Application of Machine Learning on NBA Data Sets

Jingru Wang [1, *], Qishi Fan [2]

[1] University of Maryland, College Park, Maryland, 20740, United State
[2] University of California, Irvine, Surrey, V3W2M8, Canada

*Corresponding author e-mail: jwang144@terpmail.umd.edu

Abstract. Machine learning is known as the most popular methodology to do prediction on large data set while NBA's data sets consists of plentiful statistics. Since predictions of various events are important, our research would investigate whether machine learning algorithms are efficient in doing prediction on certain NBA data sets and tasks. We are focus on mainly three supervised tasks, namely: All-Star Prediction, Playoff Prediction and Hot Streak Fallacy. For Playoff Prediction, we predict the team performance by doing machine learning on two data sets consisting of distinct well-selected features and compare the result to show which data set are more suitable for the machine learning to work. The results show that advanced statistics outperform the elementary ones. For Hot Streak Fallacy, we build the model based on multiple-linear regression to address the question: is hot streak a fallacy? It turns out that there is a lack of evidence to support 'Hot Streak Phenomenon'. For the NBA Trend, we try to view how the games involve for the past decade, and analyze the correlation of playoff tickets and other data.

Keywords: Machine learning, prediction, NBA playoffs.

1. Introduction

The main body of the report contains three parts: II.All-Star Prediction, III.Playoff Prediction, IV.Hot Streak Fallacy and V.NBA Trend Analysis.

1) In the All-Star Prediction section, we mainly employ classification to predict the results of whether a player can go into all-star lineup. We adopt K-Nearest Neighbor Classification, Logistic regression, and decision tree respectively test the accuracy of prediction on a well-cleaned players' status data set, in which we replaced string type data into float data and set null values to 0.
2) Playoff prediction contributes as an independent part. We employ two types of data set here: description data set and pre-processing data upper elementary stats. Three machine learning algorithms are put into usage: Decision tree, random forest, and gradient boosting. With the comparison of results, basically prediction accuracy, between two data sets, we get our conclusion that in multiple measurements advanced data is more advance in machine learning model.
3) In this section we examine whether hot streak phe- nomenon is a fallacy or not. With 27 selected players in 2018-19 season, we implement the linear regression model to prove to disprove the

statement, and provide a simple model as comparison to the result. In either way, due to the low test accuracy, we attribute that hot streak is actually a fallacy.

4) In this section we use PCA to see how the games involve and use correlation analysis.

2. All-Star Prediction

2.1. Introduction of the data set

The data set is retrieve from the NBA reference website [1] that records all the player's status in season 2019. The data set is well cleaned: all the string data type such as name, Team and position are converted into float data type, the outliers replaced and the null values are set into 0. The classifier is whether the player can go into the all star team. The trail of the K-Nearest Neighbours (KNN) Cluster is a failure, the maximum successfully rate is only 45 percent and cluster is separate into too many groups that tangling each other. (Fig 1) Therefore, the prediction is based on the Classification.

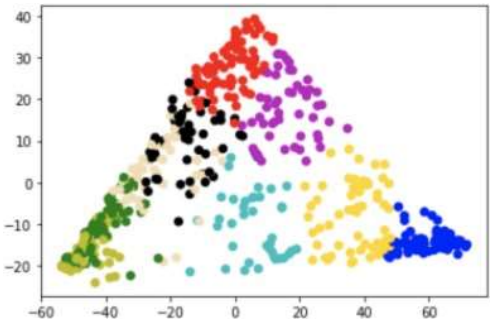

Fig. 1 Knn Cluster

2.2. K-Nearest Neighbor Classification

KNN can be used for both classification and regression predic- tive problems. KNN falls in the supervised learning family of algorithms. Informally, this means that we are given a labelled dataset consiting of training observations (x,y) and would like to capture the relationship between x and y. More formally, our goal is to learn a function h:X-Y so that given an unseen observation x, h(x) can confidently predict the corresponding output y. First applying KKN classifier on dataset, and get k=3 it has the largest accuracy score. Then apply that model to test dataset finding that accuracy of our model is equal 0.9653. Then using cross-validation to improve it, and find that the optimal number of neighbors is 27. Accuracy of our model is equal 0.9583.

2.3. Logistic Regerssion

Logistic Regerssion is also a classification model. The model is trained on the dataset and get the result of 0.979 accuracy when testing.

2.4. Decision Tree and Prediction

The accuracy of the decision tree is around 95 to 94 percent which is highly accurately. The classifier has separate into two groups: all star player and non all star player. (Fig 2) As the data set is from season 2019, instead of using 2019 data to predict, we select a few players that with highly potential in 2020 as test data set: Zion Williamson, Zach Lavine and Bradly Beal. The result shows that they are all qualified to became an all star player.

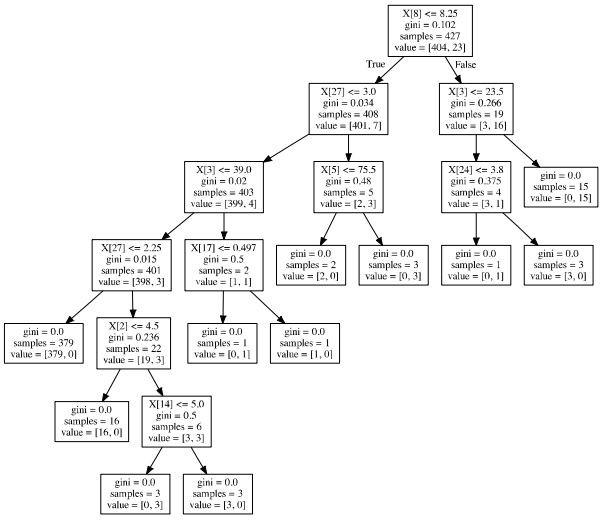

Fig.2 Decision Tree

	Player	Pos	Age	Tm	G	GS	MP	FG	FGA	FG%	...	Po
0	Zion Williamson	PF	19	NOP	24	24	27.8	8.8	15.0	0.583	...	
1	Zach LaVine	SG	24	CHI	60	60	34.8	9.0	20.0	0.450	...	
2	Bradley Beal	SG	26	WAS	57	57	36.0	10.4	22.9	0.455	...	

Fig. 3 Player Prediction

2.5. Player age distribution

Most of the player's age is around 21 to 28, the dash line represent the average age of the all player. As you can see the number of player has largely decrease after age 26, which means age is a crucial factor that determines whether the players can become all-star or not. The interpretation here is that maybe for younger players, their real-game performance is much more 'explosive' than the veterans. In short, they prefer to dunk than just to shoot. Hence voters are inclined to vote for them in the All-Star vote.

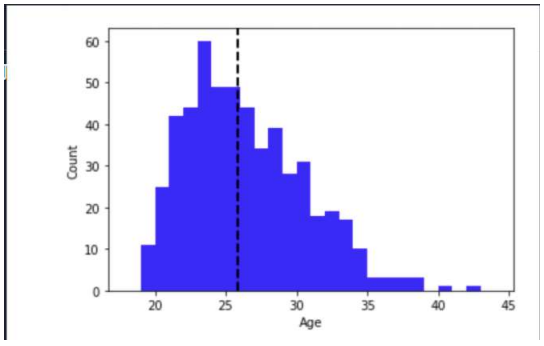

Fig. 4 Age distribution

3. Playoff Prediction

3.1. Background and Prior works

Numerous algorithms have been carried out either by re- searchers or basketball data analysts to predict whether an NBA team could enter the playoff or not based on some statistics of their team performance. But there is a lack of research regarding the comparison between the suitability of data sets when a

supervised learning algorithm is implemented to predict. Hence we would compare the efficiency of machine learning algorithms on two data sets consisting of completely different sets of well-selected features. At last, we compare the results and select the more robust data set.

3.2. Description and Pre-Processing of the Data Sets

Two data sets are collected from basketball-reference.com [1] and used separately for the task.
The first data set only contained elementary statistics of each team in each regular season from season 2009 to 2019. In short, 'elementary' is in the sense that those statistics is collected simply by counting the players' action in each game without doing further analysis.

In contrast, the second set of data contains a set of advanced metrics which is a result of the pre-processing of elementary statistics. Furthermore, the indicator of the performance of a team is changed by number of games wined instead of playoff's qualification. We then use the built-in MinMaxScaler function in sklearn to normalize the features and split the data set into 70% train data set and 30% test data set.

3.3. Machine Learning Algorithms

Based on the decision tree model, three machine learning algorithms are carried out: decision tree, random forest and gradient boosting.

3.4. Experiments and Results

The accuracy of the result for data set is calculated simply by the proportion of right predictions.

Decision Tree:

An experiment of investigating the accuracy against maximum depth allowed for the decision trees is carried out, the result is shown in Fig 5. The optimal depth turns out to be 7 which leads to an accuracy of 70%

A features importance's plot is also included. It could be seen that the most important feature is Field Goal Percentage as expected. The efficiency of gaining points by shooting the ball should be the most direct evaluation of a team after all. Moreover, BLK (number of blocks) is one of the least important features. We are not at all surprised as blocking does not quite contribute a lot to the result of a match. It is merely a fancy trick to the audience most of the time.

3.5. Random Forest:

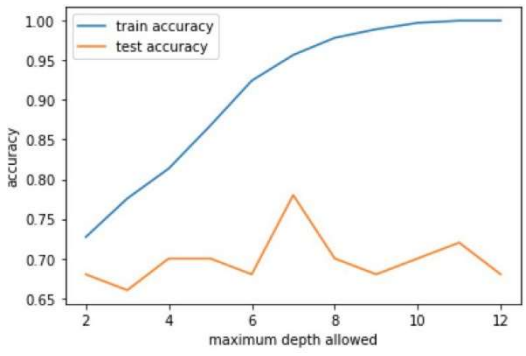

Fig. 5 Decision Tree

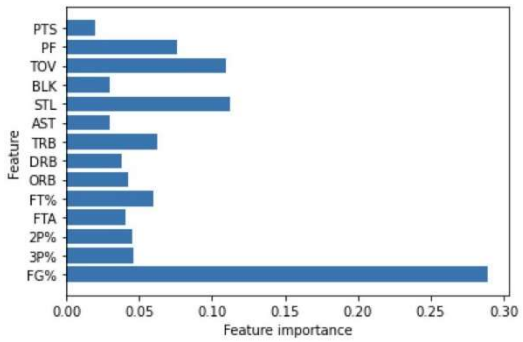

Fig. 6 Feature importance

An experiment of investigating the accuracy against the num- ber of independent decision trees is carried out. The optimal number turns out to be 8 which leads to an accuracy of 0.85.

3.6. Conclusion

Clearly from TABLEI, Data Set2, which is the data sets that contains advanced data only, results in a much higher accuracy than Data Set1 where only elementary statistics is used. So we are confident to conclude that an advanced data set is indeed 'advanced' when a machine learning model is implemented. This result is within expectation as the advanced metrics are designed to be the group of more superior indicators about a. team's performance. From the feature importance's plot, we can see that some statistics contribute almost nothing to the overall performance of a team. So it is good practiced to get rid of those features before doing the machine learning task. The optimal choices of parameters in decision tree and random forest are not so significant: From the plots, the accuracy does not alter a lot with respect to change in parameters, at least much lesser than the change brought by a replacement of data set. Moreover, it is highly dependent on the random train-test split procedure.

In the future, we could certainly apply more complex algo- rithms on the sets instead of those based merely on decision trees. Moreover, since the field of NBA data analysis becomes more mature over time, new metrics would be invented which means we can re-collect new data sets for our experiments.

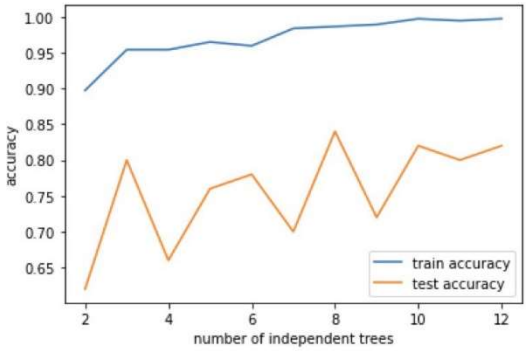

Fig. 7. random forest

Table 1. Test accuracy of the experiments

Data set \ Model	Decision Tree	Random Forest	Boosting
Data Set1	0.75	0.85	0.82
Data Set2	0.91	0.88	0.92

4. Hot Streak Fallacy

4.1. Background and Prior Works

The" hot streak fallacy" (also known as the" hot hand phe- nomenon") was considered a cognitive social bias that a person who experiences a successful outcome has a greater chance of success in further attempts. In the context of basketball, people who believe in" Hot Streak" would agree that those who just made a successful shot should attempt shooting again since they got a better chance of making the shot than others. If" Hot Streak" is proved not to be fallacy, there would be a huge change to a team's both defensive and offensive tactics. Many researches have been carried out to prove or disprove the" hot hand phenomenon". Some have worked based on the three-point contest held on all-star weekend (Miller, Joshua B.and Sanjurjo, Adam 2019) [2] while other have invested the re- lationship between field goal percentage and previous games' performance by building conditional probability model (Chang S-C 2018) [3]. However, few researches focus on finding the multi-linear relation between consecutive games' three-point shooting percentage (Caspo, Peter and Raab, Markus 2014) [4]. Therefore, we would investigate the hot streak fallacy by implementing linear regression model on players' consecutive games' shooting percentage.

4.2. Description and Pre-Processing of the Data Set

27 players are selected from the rank of number of three- point shoots attempted in 2018-19 NBA regular season. We than divide their three-point shooting percentage for each game into group of 3 consecutive games and view the third game as the dependent variable depending on the first 2 games (Miller, Joshua and Sanjurjo 2019) [5]. As usual, we divide the data set into 70% of train data and 30% of test data. For comparison, we also adapt a simple model using only 1 game's statistics to predict the next game (Chang S-C 2018) [6].

4.3. Machine Learning Algorithms

We implement the linear regression model for 2 independents variables:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where X1 and X2stand for the first two games' percentage while we use Y to predict the third game' percentage. For the simple model, we compare the accuracy of prediction using the mean estimator of the player's percentage over the whole season and the percentage of his previous game.

4.4. Experiments and Results

The baseline we set up for the experiment is that the prediction should be at least better than the mean estimator. Hence we define one prediction to be 'successful' if

$$\hat{Y} - Y \leq \bar{Y} - Y$$

Where Yˆ, $\bar{Y}$, Y stand for the prediction value, the mean estimator and the actual shooting percentage respectively. The metrics of accuracy is simply the proportion of the 'successful' predictions over the total number of predictions. The parameter of linear regression is shown in TABLE II: If the hot streak

Table 2. Parameters of the linear model

β_0	0.709
β_1	-0.0790
β_2	-0.522

is not a fallacy, we should at least expect that both β_1 and β_2 being positive. However, since neither of then is, the hypothesis of hot streak is completely rendered in jeopardy. The summary of accuracy is

shown in TABLE III: The fact that both of the models resulting in accuracy below 40% leads us to reject the hot streak model.

Table 3. Accuracy

Linear Model	0.36
Simple Model	0.37

4.5. E. Conclusion

With the low accuracy of both models at hands, we have to conclude that there is no positive correlation between consecutive games — 'Hot Streak Phenomenon' is indeed a fallacy! This is in consistent with the previous conclusion reached by statisticians who did the similar experiments.

5. NBA Trend Analysis

data. We can clearly see the trend for the past decade, which does not show clear preference for offense or defense. All the data lies in a relatively linear line and as the time goes by, all the data becomes larger.

5.1. Background

Since 2015, when we witnessed the success of Golden States Warriors, the league is said to play smaller and smaller, which means many teams are trying to let shorter players to play higher positions. This trend is often referred as" small ball" that is actually derived from three-point revolution, where players make more and more three points than long-range two points for the win, regardless of distance (Romanowich, P., Bourret, J., & Vollmer, T. R., 2007) [7]. Therefore, we would like to analyse the trend during the past decade, and find out which data really influences the possibility to get into playoffs (Santos, 2020) [8].

5.2. Description and Pre-Processing of the Data Sets

We use the first dataset as the III part, which contains some elementary data in the regular season. We split this dataset by year, and we can give out a result involving with the time. Besides we add a column" LOC", which shows which con- ference the team is in, and 1 denotes the Eastern Conference, and 0 denotes the Western Conference (Soliman, 2017) [9].

5.3. Machine Learning Algorithms

We now use PCA to analyse the data trend for the past ten years, and we split all the data into the offensive data and defensive data. The former one contains FGA FG%, 3PA, 3P%, 2PA, 2P%, FTA, FT%, PTS and AST, while the latter one contains BLK, STL, ORB, DRB and TOV. Then we use PCA to visualize them both onto the 1-dimensional data, so we can see them in a 2-dimensional surface. Moreover, we wonder which data is really related to the ticket of playoffs, so we can do correlation analysis for each data and we use interpolation formula to make the curve smooth.

5.4. Experiments and Results

PCA:

For the first part, we use PCA to do dimension reduction, and we use different colors to describe the time for each data, the lighter one means that it is early data, and the darker one means recent data. Hence we get Figure 8, where the x-axis refers to the offensive data, and the y-axis refers to the defensive data. The result of PCA for each dimension has positive correlation for most of the data, and negative correlation only for several data. Therefore, we can just think of that as positively related to offensive data and defensive

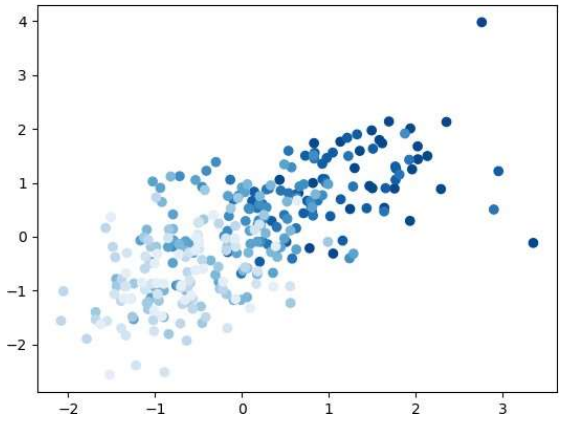

Fig. 8 PCA

This result tells us the data does become greater than the past for each team, so the game is really involving. This result somehow shows that NBA teams are now performing better, since they may have better data. This is easy to understand to some extent, for players now have better food and health care. The level of competence surly becomes higher. However, this is also highly related to the whole environment. Fans are pleased to see more rounds during a game, and most of them are just watching the highlights, rather than the whole game. Besides, with the data analysis tools, many more teams shows great care for data of each player, which makes some players focus on their data, and are trying to find ways to improve them with all means. That is to say, the data cannot show all the aspects of the real situation.

Correlation Analysis:

For the second part, we can solve the second question that is “which data is really related to the ticket of playoffs”. We can answer this problem by showing the correlation for each data, and the result can be interesting.

We can see that the playoff ticket has little correlation to the Conference of each team, so it means that the argument that in some conference it is easier to get into playoffs cannot be supported by the data. Besides, it is astonishing that 2PA has negative correlation with 2P%, while 3PA has negative correlation with 3P%. No wonder most teams now are shooting more and more 3P. Indeed, shooting more 3P will makes a higher PTS, and more 2P can only lead to lower PTS. However, the games is about having more points than the opposite. In this sense, we can know the main reason for the 3P revolution. For AST, we can know from the result that it has positive correlation for almost all the offensive data, except 2PA and FTA, which means the game indeed becomes more fluent if the team passes the ball more.

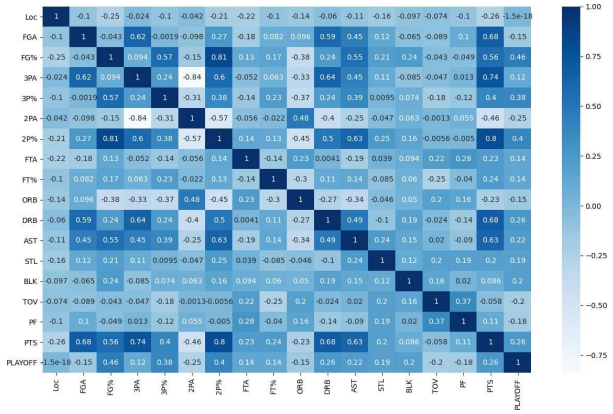

Fig. 9 Correlation Analysis

Then we can show all the correlation between Playoff and other data for the past decade.

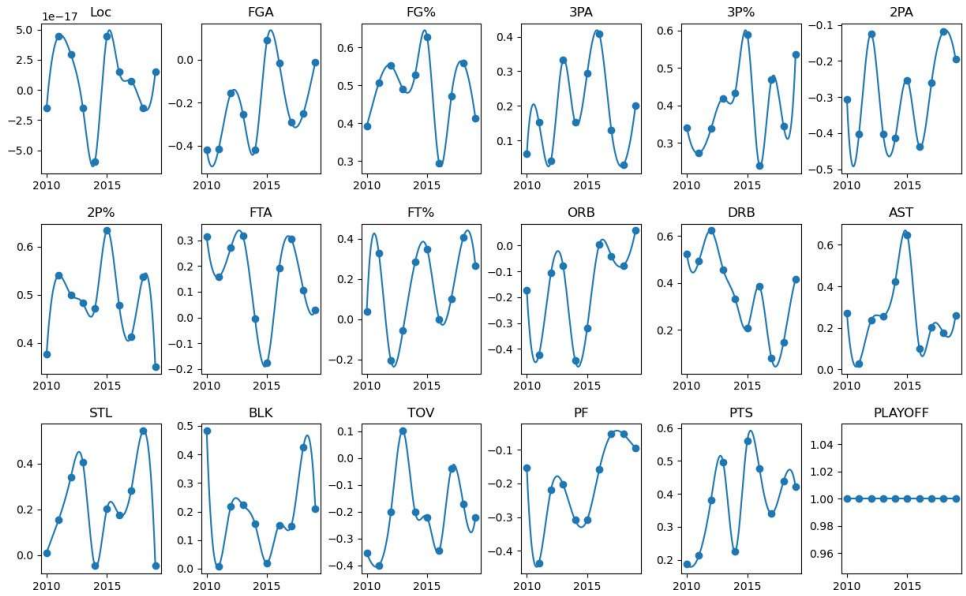

Fig. 10 Correlation between Playoff and others

For this result, we can see more clearly of the three point revolution. Indeed, shooting more 3P can makes it more likely to go into playoffs, while more 2P makes it less likely. Besides, there is no evidence shows that teams can go into playoffs by more ORB or FGA. Hence, in some sense, it is more important to protect DRB rather than grab some ORB. Moreover, for PF and TOV, the correlation is always negative, which makes sense. Controlling PF and TOV is very important to go into playoffs.

6. Conclusion

In this section, we try to analyse the NBA trend, and we do see the evolution of NBA games. However, for the correlation section, we cannot see clear changes for the past decade. The relation is changing for different years, which does not show linear relation with the time. That is to say, the games do have greater data, but in order to win the games to go into playoffs, NBA teams still need to do all the old stuff: controlling TOV, etc (Lantis, 2019) [10].

References

[1] Sports Reference LLC. Basketball-Reference.com - Basketball Statistics and History. https://www.basketball-reference.com/. September 15, 2020

[2] Michael B.E, Simcha A., Markus R. (2006) Twenty years of “hot hand” research: Review and critique, Psychology of Sport and Exercise, Volume 7, Issue 6, pp. 525-553, https://doi.org/10.1016/j.psychsport.2006.03.001.

[3] Daks, A., Desai, N. and Goldberg, L.R. Do the Golden State Warriors Have Hot Hands?. Math Intelligencer 40, 1–5 (2018). https://doi-org /10.1007/s00283-018-9825-3

[4] Caspo, P. and Raab, M. (2014) “Hand down, man down.” analysis of defensive adjustments in response to the hot hand in basketball using novel defense metrics. PLoS One, 9(12). http://dx.doi.org /10.1371/journal.pone.0114184

[5] Miller, Joshua B. and Sanjurjo, Adam, Is It a Fallacy to Believe in the Hot Hand in the NBA Three-Point Contest? (September 19, 2019). IGIER Working Paper No. 548, Available at SSRN: https://ssrn.com/abstract=2611987 or http://dx.doi.org/10.2139/ssrn.2611987

[6] Chang S-C (2018) Capability and opportunity in hot shooting perfor- mance: Evidence from top-scoring NBA leaders. PLoS ONE 13(2): e0179154.

https://doi.org/10.1371/journal.pone.0179154
[7] Romanowich, P., Bourret, J., & Vollmer, T. R. (2007). Further analysis of the matching law to describe two- and three-point shot allocation by professional basketball players. Journal of applied behavior analysis, 40(2), 311–315. https://doi.org/10.1901/jaba.2007.119-05
[8] Santos, J., Mendez-Domínguez, C., Nunes, C., Gómez, M.A., Travassos, B. (2020) Examining the key performance indicators of all-star players and winning teams in elite futsal, International Journal of Performance Analysis in Sport, 20:1, 78-89.
[9] Soliman, G., El-Nabaway, A., Misbah, A., Eldawlatly, S. (2017) Predicting All Star Player in the National Basketball Association using Random Forest. Intelligent Systems Conference 2017,706-713.
[10] Lantis, R. M., & NBER Working Papers. (2019). Hot shots: An analysis of the 'hot hand' in NBA field goal and free throw shooting. Place of publication not identified: National Bureau of Economic Research