NBA MVP Prediction with Principal Component Analysis and Recurrent Neural Network

Lei Liu, Yuchen He

Abstract

This project is aimed to build a PCA-RNN model to predict the winner of The National Basketball Association's (NBA) the Most Valuable Player (MVP) for 2017-18 regular season. The present selection method is objective, while a more subjective and rational technique needs to be accomplished. Both NBA players individual statistics and MVP voting history from 2000-01 to 2016-17 season will be utilized. After scaling the dataset via principal component analysis (PCA), we will develop a recurrent neural network (RNN) model in order to estimate who will receive the highest MVP votes in the current season. The results show that PCA can significantly ascertain appropriate input parameters for our RNN model and the RNN model is a cogent predicator of NBA's MVP.

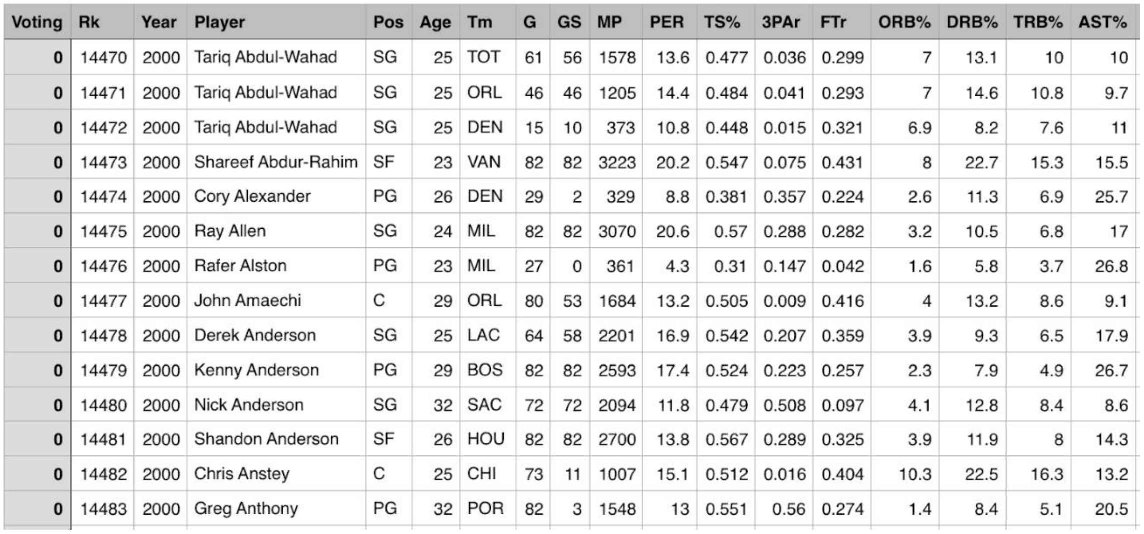
Keywords
Deep Learning, RNN, PCA, MVP Prediction, Data Mining

1. Introduction

The National Basketball Association (NBA) establishes various of awards to reward those who has the excellent performance and extraordinary contribution to the league [1]. Among all this, the Most Valuable Player (MVP) is considered the most remarkable. The NBA's MVP has been decided by a panel of sportswriters and broadcasters throughout the United States and Canada since the 1980-81 season. Each adjudicator, elected by the league, is allowed to put five players on their MVP ballot,
depending on the place, the players can receive points on a 10-7-5-3-1 scale [2]. The problem for this method is that the selection is subjective which means the most essential awards for NBA players is not quite fair. Therefore, we want to implement a PCA-RNN model on predicting the real MVP more objectively.

After counting and analyzing all MVP data of past seasons, we find that MVP winners have to be all-around productive and follow many common rules. For example, most of MVP players' team rank top 3 places in the regular season; Every MVP in the modern era (since 1979-80) put up a PER of at least 18.5 in the season prior [3]; Individual advanced data is key and 84 percent of MVPs averaged at least 0.20 WS/48 during the season, etc. This could contribute to the result that MVP award is related to many advanced data statistics.

In order to predict the MVP of 2017-2018 season, we collected 2000-2017 player data from Kaggle [4], 2017-2018 player data, voting rates and other advanced statistic data from Basketball Reference. After that, we jointed these data to adapt to our needs. In below Figure 1, our training and testing dataset has more than 10,000 pieces of data and 52 types of statistics.



Voting	Rk	Year	Player	Pos	Age	Tm	G	GS	MP	PER	TS%	3PAr	FTr	ORB%	DRB%	TRB%	AST%
0	14470	2000	Tariq Abdul-Wahad	SG	25	TOT	61	56	1578	13.6	0.477	0.036	0.299	7	13.1	10	10
0	14471	2000	Tariq Abdul-Wahad	SG	25	ORL	46	46	1205	14.4	0.484	0.041	0.293	7	14.6	10.8	9.7
0	14472	2000	Tariq Abdul-Wahad	SG	25	DEN	15	10	373	10.8	0.448	0.015	0.321	6.9	8.2	7.6	11
0	14473	2000	Shareef Abdur-Rahim	SF	23	VAN	82	82	3223	20.2	0.547	0.075	0.431	8	22.7	15.3	15.5
0	14474	2000	Cory Alexander	PG	26	DEN	29	2	329	8.8	0.381	0.357	0.224	2.6	11.3	6.9	25.7
0	14475	2000	Ray Allen	SG	24	MIL	82	82	3070	20.6	0.57	0.288	0.282	3.2	10.5	6.8	17
0	14476	2000	Rafer Alston	PG	23	MIL	27	0	361	4.3	0.31	0.147	0.042	1.6	5.8	3.7	26.8
0	14477	2000	John Amaechi	C	29	ORL	80	53	1684	13.2	0.505	0.009	0.416	4	13.2	8.6	9.1
0	14478	2000	Derek Anderson	SG	25	LAC	64	58	2201	16.9	0.542	0.207	0.359	3.9	9.3	6.5	17.9
0	14479	2000	Kenny Anderson	PG	29	BOS	82	82	2593	17.4	0.524	0.223	0.257	2.3	7.9	4.9	26.7
0	14480	2000	Nick Anderson	SG	32	SAC	72	72	2094	11.8	0.479	0.508	0.097	4.1	12.8	8.4	8.6
0	14481	2000	Shandon Anderson	SF	26	HOU	82	82	2700	13.8	0.567	0.289	0.325	3.9	11.9	8	14.3
0	14482	2000	Chris Anstey	C	25	CHI	73	11	1007	15.1	0.512	0.016	0.404	10.3	22.5	16.3	13.2
0	14483	2000	Greg Anthony	PG	32	POR	82	3	1548	13	0.551	0.56	0.274	1.4	8.4	5.1	20.5

Fig. 1. Training & Testing dataset

2. Code with Documentation

https://github.com/ll1195831146/Big-Data-Systems-Intelligence-Analytics/blob/master/Protfolio%20Blog.ipynb

3. Results

We create two models which are all progressed by the same RNN, the only difference is that PCA is implemented on the second model, which select 15 components that have more significant impact on the voting of MVP. The comparison of implementing PCA on our dataset is shown in both Figure 2 and Figure 3. The horizontal axis represents the players, while the vertical axis represents the final votes they will receive. The blue dots represent the actual voting, which can be considered as the test set. The red dots stand for the predictive voting after training by our RNN model [7]. It is quite obvious that the coincidence rate on the second plot is higher than the first.

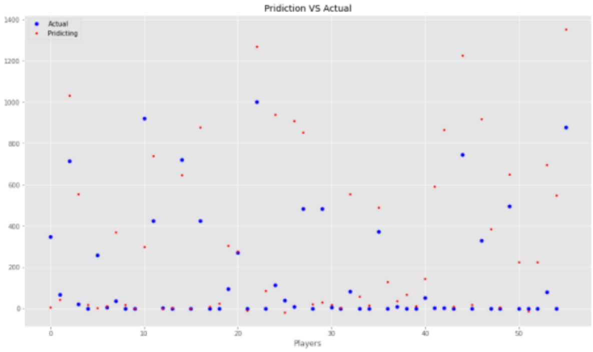


Fig. 2. Result not using PCA

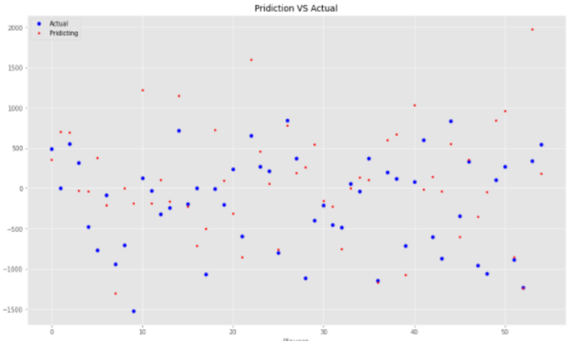
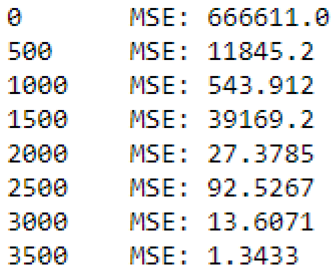


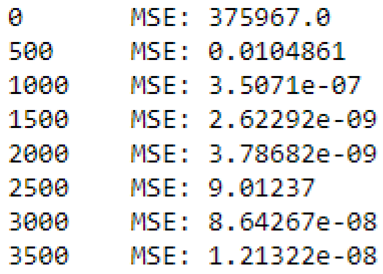
Fig. 3. Result using PCA

To evaluate prediction errors for our model, we also implement the mean squared error (MSE) to testify the difference between predicted and actual value. The lower the MSE values are, the less error the prediction has [5]. Therefore, MSE can be another evidence of proving that implementing PCA has effectively increased the accuracy. As shown in Figure 4, the MSE of the model which does not use PCA is way lower than the one used PCA, as shown in Figure 5, on every epoch.



```
0       MSE: 666611.0
500     MSE: 11845.2
1000    MSE: 543.912
1500    MSE: 39169.2
2000    MSE: 27.3785
2500    MSE: 92.5267
3000    MSE: 13.6071
3500    MSE: 1.3433
```

Fig. 4. Mean-squared-error of every 500 epochs



```
0       MSE: 375967.0
500     MSE: 0.0104861
1000    MSE: 3.5071e-07
1500    MSE: 2.62292e-09
2000    MSE: 3.78682e-09
2500    MSE: 9.01237
3000    MSE: 8.64267e-08
3500    MSE: 1.21322e-08
```

Fig. 5. Mean-squared-error of every 500 epochs after using PCA

We decide to repeat our training cycle for 4000 times and display the changing process every 500 epochs via MSE. As shown in Figure 5, the MSE value has a rapidly decrease from the beginning, and eventually remain as low numbers, which means a high precision comparing our predicted value with the test value. Thus, our final results can be trustworthy.

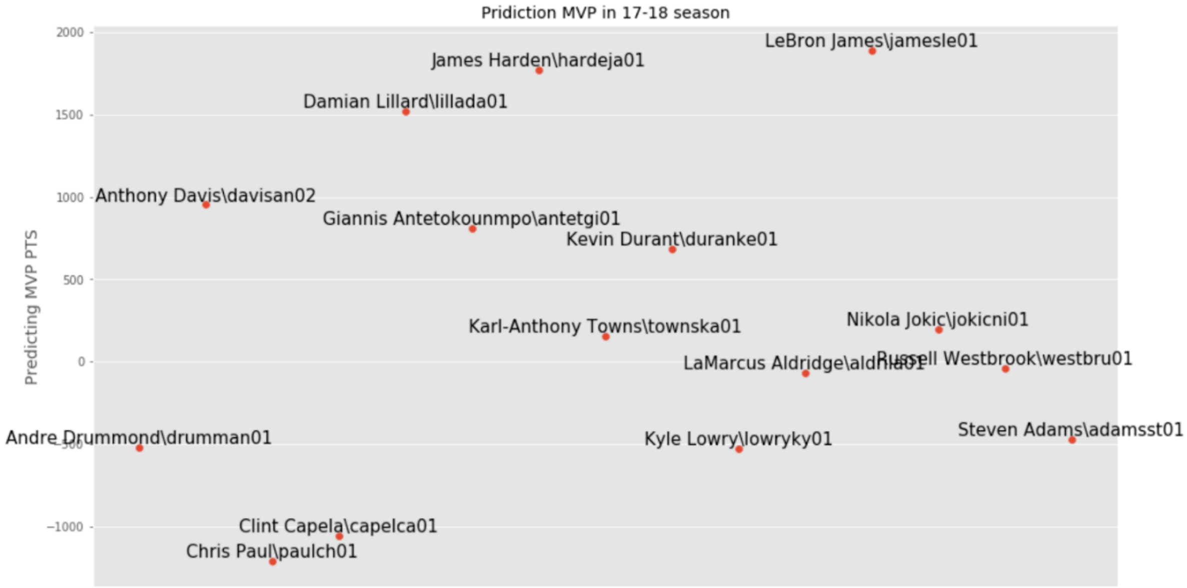


Fig. 6. Final results of NBA 2017-18 regular season MVP winner

Our project ultimately forms the possible votes that every candidate can obtain for 2017-18 NBA regular season. The predicting MVP points that players receive shows the possibility he can be chosen as the MVP of this season (because our dataset is processed with PCA, the scaling is quite different and points can be negative). The Figure 6 represents the voting result of some players who are more likely to receive high numbers of votes, and the vertical shows the vote they may receive. As Figure 6 shows, the most likely winner will be LeBron James with the unique ID of jamesle01, followed by James Harden (hardeja01). This result makes sense since LeBron James is actually one of the best player in the league and considered as a high-rank candidate for the NBA MVP award. Besides, Harden is also one of the MVP contenders and led the team rockets into a historic season.

4. Discussion

In closing, the main idea of the project is to provide a more objective, progressive and precise method for those who are interested in the winner of NBA MVP. Because of the subjectivity of NBA MVP voting, a machine learning model seems to be more suitable for prediction. Unlike the original statistical model, which is the formalization of relationships between variables in the form of mathematical equations, a machine learning model can learn from data without relying on rules-based programming [6].

Indeed, the player statistics have huge impact on the MVP election, and we comprehensively utilized all these data and acquired satisfying results. However, some other elements, which are hardly expressed by numbers, may still affect this award, such as a player's commercial value. The operation of NBA league needs large sum of funds, which may cause a covert transaction between the league and the players' sponsors. In addition, the influence of the player may also be considered, Bill Russell, Michael Jordan and Lebron James are examples. We need to consider more about how we can transfer these elements into numerical value which can be processed by our model.

We hope, one day, our modeling can be used by the league of NBA to determine the MVP winner as well as the other awards rather than continuing to use personal votes for selection. "Everyone has their own Hamlet", the voters may choose somebody who is not qualified just because of their preference, which is not fair for those who work hard and bring excellent games to the audiences.

References

[1] "2018 NBA Awards." *NBA.com*, 18 Apr. 2018, www.nba.com/NBAawards#/. Accessed 21 Apr. 2018.

[2] Hertzog, Aaron. "The NBA MVP Rubric." 13 Apr. 2017, medium.com/holding-court/the-nba-mvp-rubric-c58f050a8e9c. Accessed 21 Apr. 2018.

[3] "500 Players, 1 Winner: Predicting the 2017-18 MVP." *ESPN.com*, www.espn.com/espn/feature/story/_/id/20982956/predicting-2017-18-nba-mvp. Accessed 21 Apr. 2018.

[4] "NBA Players Stats Since 1950." *Kaggle*, www.kaggle.com/drgilermo/nba-players-stats. Accessed 21 Apr. 2018.

[5] Chai, T., and R. R. Draxler. "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature." *Geoscientific Model Development*, vol. 7, no. 3, 2014, pp. 1247-1250.

[6] "Machine Learning Vs Statistical Modeling: The Real Difference." 15 Sept. 2016, www.infogix.com/blog/machine-learning-vs-statistical-modeling-the-real-difference/. Accessed 21 Apr. 2018.

[7] Heinz, Sebastian. "A Simple Deep Learning Model for Stock Price Prediction Using TensorFlow." 9 Nov. 2017, medium.com/mlreview/a-simple-deep-learning-model-for-stock-price-prediction-using-tensorflow-30505541d877. Accessed 21 Apr. 2018.