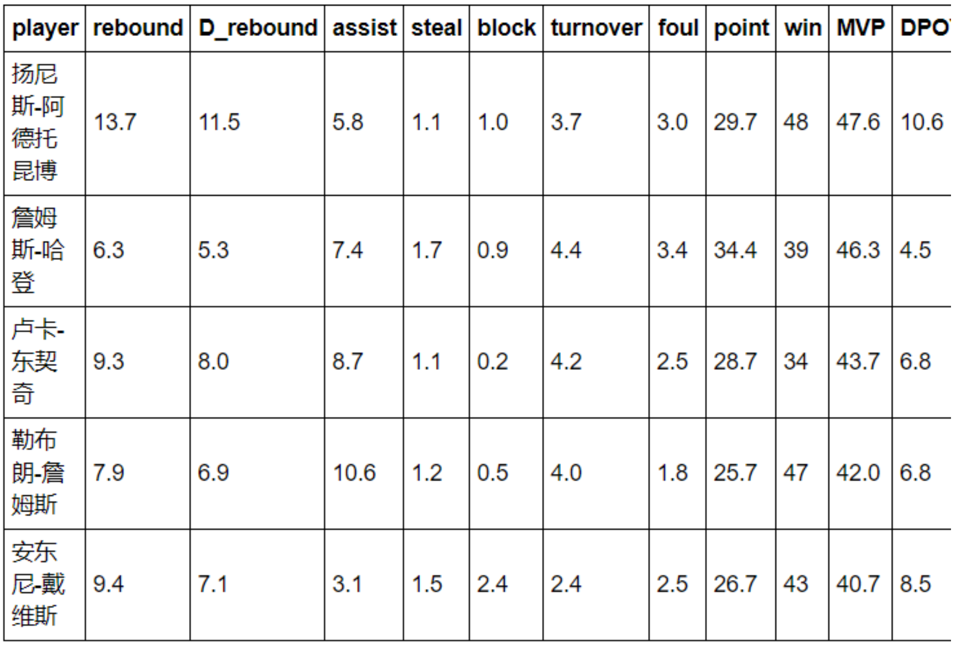
In the process of cleaning the data, I need to delete the extra data which I do not need at first, such as team, the number of starting and that sort of things. Secondly, the data collected from the website are Chinese, so I need to convert the language to English. Finally, I need to compute some index (MVP and DPOY) and add them to the data frame.

2.3 feature

In the final data frame, there are 151 rows (1 head and 150 players) and 12 columns (player, rebound, D_rebound, assist, steal, block, turnover, foul, point, win, MVP, DPOY). The explanations of MVP and DPOY are as following:

MVP = point + assist + rebound + steal + block – turnover

DPOY = Defensive rebound + steal + block – foul



player	rebound	D_rebound	assist	steal	block	turnover	foul	point	win	MVP	DPO
扬尼斯-阿德托昆博	13.7	11.5	5.8	1.1	1.0	3.7	3.0	29.7	48	47.6	10.6
詹姆斯-哈登	6.3	5.3	7.4	1.7	0.9	4.4	3.4	34.4	39	46.3	4.5
卢卡-东契奇	9.3	8.0	8.7	1.1	0.2	4.2	2.5	28.7	34	43.7	6.8
勒布朗-詹姆斯	7.9	6.9	10.6	1.2	0.5	4.0	1.8	25.7	47	42.0	6.8
安东尼-戴维斯	9.4	7.1	3.1	1.5	2.4	2.4	2.5	26.7	43	40.7	8.5

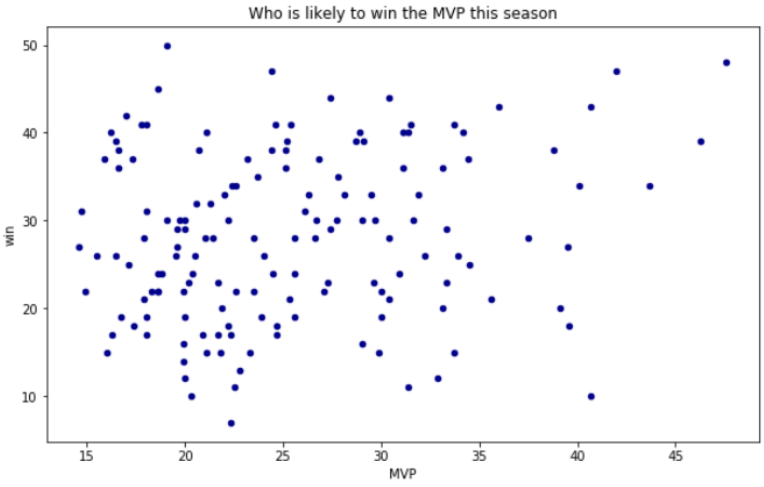
3. Methodology

3.1 machine learning method

In this project, I use k-means clustering to divide the player into several group, through which people can know which players are the potential competitors of the rewards. Besides, not only k-means clustering is the most suitable model for the data, but also can it help one to find which level does one's favorite player belong to.

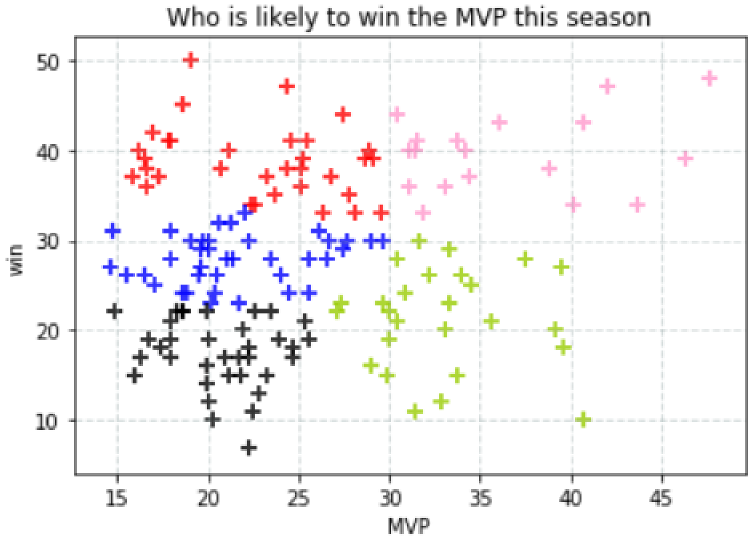
3.2 data analysis for MVP

At first, I set a sub data frame with the data of "MVP" and "win". Then, I draw the scatter plot of this data frame to see the approximate distribution of players' data.



Secondly, I set the clusters of the data and adjust the number of clusters in order to Make the boundaries of players clearer. In the Prediction of MVP, the number of clusters is 5.

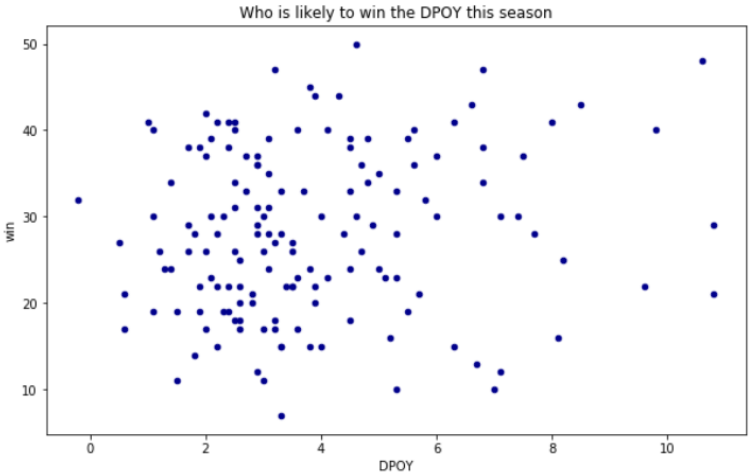
Thirdly, after dividing the players into 5 different levels, I draw the scatter point of k-means clustering:



From the picture above, we can see the 5 groups are: {MVP: [30~50], win:[30~50]}, {MVP: [25~40], win:[10~30]}, {MVP: [15~30], win:[30~50]}, {MVP: [15~30], win:[10~25]}, {MVP: [15~30], win:[25~35]}. It seems that the MVP consider both the value of MVP and the number of winning games.

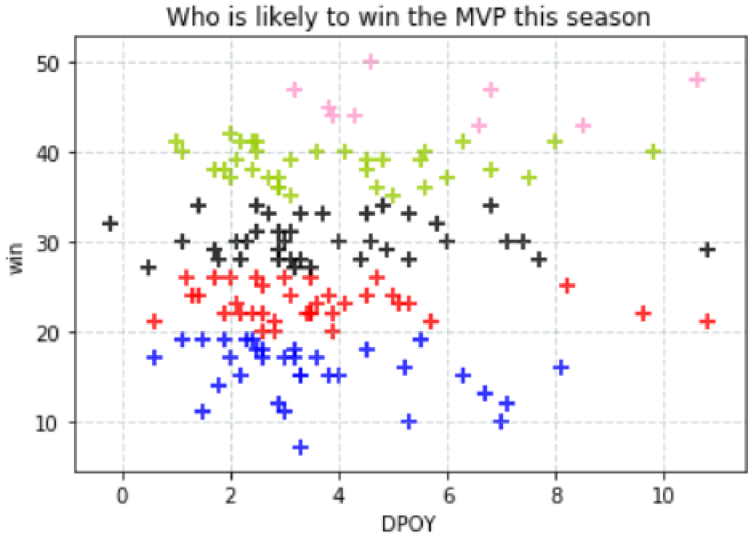
3.3 data analysis for DPOY

At first, I set a sub data frame with the data of "DPOY" and "win". Then, I draw the scatter plot of this data frame to see the approximate distribution of players' data.



Secondly, I set the clusters of the data and adjust the number of clusters in order to Make the boundaries of players clearer. In the Prediction of DPOY, the number of clusters is also 5.

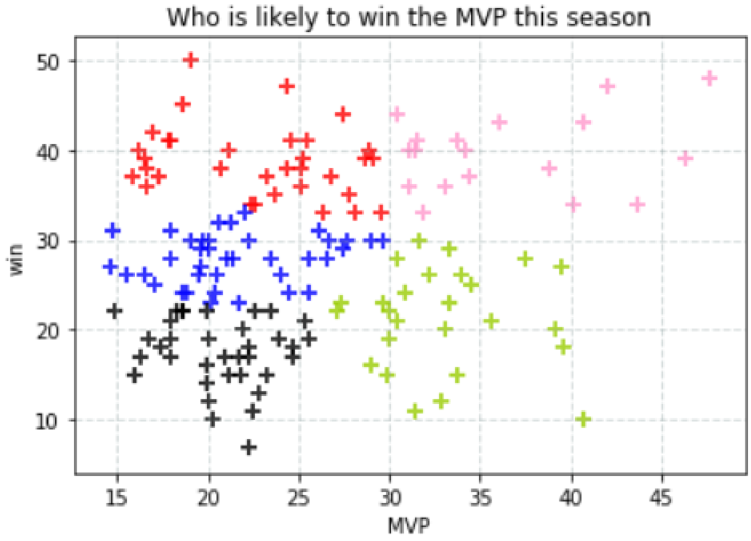
Thirdly, after dividing the players into 5 different levels, I draw the scatter point of k-means clustering:



From the picture above, we can see the 5 groups are: {DPOY: [0~12], win:[42~50]}, {DPOY: [0~12], win:[35~42]}, {DPOY: [0~12], win:[27~35]}, {DPOY: [0~12], win:[20~27]}, {DPOY: [0~12], win:[0~20]}. It seems that the number of winning games have a big influence on the DPOY.

4. Results

4.1 the result of MVP

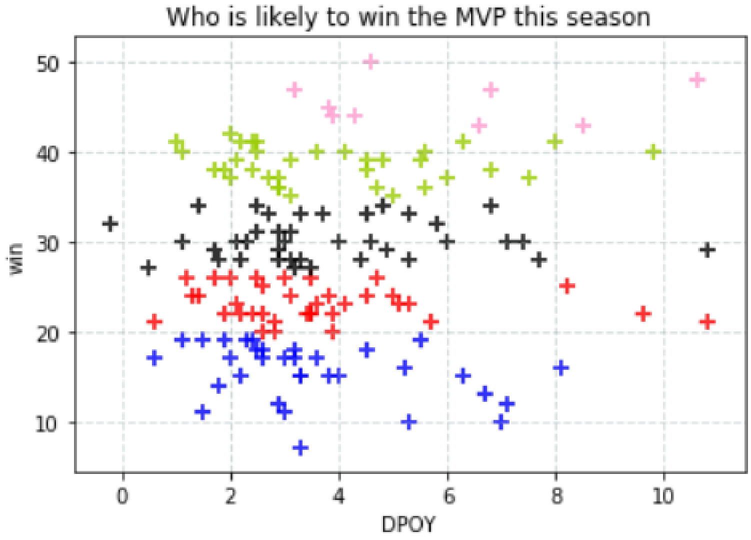


I divided 150 players into 5 groups:
MVP candidate: {MVP: [30~50], win: [30~50]},
Star in week team (some of them may be data brush playe): MVP: [25~40], win: [10~30]},
Main player in strong teams: {MVP: [15~30], win: [30~50]},
Main player in weak teams: {MVP: [15~30], win:[10~25]},
Main player in normal teams: {MVP: [15~30], win:[25~35]}.

In the MVP candidate, I think these four players may win the MVP:

Player	The Value of MVP	The Number of Winning
Giannis Antetokounmpo	47.6 1st	48 5th
Lebron James	41.9 4th	47 7th
Luka Doncic	43.8 3rd	34 65th
James Harden	46.3 2nd	39 32th

4.2 the result of DPOY



We can see that the 5 groups are divided mainly by the number of winning, which seems to remind us the importance of the number of wining while predicting the winner of DPOY. In addition, there is an interesting thing: Rudy Gobert belongs to the second group! It may be accident, but it can also be the truth. We can see who will win the DPOY in the end.

In the MVP candidate, I think these four players may win the DPOY:

Player	The Value of DPOY	The Number of Winning
Giannis Antetokounmpo	10.6 3rd	48 5th
Anthony Davis	8.5 6th	43 13th
Rudy Gobert	9.8 4th	40 25th
Bam Adebayo	8 9th	41 18th

5. Discussion

5.1 the data

In this project, I do not consider WS, PER, GMSC and that sort of things because I think the simple data such as point, rebound and assist may be more direct. However, high-level data also has its scientific aspect, so I wonder if the results will be different if high-level data is added for prediction.

5.2 the model of DPOY

In the actual prediction, I found that k-means clustering is mainly through the number of winning. Such clustering cannot be regarded as the weak one, because without victory, the defensive data will be unconvincing. However, is there a better clustering method or machine learning method to model the data?

6. Conclusion

This project is the first time for me to use the machine learning method to do the data analysis of NBA. I found that such method is indeed more intuitive than my previous regression analysis, and it is more beneficial to both the analyst and the audience. Completing such a project is also different from the lab and assignment in the course, because I am not completing the project according to the teacher's requirements, but completing the project according to my own ideas. When writing here, a sense of satisfaction emerged spontaneously, which also inspired me to continue to learn more knowledge about data science!