Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230413

Investigating the factors that influence NBA draft rankings

Yanbing Ren[1,4], Bailin Li[2,5], Yangsheng Li[3, 6]

[1]School of Economics, Jinan University, Guangzhou, 510632, China,
[2]Nottingham University Business School China, University of Nottingham Ningbo China, Ningbo, 315100, China,
[3]College of Agriculture and Biology, Zhongkai College of Agricultural Engineering, Guangzhou, 510550, China

[4]2863099142@163.com
[5]biybl8@nottingham.edu.cn
[6]1901433156@qq.com

Abstract. The NBA Draft is an annual event for the NBA to select new players, which is significant for both players and NBA teams. In this work, the main objective is to explore whether player draft rankings could reflect a player's career development in the NBA. With this aim, we built three multiple linear regression models based on the 10-year draft rankings collected from 2009-2018 and the NBA career data and college data for these players. The results of these analyses indicate that these variables of NBA players: G, PTS and BLK have the strongest associations with their NBA draft ranks in NBA games and that these variables: TG, FGP, P3PP, PPTS and PSTL in college statistics have strong associations with the future NBA draft rank. However, there are some outliers in the prediction results, which indicates that some abilities valued by the team cannot be reflected by data values.

Keywords: NBA draft rank, Factor, Multiple linear regression, Predict.

1. Introduction

National Basketball Association (NBA) was originated from Basketball Association of America (BBA) and National Basketball league (NBL) [1]. After the merge of BBA and NBL, NBA stepped into a developing period. With the development of NBA, the draft system is also reforming and changing [2]. The draft system of the NBA is an essential measure for the sustainable development of the NBA because it allows the weaker teams to have access to the future stars. This reverse order of the draft was determined by the NBA's predecessor, the BAA, which followed the lead of the NFL [3]. Each spring, a random ping pong ball drawing will determine the fate of the fourteen teams that will not make the playoffs that season [4]. The team with the last record has the highest probability of winning the lottery, with the three ping-pong balls representing the first, second, and third picks, and the subsequent picks directly following the reverse order of record [4]. After the 14th pick, the teams advancing to the playoffs that season are also listed in reverse order of record [4]. However, there is a relative lack of research on the NBA draft, with a number of questions that have not been clearly answered. Based on this, we collected ten-year draft rankings from 2009-2018 and these players' career data and college data to explore whether the NBA draft rankings reflect players' future career development. The draft rankings are also predicted by their college data to examine which indicators have a significant relationship with

DOI: 10.54254/2755-2721/6/20230413

draft rankings. The rest of this essay is divided into three sections. In the next section, we describe the data we collected. Following that, two backward linear models built on NBA career data and rookie season data, respectively, are presented and the models are further analyzed. Subsequently, in the same section, two forward linear models built on college career data and data from the last college season before entering the NBA are presented and the better one is selected for prediction ranking through analysis. The essay ends with a summary of the findings and a discussion of the shortcomings of our models and possible future research directions.

2. Materials

The main source of data used for this essay is the five different types of tables on the Basketball-Reference website https://www.basketball-reference.com in excel format. It includes NBA Draft, Per Game, Advanced, Shooting, College Stats [5]. The following data used is extracted from these five tables. The abbreviation and meanings of indicators are shown in table 1.

Table 1. Abbreviation and Specific explanation.

Column name	Specific explanation
Rank	Draft rank.
G	Total games.
ASTP	Assist percentage. An estimate of the percentage of teammate field goals a player.
PMP	Minutes played per game.
ORBP	Offensive rebound percentage. An estimate of the percentage of available offensive rebounds a player grabbed while they were on the floor.
BLKP	Block percentage. An estimate of the percentage of opponent two-point field goal attempts blocked by the player while they were on the floor.
PPTS	Points per game.
USGP	Usage percentage. An estimate of the percentage of team plays used by a player while they were on the floor.
DRBP	Defensive rebound percentage. An estimate of the percentage of available defensive rebounds a player grabbed while they were on the floor.
TOVP	Turnover percentage. An estimate of turnovers committed per 100 plays.
pros	The proportion of total games started to total games.
P2PP	2-point field goal percentage.
PDRB	Defensive rebounds per game.
OWS	Offensive win shares. An estimate of the number of wins contributed by a player due to offensive.
DWS	Defensive win shares. An estimate of the number of wins contributed by a player due to defense.
D0.3FGAP.	Percentage of field goal attempts that are 0-3 feet from the basket.
D16.3PFGAP	Percentage of field goal attempts that are 2-pt shots and 16+ feet from the basket.
GS	Total games started.
STLP	Steal percentage. An estimate of the percentage of opponent possessions that end with a steal by the player while they were on the floor.
P3PP	3-point field goal percentage.
TSP	True shooting percentage. A measure of shooting efficiency that takes into account 2-point field goals, 3-point field goals, and free throws.

DOI: 10.54254/2755-2721/6/20230413

Table 1. (continued).

PBLK	Blocks per game.
PTOV	Turnovers per game.
FG3P	field goal percentage on 3-pt field goal attempts.
FGP	Field goal percentage.
FTP	Free throw percentage per game.
PAST	Assists Per Game.
PORB	Offensive Rebounds Per game.
PSTL	Per game Steals.
PPF	Per game Personal Fouls.
TG	Total Games.

3. Models

The goals of the essay are to find whether some factors in NBA games' statistics have some associations with NBA draft rank and whether NBA draft rank can represent each NBA players' ability. Thus, try to use two types of multiple linear regression model: Backward model and forward model. Each model is fitted by different sets of data.

3.1. Backward model

Use two sets of data to fit two multiple linear regression models and to investigate the associations between explanatory variables and dependent variable. LMC is the first model using career data of each NBA player who is on the NBA draft from 2009 to 2018, which may showcase during a player's career, which factor may have an association with NBA draft rank. LMRS is the second model using rookie season data of each NBA player who is on the NBA draft from 2009 to 2018, which may show in their rookie seasons, what factor may have an association with NBA draft rank.

3.1.1. LMC

Use the career data of each NBA players who are on the draft from 2009 to 2018 to fit the multiple linear regression model LMC. Choose ten important variables essential for evaluating a player's ability as explanatory variables: G, ASTP, PMP, ORBP, BLKP, PPTS, USGP, DRBP, TOVP, pros. Choose NBA draft rank as dependent variables. Use stepwise regression and find some new variables: P2PP, PDRB, OWS, DWS, D0.3FGAP, D16.3PFGAP. Add these variables to original variables, filter out those variables whose p-values are larger than 0.05 and acquire the final linear model LMC.

$$\widehat{Rank} = \hat{\alpha} + \widehat{\beta_1} * G + \widehat{\beta_2} * PPTS + \widehat{\beta_3} * PDRB + \widehat{\beta_4} * OWS + \widehat{\beta_5} * D0.3FGAP + \widehat{\beta_6} * D16.3PFGAP + \hat{\varepsilon} \quad (1)$$

Estimated coefficients of LMC are shown in table 2.

Table 2. Estimated Coefficients of LMC.

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	51.078657	1.181233	43.242	< 2e-16 ***
G	-0.021597	0.003645	-5.925	5.29e-09 ***
PPTS	-1.345871	0.179351	-7.504	2.28e-13 ***
PDRB	-2.245719	0.560333	-4.008	6.91e-05 ***
OWS	0.429265	0.065054	6.599	9.21e-11 ***
D0.3FGAP	-8.779985	3.150181	-2.787	0.00549 **
D16.3PFGAP	-15.427794	5.816568	-2.652	0.00821 **

DOI: 10.54254/2755-2721/6/20230413

Assume that NBA draft rank, the dependent variable, is a numerical variable. Given this, a small number of the dependent variable suggests that the player has a high rank. The smaller number of the dependent variable, the higher rank a player is. Therefore, a negative coefficient is a normal value, which indicates that the variable makes a positive contribution to rank.

From the result, PPTS has the strongest association with NBA draft rank, OWS has the second strongest association with NBA draft rank, G has the third strongest association with NBA draft rank, PDRB has the fourth strongest association with NBA draft rank, D0.3FGAP has the fifth strongest association with NBA draft rank and D16.3PFGAP has the sixth strongest association with NBA draft rank. At the same time, G, PPTS, PDRB, D0.3FGAP, D16.3PFGAP have negative coefficients and OWS has a positive coefficient. The explanation of each coefficient is in the following part.

PPTS: When NBA teams choose new players to enhance their strength, the first thing that they care about is the ability of getting points. They are inclined to choose those players who will have higher points in each game. It's not difficult to find that almost every super star in NBA history like Kobe, James and Jordan can acquire high points in each game throughout his career and that almost every super star has a high rank in NBA draft, for example, Kobe was the 13th pick, James was the 1st pick and Jordan was the 3th pick. Consequently, it's easy to explain that PPTS has the strongest relationship with NBA draft rank and that the coefficient of PPTS is negative.

OWS: Only when a player makes great contributions to the winning games, people will deem that this player is an extraordinary player. Provided this, NBA teams intend to choose players who may make great contributions to the future games. Accordingly, OWS has a strong association with NBA draft rank. But it's interesting that the coefficient of OWS is a positive number, which means that if one player is a highly picked one, his career offensive win share is lower than that of other players who is a lowly picked one contemporarily. Writers reckon that those highly picked players own strong offensive ability before they attend NBA league, thus after they are chosen by NBA teams, they might train their defensive skills to make their opponents acquire low points in every game. However, for those lowly picked players, they have relatively low offensive ability before they attend NBA league, thus after they are chosen by NBA teams, they must train their offensive skills first to get points effectively so that they won't be discarded by their teams. This might be the reason why the coefficient of OWS is a positive number.

G: The number of total games reflect the career length of one player. Usually, only those good players are needed by NBA teams, thus the career length of one player can represent the quality of the player to some degree. NBA draft rank virtually is the expectations of NBA teams to those players on the draft. A high rank means NBA teams believe the player will have a good performance during his career. The longer the length of a player's career is, the better performance the player has and the higher rank the player is. This explains why G has a strong association with rank and has a negative coefficient.

PDRB: When opponents fail to make shoots, players often need to grab rebounds before their opponents grab rebounds to ensure that their opponents won't offend again. In a sense stopping opponents from offending again is another way of acquiring points. Therefore, like PPTS, the more rebounds one player grabs, the higher rank he is. That's the reason that PDRB has a strong association with rank and that the coefficient of PDRB is a negative one.

D0.3FGAP: When playing games, some players sometimes may make a terrible shooting decision like logo shoot. Therefore, whether players can make a good shooting decision is always considered by NBA teams. Attempting to shoot 0-3 feet from the backboard usually indicates that players choose to lay up or choose to offend again after they grab offensive rebounds. NBA teams believe that this way of getting points is the most effective scoring method. The more times one player chooses this scoring method, the more effective player he is. NBA teams tend to choose players having effective scoring methods. Therefore, the times of adopting this scoring method for those highly picked players always are bigger than that for those lowly picked players. This explains why D0.3FGAP has a strong association with rank and the coefficient of it is a negative one.

D16.3PFGAP: In addition to choose the scoring way of attempting to shoot 0-3 feet from the backboard, players will choose to shoot over 16 feet from the backboard but under three-point line. This

DOI: 10.54254/2755-2721/6/20230413

is considered as another effective scoring method by NBA teams. Like D0.3FGAP, the more times one player chooses this scoring method, the higher rank he is. This explains the strong association between D16.3PFGAP and rank and the coefficient of it is a negative value.

3.1.2. LMRS

Use the rookie season data of each NBA players who are on the draft from 2009 to 2018 to fit the multiple linear regression model LMC. Choose nine important variables essential for evaluating a player's ability as explanatory variables: G, GS, DRBP, USGP, ASTP, PMP, ORBP, STLP, OWS. Choose NBA draft rank as dependent variables. Use stepwise regression and find some new variables: P3PP, PPTS, TSP, PBLK, PTOV, FG3P. Add these variables to original variables, filter out those variables whose p-values are larger than 0.05 and acquire the final linear model LMC.

$$\widehat{Rank} = \hat{\alpha} + \widehat{\beta_1} * G + \widehat{\beta_2} * PPTS + \widehat{\beta_3} * TSP + \widehat{\beta_4} * DRBP + \widehat{\beta_5} * PBLK + \widehat{\beta_6} * OWS + \hat{\varepsilon} \quad (1)$$

Estimated coefficients of LMRS are shown in table 3.

Table 3. Estimated Coefficients of LMRS.

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	37.2831	2.637	14.138	< 2e-16 ***
G	-0.132	0.0303	-4.356	1.56e-05 ***
PPTS	-1.4637	0.1712	-8.549	< 2e-16 ***
TSP	9.28	4.3152	2.151	0.031915 *
DRBP	-0.2555	0.1039	-2.459	0.014208 *
PBLK	-2.872	0.7961	-3.608	0.000335 ***
OWS	1.9842	0.5056	3.924	9.72e-05 ***

The result indicates that PPTS has the strongest association with rank, G has the second strongest association with rank, OWS has the third strongest association with rank, PBLK has the fourth strongest association with rank, DRBP has the fifth strongest association with rank and TSP has the sixth strongest association with rank. In addition to this, the model also showcases that the coefficients of G, PPTS, DRBP, PBLK are negative values and the coefficients of TSP and OWS are positive values. In LMC, G and PPTS also have strong associations with rank. Besides, in LMC, PDRB shows a strong association with rank and in LMRS, DRBP also has a strong association with rank. The difference between these two variables is that PDRB is a number and DRBP is a percentage. They reflect the same ability of players. Consequently, it's reasonable to believe the reason that why these three variables have strong associations with rank and why the coefficients of these three are negative values in LMRS is the same as that in LMC. Writers will only explain other three variables. Explain the coefficients of TSP, OWS, PBLK in the following part.

TSP: An unparalleled player always has many offensive methods, like shooting, lay-out, dunk… Among them, shooting is an important offensive method and TSP evaluates whether a player is good at shooting, for it is a comprehensive index that is calculated by three types of field goal percentage: 2-point shooting percentage, 3-point shooting percentage and free throw shooting percentage. Accordingly, NBA teams always concern that if the player that they choose will have a high TSP in his future game. Provided this concern, NBA teams always choose those players who will have high TSP in their future game. Obviously, the coefficient might be a negative value, but in LMRS, the coefficient is a positive value. This originates from the influence of other variables. That the association between G and rank is stronger than the association between TSP and rank makes the coefficient of TSP is a positive value, that is, the less contribution of TSP to rank causes the positive coefficient.

OWS: Like the same coefficient in LMC, it is obvious that OWS has a strong association with rank. In LMC, the data is all players' career data. However, in LMRS, the data is all players' rookie season

DOI: 10.54254/2755-2721/6/20230413

data. Thus, the explanation in LMC is not effective. The explanation of the positive coefficient is that OWS is influenced by PTS and the less contribution of OWS to rank makes the positive coefficient. It is called net suppression [6]. Net suppression happens when an explanatory variable has a negative correlation with the dependent variable but the coefficient of the explanatory variable is positive. This is consistent with the situation of OWS.

PBLK: The number of blocks reflects one player's defensive ability. Not only will NBA teams choose those players having offensive talents, but choose those players having defensive ability. If one player can block opponents many times, it's no doubt that he is a good defensive player. NBA teams will choose those players who can block opponents as many times as possible. Given this, it's clear that PBLK has a strong association with rank and the coefficient of PBLK is a positive value.

3.2. Forward model

Based on the collected data from college careers and the previous season of entering the NBA, separate linear models could be built to predict draft rankings. However, since it is confused that which set of data can fit a better linear model, a good solution is to try to fit linear models respectively, to compare the results and to choose the best fitted model. Accordingly, LMCC which uses college career data and LMCS which uses college season data are built.

3.2.1. LMCC

Use college career data for NBA draft players from 2009-2018 to fit the multiple linear regression model LMCC.

There are 14 variables that might be used in the linear model: Rank, TG, FGP, P3PP, FTP, PMP, PPTS, PAST, PORB, PDRB, PSTL, PBLK, PTOV, PPF. Among them, Rank is the dependent variable and other variables are explanatory variables. P3PP has some strange values which are greater than 0.5. It's unreasonable that one player's 3-point percentage per game exceeds 0.5. To fix these unreasonable values, change those values to 0.5. Next, use stepwise regression. It is shown that some explanatory variables whose p-values are greater than 0.05 are filtered out. The rest of explanatory variables are TG, FGP, P3PP, PPTS and PSTL. They are used to fit the linear model LMCC.

$$\widehat{Rank} = \hat{\alpha} + \hat{\beta} * TG + \widehat{\beta_2} * FGP + \widehat{\beta_3} * P3PP + \widehat{\beta_4} * PPTS + \widehat{\beta_5} * PSTL + \hat{\varepsilon} \quad (3)$$

Estimated coefficients of LMCC are shown in table 4.

Table 4. Estimated Coefficients of LMCC.

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	73.58084	8.09070	9.094	< 2e-16 ***
TG	0.18267	0.01713	10.661	< 2e-16 ***
FGP	-76.77532	12.43775	-6.173	1.43e-09 ***
P3PP	-15.81239	7.10728	-2.225	0.026560*
PPTS	-0.94674	0.20325	-4.658	4.15e-06 ***
PSTL	-5.57666	1.65905	-3.361	0.000838 ***

The result shows that TG has the strongest association with rank, FGP has the second strongest association with rank, PPTS has the third strongest association with rank, PSTL has the fourth strongest association with rank and P3PP has the fifth strongest association with rank. The negative coefficient of FGP shows that FGP makes a positive contribution to a higher rank; the negative coefficient of P3PP explains that P3PP makes a positive contribution to a higher rank; the negative coefficient of PPTS showcases that PPTS makes a positive contribution to a higher rank; the negative coefficient of PSTL manifests that PSTL makes a positive contribution to a higher rank. The most interesting thing is the coefficient of TG is a positive value, which means that TG makes a negative contribution to a higher rank. This can be explained by the fact that many great players always attend NBA draft after they finish

DOI: 10.54254/2755-2721/6/20230413

their first college season and that only those not that good players will play four seasons. Accordingly, the more games one player attends in college, the worse player he is. This is the reason why the coefficient of TG is a positive number.

3.2.2. LMCS

Use the last season data in college for NBA draft players from 2009-2018 to fit the multiple linear regression model LMCS. Choose fourteen important variables essential for evaluating a player's ability as explanatory variables: Rank, TG, FGP, P3PP, FTP, PMP, PPTS, PAST, PORB, PDRB, PSTL, PBLK, PTOV, PPF. Among them, Rank is the dependent variable and other variables are explanatory variables. Next, use stepwise regression. It is shown that some explanatory variables whose p-values are greater than 0.05 are filtered out. The rest of explanatory variables are FGP, PSTL and PTOV. They are used to fit the linear model LMCC.

$$\widehat{Rank} = \hat{\alpha} + \widehat{\beta_1} * FGP + \widehat{\beta_2} * PSTL + \widehat{\beta_3} * PTOV + \hat{\varepsilon} \quad (4)$$

Estimated coefficients of LMCS are shown in table 5.

Table 5. Estimated Coefficients of LMCS.

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept)	69.855	7.802	8.954	< 2e-16 ***
FGP	-57.756	13.383	-4.314	1.95e-05 ***
PSTL	-4.36	1.674	-2.604	0.00949 **
PTOV	-3.395	1.238	-2.743	0.00632 **

The result shows that FGP has the strongest association with rank, PTOV has the second strongest association with rank and PSTL had the third strongest association with rank. The coefficients of FGP, PSTL and PTOV are negative values, which indicates that these three variables make positive contribution to a higher rank.

After comparing the multiple R-squared of LMCC and the multiple R-squared of LMCS, it is believed that LMCC is a better linear model. Given this, adopt LMCC as the final linear model.

3.2.3. Prediction

After using LMCC to predict the draft rankings of NBA players, it is found that the predicted rankings of several players in LMCC are so different from the actual draft rankings that these seem to be outliers. Two obvious outliers are chosen from those outliers: Lester Hudson whose predicted ranking is the 1st picked player and actual ranking is the 58th picked player and Ekpe Udoh whose predicted ranking is the 49th picked player and actual ranking is 6th picked player. Analysis about these two players is shown in the following part.

(1) Lester Hudson

Lester Hudson is an overestimated player in the prediction. In his scouting report [7], he does not have experience in the high-level basketball league, thus he might not adapt NBA match's pace and intensity. When in college, his FGP and P3PP were 0.456 and 0.37 [8]. However, after he entered NBA, his career FGP and P3PP decreased to 0.375 and 0.277 [9], respectively. This strongly demonstrates that he doesn't adapt NBA league at all. What's more, he lacks self-discipline. It's a disastrous disadvantage for him, a Point Guard. As a Point Guard, he needs to organize the whole team's offense. This organization ability will help the team to win the match more easily. As a result of the lack of the ability, he cannot run the team. The last point about his low rank is that when he attended NBA draft, he was very old and he had reached his peak of basketball ability. Therefore, he doesn't have a lot of potential to be explored. NBA teams are more willing to explore the potential of young players.

(2) Ekpe Udoh

Ekpe Udoh is an underrated player. Compared with other players, his FGP (total goals scored on the

DOI: 10.54254/2755-2721/6/20230413

field), PPTS (Per game Points Per Game) and P3PP (3-Point Field Goal Percentage) are not very high statistically. According to his scouting report [10], this could be due to his excellent passing ability, field vision and his willingness to pass the ball to his teammates to create scoring opportunities. These factors are not considered into LMCC. In addition to this, in LMCC, PSTL is also an essential factor influencing NBA draft rankings. However, he isn't good at stealing basketballs in that he is an amazing offensive rebounder, which also greatly affects his ranking in LMCC.

4. Conclusion

It's a tough task to investigate factors that influence NBA draft rankings, for NBA draft rankings are a combination of many quantifiable factors and non-quantifiable factors. This study only investigates those quantifiable factors. The results from LMS and LMRS showcase that G, PPTS and PDRB have the strongest associations with NBA draft rankings. When researchers do some other experiments related to NBA draft and game performances of NBA players, they can think about these three factors preferentially. In LMCC, TG, FGP, P3PP, PPTS and PSTL have the strongest associations with NBA draft rankings. When researchers are inclined to investigate the relationship between NBA draft rankings and NBA players' college performances, they can consider these five factors firstly. However, in LMCC, only those quantifiable factors have been considered into the model and those factors that are not statistically measurable but are essential ones evaluating NBA players' potential and future developments aren't be considered, which causes that some players' abilities cannot be reflected by LMCC, such as static ability, athletic abilities, players' attitude and so on. Consequently, when investigating this, people can adopt other machine learning models to see whether using other methods can obtain more accurate results.

Acknowledgement

It's our pleasure to take Professor John Emerson's program. We have learned many practical skills from Professor John Emerson. We also acknowledge the contribution of Hong Pan, for she provides some useful solutions. Besides, Xi Zheng and Sicheng Zhou also offer some suggestions. Given that, we are sincerely grateful for their contributions.

References

[1] Fang, W., Liu, M. R. (2007) The Origin of NBA and Its Recipe for Success. Hubei Sports Science, (04), 414-415.
[2] Xie, K. (2014) The analysis of the defects about the system of selecting new talents in NBA. Journal of Shaoguan University, 35(06), 60-65.
[3] Berri, D. J., Brook, S. L., Fenn, A. J. (2011) From college to the pros: Predicting the NBA amateur player draft. Journal of Productivity Analysis, 35(1), 25-35.
[4] Staffo, D. F. (1998) The development of professional basketball in the United.. Physical Educator, 55(1), 9.
[5] Basketball-reference.com. https://www.basketball-reference.com.
[6] Smith, R. L., Ager, J. W., Williams, D. L. (1992) Suppressor Variables in Multiple Regression/Correlation. Educational and Psychological Measurement, 52(1), 17–29.
[7] Lester Hudson's NBA scouting report (n.d.) https://www.nbadraft.net/players/lester-hudson/NBA China (n.d.). NBA Frequently Asked Questions. https://china.nba.cn/topics/faq/index.htm
[8] Espn.com. Lester Hudson. https://www.espn.com/nba/player/stats/_/id/3996/lester-hudson.
[9] Nba.com. Lester Hudson. https://www.nba.com/stats/player/201991/career.
[10] Ekpe Udoh's NBA scouting report (n.d.) https://www.nbadraft.net/players/ekpe-udoh/