PJSS

ISSN: 1309-0356

Pamukkale Journal of Sport Sciences
2014, Vol.5, No.1, Pg:48-60

Received : 04.04.2013
Accepted : 17.12.2013

Barış Ergül[1], Arzu Altın Yavuz[1], Hasan Serhan Yavuz[2]

[1]Eskisehir Osmangazi University, Art and Science Faculty, Statistics Department, Eskisehir, Turkey
.[2]Eskisehir Osmangazi University, Engineering and Architecture Faculty, Electrical and Electronics Engineering Department, Eskisehir, Turkey

bergul@ogu.edu.tr

ORIGINAL ARTICLE

CLASSIFICATION OF NBA LEAGUE TEAMS USING DISCRIMINANT AND LOGISTIC REGRESSION ANALYSES

Abstract

The National Basketball Association (NBA) is one of the most popular and well-established men's professional basketball leagues in the world. Predicting of NBA playoffs between NBA teams poses a challenging problem of interest to statistical science as well as the general public. We concentrated on modeling and determining the variables from game-related statistics that affect the teams playing in the playoffs by using the performance values of the teams during the regular season. Finally we show how it could have been used to predict the 2011-2012 NBA playoff results with Discriminant and Logistic Regression Analyses.

Keywords: NBA teams, playoff, discriminant analysis, logistic regression analysis

INTRODUCTION

In the most general sense, sports can be defined as a "game, procrastination, and being away from working" (Mass, 2002). Sports require intensifying an individual's effort, and this effort in conjunction with the aim of being superior results in the selection of those who are more capable of intense physical work for increasingly competitive leagues or competitions. (Fisek, 1998). Historically, sports have undergone some changes, and sports professionals have extensively been involved in these changes. Professional team sports are characterized by an abundance of mineable data, defined objectives, and net results (Berri and Schmidt, 2002).

It is difficult to remember all the details of the events during a competition, and there may be no need for this because the entirety of the details generally contains redundant information. However, some points could be critical, and they could provide information important to measuring performance. Performance measurement is generally conducted and interpreted in terms of sports statistics. At this point, sports statistics is relied on as one of the most important tools with which to collect information related to the observed movements taking place in sports competitions, to organize and evaluate them in line with the determined targets, and to realize the required modifications to increase the performance of a team (Smith et al., 1996). The statistical information generated can be effective for increasing the productivity of an individual player's performance and a coach's decision-making process with regard to the team's goals and strategies.

Studies investigating the collection and analysis of information related to basketball players and team performance in terms of statistical productivity are quite recent. In the 1946-1947 NBA season, statistics regarding assists, personal fouls, and factors associated with scoring were first collected. In the 1977-1978 season, some factors such as turnovers, steals, and blocks were added (Berri and Schmidt, 2002). It is quite difficult to analyze the data collected for basketball, to determine basic performance indicators related to success, and to interpret them. To find performance indicators for players and their teams, basketball researchers generally compare specific performance parameters with each other (Tavares and Gomes, 2003).

The research conducted in this area in recent years has become increasingly pursued by scientists, professional sports teams and coaches. Schwertman et al. (1996) and Carlin (1996) tackled NCAA basketball. Both papers estimated the probability that team i beats j. The first used a logistic regression analysis of win - loss records. Trninic et al. (1997) analyzed the differences between the winning and losing teams in basketball. Sampaio et al. (2006) utilized discriminant analysis to reveal the relationship between the positions of players and game-related statistics in three professional basketball leagues. . Ibanez et al. (2008) determined the game-related statistics affecting teams' successes and failures throughout the season by using discriminant analysis. Gomez and Pollard (2011) investigated the impact of home advantages of basketball teams using statistical techniques.

NBA league teams obtain rights to play in the playoffs according to their performances in the regular season. In this study, we concentrated on modeling and determining the variables from game-related statistics that affect the teams playing in the playoffs by using the performance values of the teams during the regular season. First, the models were obtained by using discriminant analysis and logistic regression; then, the variables affecting a team's success in joining the playoff in NBA were determined. After the determination of the variables, we also compared the classification performances of the methods to find the technique that best approximated actual historical outcomes.

METHODS

Sample

In this study, we considered the data of the 30 teams playing in the 2011-2012 NBA season from the official web site of the NBA (http://www.nba.com (2012)) The game-related statistics collected from the site consist of (1) own points per game, (2) opponent's points per game, (3) own field goal percentage, (4) opponent's field goal percentage, (5) own three-point field goal percentage, (6) opponent's three-point field goal percentage, (7) own free throw percentage, (8) opponent's free throw percentage, (9) own assists per game, (10) opponent's assists per game, (11) own total rebounds per game, (12) opponent's total rebounds per game, (13) own blocks per game, (14) opponent's blocks per game, (15) own steals per game, (16) opponent's steals per game, (17) own fouls, and (18) opponent's fouls. As a result, the data included 18 independent variables that were thought to affect the performance of the teams.

Quantitative analysis of basketball performance through game-related statistics is being widely used amongst coaches and teams in order to analyse game events with more valid data. (Ibanez et al. ,2008).

The dependent variable was then defined to be a team's playoff status, coded as follows:

$$Y_i = \begin{cases} 1, & \text{A team that failed to make the playoffs} \\ 2, & \text{A team that played in the playoffs} \end{cases}$$

Data analysis

After determining the dependent and independent variables, the problem of classification simply becomes a statistical decision-making process. This decision-making process has two stages in general. In the first stage, the variables that are effective for the distinctive features of the group are determined; in the second stage, elements are assigned into the proper group by using some discriminative functions inferred from the class model. In this study, due to the availability of prior information, we identified discriminant analysis and logistic regression analysis as the classification methods most suited for our objectives from the myriad methods proposed in the literature.

Discriminant analysis and logistic regression analysis are considered to be robust with selected variables and we want to compare this analyses which has good perfomance.

Discriminant analysis is one of the most efficient techniques used in classification. Under the assumption of normal distribution, discriminant analysis aims to minimize the within-class scatter of observations from the same class while maximizing the between-class scatter of observations belonging to different classes. To achieve this end, the method uses some mathematical equations, which are called the discriminant function (Johnson and Wichern, 2002; Klecka 1980; Lachenbruch, 1975). The main objective of the method is to classify like-observations in the same group and unlike-observations in different groups.

Logistic regression analysis is an alternative technique when discriminant analysis assumptions are not met. The main purpose of logistic regression analysis is to explain the relationship between the independent variables and the dependent variable with the help of the fewest number of variables. There is no assumption in logistic regression analysis that the

dependent variable should be continuous. It is used especially when the dependent variable takes two or more qualitative values (Albert and Lesaffre, 1986; Aldrich and Nelson, 1984; Hair et al., 1995; Lemeshow and Hosmer, 2000).

In this study, we classified the NBA teams into two groups—playoff and non-playoff—by examining their regular season performance statistics via discriminant analysis and logistic regression analysis methods. The statistical analysis was performed by using SPSS software, version 18.0.

RESULTS

The descriptive statistics of the independent variables are given in Table 1. A *t*-test for independent samples was performed to identify univariate differences between the game-related statistics between playoff and non-playoff teams.

Playoff teams have statistically significant differences ($p \leq 0.05$) from non-playoff teams in own points per game, opponent points per game, own field goal percentage, opponent field goal percentage, opponent assists per game, own total rebounds per game, opponent total rebounds per game, own steals per game.

Utilizing the technique of variable selection in discriminant analysis, we determined the variables that should be included in the model. The results are given in Table 2.

Because there were two groups (playoff and non-playoff) in the problem, only one discriminant function was obtained. The obtained discriminant function was found to be statistically significant ($p \leq 0.05$). Having a large eigenvalue implies that the independent variables explain the dependent variable to a high degree. As given in Table 2, the eigenvalue was found to be 4.991 in the model and thus explained 88.3% of the variance. In addition, the canonical correlation coefficient was found to be 0.913. The square of this coefficient (R^2) was 0.986. It demonstrates that independent variables explained the dependent variable with the ratio of 98.6%. Wilks' lambda is computed as the ratio of the group variances that are not explained by the dependent variable to the total variance. In the model, this value was 0.167. It shows that 16.7% of the variance could not be explained by the differences between groups.

Table 1: Means And Standard Deviations For Playoff and Non-Playoff NBA Teams

Game Related Statistics	Playoff	Non-Playoff
Own Points Per Game*	97.67	94.65
	(3.53)	(4.00)
Opponent's Points Per Game*	94.21	98.60
	(3.46)	(2.90)
Own Field Goal Percentage*	0.46	0.44
	(0.01)	(0.01)
Opponent's Field Goal Percentage*	0.44	0.46
	(0.01)	(0.01)
Own Three-Point Field Goal Percentage	0.35	0.34
	(0.02)	(0.02)
Opponent's Three-Point Field Goal Percentage	0.35	0.35
	(0.02)	(0.02)
Own Free Throw Percentage	0.75	0.76
	(0.04)	(0.02)
Opponent's Free Throw Percentage	0.75	0.76
	(0.01)	(0.01)
Own Assists Per Game	21.33	20.57
	(1.74)	(1.41)
Opponent's Assists Per Game*	20.11	21.97
	(1.16)	(1.24)
Own Total Rebounds Per Game*	42.87	41.39
	(1.89)	(1.27)
Opponent's Total Rebounds Per Game*	41.48	42.98
	(1.32)	(1.80)
Own Blocks Per Game	5.07	4.90
	(0.52)	(0.66)
Opponent's Blocks Per Game	4.97	5.27
	(0.71)	(0.65)
Own Steals Per Game*	7.99	7.35
	(0.94)	(0.76)
Opponent's Steals Per Game	7.42	7.81
	(0.87)	(0.34)
Own Fouls	19.27	19.91
	(1.71)	(1.30)
Opponent's Fouls	19.75	19.35
	(1.60)	(1.24)

*Note** $p \leq 0.05$ statistically significant. Standard Deviations appear in parantheses below means.

Table 2: Discriminant Function, Structure Coefficients And Tests of Statistical Significance For NBA Teams

Game-related statistics	SC
Opponent's Field Goal Percentage	−0.386
Opponent's Assists Per Game	−0.358
Opponent's Points Per Game	−0.316
Own Field Goal Percentage	0.301
Opponent's Total Rebounds Per Game	−0.222
Own Total Rebounds Per Game	0.205
Own Points Per Game	0.186
Own Steals Per Game	0.173
Own Three-Point Field Goal Percentage	0.161
Opponent's Steals Per Game	−0.134
Own Assists Per Game	0.110
Opponent's Blocks Per Game	−0.100
Own Fouls	−0.098
Own Free Throw Percentage	−0.094
Opponent Free Throw Percentage	−0.093
Own Blocks Per Game	0.063
Opponent's Fouls	0.061
Opponent's Three-Point Field Goal Percentage	−0.034
Eigenvalue	4.991
Wilks' lambda	0.167
Canonical correlation	0.913
Chi-squared	34.016
Significance	0.013
Reclassification	86.7%

The discriminant function and the coefficients of the variables in the function obtained in the analysis were given in (1).According to these results, the variables affecting the NBA teams' playoff eligibility were found to be opponent's points per game, own field goal percentage, opponent's field goal percentage, and opponent's assists per game. The resulting discriminant function that classifies the NBA teams is given in (1).

$$v_i = -4.392 + 0.103 \times (\text{Opponent Points per Game}) - 45.539 \times (\text{Own Field Goal Percentage}) + 21.609 \times (\text{Opponent Field Goal Percentage}) + 0.248 \times (\text{Opponent Assists per Game}) \quad (1)$$

As shown in Table 2, the correct classification ratio of the obtained discriminant function was 86.7%.

In the portion of the study using logistic regression analysis, the forward-stepwise (conditional) method was used to determine which independent variables best explained significant increases in the dependent variable. In this method, independent variables are added sequentially to the model, which then attempts to identify the model that explains the dependent variable with inclusion of the fewest variables. Beyond determining the best model for classification, this method measures the degree of influence each independent variable has on the dependent variable. In accordance with various criteria, the result of the analysis is summarized in Table 3.

In the first step, while other independent variables were constant, the variables Own Points per Game and Opponent Field Goal Percentage significantly explained a team's status in the playoffs. As can be seen in Table 3, coefficients of Own Points per Game and Opponent's Field Goal Percentage were considered to be significant because p values were less than 0.05.

Table 3: Logistic Regression Results With Forward-Stepwise (Conditional) For NBA Teams

Variables	**Logit B**	**Std. error**	**Wald**	***p***	**Exp(B)**
Constant	81.282	45.974	3.126	0.077	1.998E35
Own Points Per Game	0.666	0.316	4.440	0.035	1.946
Opponent's Field Goal	-324.900	142.282	5.214	0.022	0.000
Goodness of Fit Test Results					
Cox and Snell R^2				0.586	
Nagelkerke R^2				0.783	
Hosmer and Lemeshow Test				5.203	
−2LogLikelihood				14.974	
Correct Classification Percentage for Logistic Regression Model					
Observed	Predicted Cluster				Percentage Correct
	Non-Playoff	Playoff			
Non-Playoff	12	2			85.7
Playoff	1	15			93.8
Overall Percent					90.0

In the logistic regression model, Hosmer-Lemeshow statistics (model chi-square statistics) were used for testing the goodness-of-fit of the model formed for explaining dependent variable. Hypotheses were formulated as follows:

$H_0: b_0 = b_1 = b_2 = \cdots = b_k = 0$ (Model is meaningless in general)

$H_1: b_0 \neq b_1 \neq b_2 \neq \cdots \neq b_k \neq 0$ (Model is meaningful in general)

That the *p* probability value obtained for the Hosmer-Lemeshow test was greater than the 0.05 significance level suggested that the model was meaningful. The −2LogLikelihood value was also calculated at 14.974. This value could be compared with the chi-square table value of 3.84 at 1 degree of freedom at a significance level of $\alpha = 0.05$. Because −2LogLikelihood value was greater than the critical table value, the model was significant in general. That the *p* probability value (0.736) obtained for the Hosmer-Lemeshow test was greater than the 0.05 significance level suggested that the model was meaningful. The −2LogLikelihood value was also calculated at 14.974. This value could be compared with the chi-square table value of 3.84 at 1 degree of freedom at a significance level of $\alpha = 0.05$. Because −2LogLikelihood value was greater than the critical table value, the model was significant in general.

In the analysis, the Cox and Snell and Nagelkerke R^2 statistics showed the explanation rate, which is a dependent variable of the independent variable. According to the Cox and Snell R^2 statistics, the ratio was 58.6%. According to the Nagelkerke R^2 statistic, it was approximately 78.3%. According to the results of the logistic regression analysis, Own Points per Game and Opponent's Field Goal Percentage both played the most important role in playoff eligibility prediction. The logistic regression model that provided the classification of playoff performances of the NBA teams was as given in (2):

$$\hat{\pi}(x) = 0.666 \times (\text{Own Points per Game}) - 324.9 \times (\text{Opponent Field Goal Percentage}) \quad (2)$$

When the values of other variables were held constant, a one-unit increase in the value of the variable own points per game increased the odds of playoff eligibility 1.946-fold.

The classification table is another goodness-of-fit criterion used in logistic regression analysis. By using the classification table, we evaluated that the established model correctly classified 90% of cases for the model based on two independent variables. This result showed that logistic regression model has a good prediction rate. According to the results of both analyses, the real and estimated values related to NBA teams' playoff eligibilities are summarized in Table 4.

Table 4: Classification of NBA Teams

Team Name	Real	Discriminant Analysis	Logistic Regression
Denver	Playoff	Playoff	Playoff
San Antonio	Playoff	Playoff	Playoff
Oklahoma	Playoff	Playoff	Playoff
Utah	Playoff	Non-Playoff	Playoff
Milwaukee	Non-Playoff	Non-Playoff	Playoff
Sacramento	Non-Playoff	Non-Playoff	Non-Playoff
Miami	Playoff	Playoff	Playoff
Phoenix	Non-Playoff	Non-Playoff	Non-Playoff
Houston	Non-Playoff	Non-Playoff	Non-Playoff
Minnesota	Non-Playoff	Non-Playoff	Non-Playoff
New York	Playoff	Playoff	Playoff
Golden State	Non-Playoff	Non-Playoff	Non-Playoff
Indiana	Playoff	Playoff	Playoff
L.A. Clippers	Playoff	Playoff	Playoff
L.A. Lakers	Playoff	Playoff	Playoff
Portland	Non-Playoff	Non-Playoff	Non-Playoff
Atlanta	Playoff	Playoff	Playoff
Chicago	Playoff	Playoff	Playoff
Dallas	Playoff	Playoff	Playoff
Memphis	Playoff	Playoff	Playoff
Orlando	Playoff	Playoff	Non-Playoff
Philadelphia	Playoff	Playoff	Playoff
Washington	Non-Playoff	Non-Playoff	Non-Playoff
New Jersey	Non-Playoff	Non-Playoff	Non-Playoff
Cleveland	Non-Playoff	Non-Playoff	Non-Playoff
Boston	Playoff	Playoff	Playoff
Detroit	Non-Playoff	Non-Playoff	Non-Playoff
Toronto	Non-Playoff	Playoff	Playoff
New Orleans	Non-Playoff	Non-Playoff	Non-Playoff
Charlotte	Non-Playoff	Non-Playoff	Non-Playoff

As a result of the analyses, the percentage of correct assignments, using the function obtained through discriminant analysis, was 86.7%. Using the model developed through logistic regression analysis, the correct assignment percentage was 90%. Obtaining a higher correct classification percentage of the established model through logistic regression analysis had to do with the variables utilized in the analysis.

The logistic regression analysis has not often been used in current studies of sport statistics on basketball leagues. In this study, the results of logistic regression analysis, which is an important alternative against discriminant analysis, demonstrated that playoff participation of NBA teams could be predicted by using just two variables: own points per game and opponent's field goal percentage. Although the variables in the models from both analyses were similar, the model obtained by logistic regression analysis had fewer variables and more accurately classified teams than the one obtained by discriminant analysis. For this reason, the usefulness of logistic regression analysis is evident because it does not require many assumptions and is a simple but powerful method for giving a classification result.

DISCUSSION

NBA teams are divided into playoff and non-playoff groups based on their performance in the season. In this study, based on their performance in the regular season, NBA teams' likelihoods of participation in the playoffs were modeled using discriminant and logistic regression analyses.

Although known as a game of attack, basketball is actually a defensive game. In a study by Trninic et al.(2000), steals and rebounds cases were discussed as keys to a successful team. As a result of *t*-tests conducted for playoff eligibility, statistically significant differences were found among the following variables: points per game own and opponent, field goal percentage own and opponent, opponent assist per game, total rebounds per game own and opponent, and own steals per game. These results were in parallel with Trininic et al.'s study (2000). For playoff teams, the variables of own points per game, own field goal percentage, own total rebounds per game, and own steals per game had higher values. When the fan pressure during home play and the players' psychological urges to win come together, the teams with more total rebounds, steals, and more field coverage are more likely to be eligible for the playoffs. This result was similar to the findings obtained in Lorenzo et al.'s study(2010), which compared the winning and losing teams in the U-16 league in the European Championship.

In the study by Sampaio et al(2006), the variables assists, rebounds and field goals by centers and forwards playing in the NBA league were important. In our study of NBA teams, the important variables determined with the help of discriminant analysis were field goal percentage (own and by an opponent), opponent's assists per game, and opponent's points per

game. Because a team's performance depends on its players, the similarity between these two results increases the reliability and accuracy of the study.

The variables in both the model based on the discriminant analysis and the model based on the logistic regression analysis have been found meaningful in performed *t*-tests implying that the results reveal the most important variables of basketball statistics for NBA teams to join playoff. These findings will be useful in terms of their role in reminding coaches that team performance is more important than the performance of individual players in NBA.

Also, they show that it's better for coaches to concentrate on own statistics of field goal and points per game and opponent's statistics of field goal, points per game and assists per game. They will be driving forces for teams to play toward playoff eligibility.

REFERENCES

Albert, A., Lesaffre, E. (1986). Multiple Group Logistic Discrimination, Computational Mathematics with Applications, 12 A, 2: 209-224.

Aldrich, H. J., Nelson, D. F. (1984). Linear Probability, Logit and Probit Models, Series: Quantitative Applications in the Social Sciences, No:45, USA: Sage University Paper.

Berri, D. J., Schmidt, M. B. (2002). Instrumental vs. bounded rationality: The case of Major League Baseball and the National Basketball Association, Journal of Socio-Economics, 31(3): 191-214.

Carlin, B.P. (1996). Improved NCAA basketball tournament modelling via point spread and team strength information, The Amer Statistician, 50, 39-43.

Fisek, K. (1998). Devlet Politikası ve Toplumsal Yapıyla Iliskileri Acısından Dunya'da ve Turkiye'de Spor Yonetimi, BagirganYayınevi, Ankara.

Gomez, M.A., Pollard, R. (2011). Reduced home advantage for basketball teams from capital cities in Europe, European Journal of Sport Science, 11(2): 143-148.

Hair, J., Rolph, E., Ronald, L. and William, C. (1995). Multivariate Data Analysis with Readings, New York: Prentice Hall International Editions.

Ibanez, S., Sampaio, J., Feu, S., Lorenzo, A., Gomez, M. and Ortega, E. (2008). Basketball game-related statistics that discriminate between teams season-long success, European Journal of Sport Science, 8: 369–372.

Johnson, R. A., Wichern, D.W. (2002). Applied Multivariate Statisticial Analysis, New Jersey Pearson Education Int.

Klecka, W. (1980). Discriminant Analysis, London: Sage Publications.

Lachenbruch, P.A. (1975). Discriminant Analysis, London: Hafner Press.

Lemeshow, S., Hosmer, D. (2000). Applied Logistic Regression, Wiley Series in Probability and Statistic, 2. Edition, New York: Wiley Interscience.

Lorenzo, A., Gomez, M., Ortega, E., Ibanez, S. and Sampaio, J. (2010). Game Related Statistics Which Discriminate Between Winning and Losing Under-16 Male Basketball Games, Journal of Sports Science and Medicine, 9: 664-668.

Mass, C. M. (2002). Webster's Third New International Dictionary, Unabridged, Merriam Webster, http://unabridged.merriam-webster.com, Retrieved 15 October 2012.

NBA (2012). http://www.nba.com, Retrieved 12 September 2012.

Sampaio, J., Janeira, M., Ibanez, S. and Lorenzo, A. (2006). Discriminant analysis of game-related statistics between basketball guards, forwards and centres in three professional leagues, European Journal of Sport Science, 6: 173–178.

Schwertman, N.C., Schenk, L. and Holbrook, B.C. (1996). More probability models for the NCAA regional baketball tournaments, The Amer Statistician, 50, 34-38.

Smith, N., Handford, C. and Priestly, N., (1996). Sport Analysis in Coaching. Department of Exercise and Sport Science, Crewe and Algeser Faculty, The Manchester Metropolitan University, Manchester.

Tavares, F., Gomes, N. (2003). The offensive process in basketball - a study in high performance junior teams, International Journal of Performance Analysis in Sport, 3(1): 34-39.

Trninic, S., Milanovic, D. and Dizdar, D. (1997). Where are the differences between winning and losing teams in basketball?, School of Sport , 38: 25-35. (In Italian).

Trninic, S., Dizdar, D. and Dezman, B. (2000). Empirical verification of the weighted system of criteria for the elite basketball players quality evaluation, Collegium Antropologicum, 24: 443-465.