Basketball Analytics. Data Mining for Acquiring Performances

Leila Hamdad(✉), Karima Benatchba, Fella Belkham, and Nesrine Cherairi

Ecole nationale Supérieure en Informatique ESI,
BP 68M, 16309 Oued-Smar, Alger, Algeria
l_hamdad@esi.dz
http://www.esi.dz

Abstract. Choices of decision makers in a basketball team are not limited to the strategies to be adopted during games. The most important ones are outside the field and concern team composition and talented and productive players to acquire on which the team can rely to raise its game level. In this paper, we propose to use data mining tasks to help decision makers to make appropriate decisions that will lead to the improvement of the performance of their players and their team. Tasks such as clustering, classification and regression are used to detect weaknesses of a team; best players that can help overcome these weaknesses; predict performance and salaries of players. These will be done on the NBA dataset.

Keywords: Data mining · Basket ball analytics · Clustering
Classification · Performances

1 Introduction

Data mining (DM) is used in many fields and the sports industry is no exception. In fact, sport is an ideal area for the implementation of data mining tasks and techniques. This is due to the amount of statistics collected on players, teams, games and seasons. Team managers wish to understand these data in order to extract information that will help them improve performances of their players and their teams by predicting future performances and affecting players in appropriate training group to make better profits. The Basket ball is one of those sports. A basketball team is composed of five players. A player can have a great impact on the efficiency of the team due to his talent. Indeed, in many cases, a basketball team with new acquisitions has seen its game significantly improved. Sometimes, the reverse phenomena can take place and observers might say that the added talent could not adapt his game to the rest of the team. And in other cases, one has seen teams with modest talents exceeding expectations. The difficulty of finding talents that are coherent with the rest of the team and the uncertainty of the stability of the acquired players' performances amplify

A. Amine et al. (Eds.): CIIA 2018, IFIP AICT 522, pp. 13–24, 2018.
https://doi.org/10.1007/978-3-319-89743-1_2

the risks taken by the team managers in their choices; especially if the league to which they belong is as competitive and exigent as NBA.

NBA is one of the three major leagues in the United States. The results and statistics that it provides make it an ideal subject of study. Indeed, the amount of existing data is accessible and concerns over 1200 games played in a single season (Sports Reference, 2000). Moreover, the measured performances of the players, relative to their different positions, are available. To extract significant knowledge from this amount of data, DM tasks such as clustering, classification, regression can be used. In this work, we have two objectives: 1. Our first goal is to study acquisition of new players. Before acquiring new players, it is necessary to determine team's needs. For this purpose, a clustering of the players is done according to their performances. It will give managers a description of players' performances' levels and hence allow detecting weaknesses that will help in deciding which talent to acquire. 2. The second goal is to insure the stability or performance improvement of the chosen player. This can be done by predicting his future performances and cost according to his efficiency. All the proposed tasks have been tested on the data of 10 NBA seasons (from 2005–2006 to 2014–15).

The rest of the paper is organized as follows: Sect. 2 presents related work in the literature. In Sect. 3, the data used to test different tasks is described. Then in Sect. 4, we present some DM tasks and techniques to be used for Basket ball and run them on NBA dataset in Sect. 5.

2 Related Works

The NBA is one of the major leagues of the US. It represents an extremely competitive environment for players and teams. Moreover, it is the subject of intensive research and perfect for the application of DM due to the amount of data generated by multiple meetings and sport events. In Basketball, it was Dean Oliver who introduced statistical basketball analysis in his book "Basketball on Paper" [4]. Nowadays, it is regarded as a revolution for the Basketball statistical analysis. The literature concerning Basketball analysis is quite abundant. Among these works, one can find those of [5]. They were interested in identifying the variables which had a potential effect on players' performance from a game to another, using linear regression. They, first, used a linear mixed model (LMM), to model Points (P) according to variables with both fix and random effects. Then, they used a generalized linear mixed model (GLMM) to model Win Score (WS). They concluded that minutes played, percent utilization, and team quality difference variables were among the ones that had an impact on P and WS.

[1] studied the impact of a group of two or three players (called *Big 2's* or *Big 3's*) on successive wins of a team. To this end, [1] grouped players, according to their level, in appropriate groups of two or three, using k-means algorithm. This clustering is done according to some features as points per game, offensive rebounds per game, defensive rebounds per game, assists per game, etc. A regression model was then used to measure the impact of the composition of

"*Big 2s*" and "*Big 3s*" on winning teams. He showed that the composition of a team's top 2 and top 3 players is a factor with a high statistical significant in the success of a team, and showed which combinations yielded over-performance, and which combinations yielded underperformance, relative to the team's talent and coaching quality.

The successive victories of teams have also been studied in [8], where their income and victories were evaluated according to the performance of their players. A regression method was used to study the relationship that may exist between *PER* (Player Efficiency Rating) and variables that determine team's wins and earnings. He concluded that the defensive abilities contributed to win games and thus, ensured greater income, knowing that every win brings 3% more revenue. Among the existing works, some focused on prediction. Among them, [6] studied which teams would make the NBA playoffs. They collected and analysed team data using Principal Components Analysis to reduce the dimensionality of the data set and then a Discriminant Analysis to predict the classifications of teams into playoffs or non-playoffs. In their paper, NBA Oracle [2], supervised and unsupervised learning methods are applied for predicting game outcomes and providing guidance and advice for common decisions in the field of professional basketball. For game outcome prediction, four different binary classification techniques are compared according to their accuracy prediction: Linear Regression, SVM (Support Vector Machines), Logistic Regression, and ANN (Artificial Neural Networks). In the same paper, k-Means is applied to infer optimal player positions and Outlier Detection to identify outstanding players. Neural networks have also been used for predictive end in [9]. Some others works focused on the prediction of points scored and therefore the results of games using regression models [11].

The clustering was mainly used in the construction of teams [1]. However, in our work, studying similarities between performances of players is only a first step in the acquisition process. Indeed, the resulting groups aim to determine the weaknesses of a team and help choose the right player(s) that will reinforce the team. This will be done using clustering and classification. Thereafter, our goal is to predict the performance of a player (assume selected) and his salary to determine whether he is a good option for the team. Regression and time series are used for this purpose.

3 Data

We have chosen to use the data source "Basketball-reference.com", a web site with a rich data base. One can find there, statistics on over 60 NBA seasons. They are reliable and contain few missing values [3]. The elements that point out the performance of a player during the game are divided into four categories:

- Defensive performance: these are indicators that ensure the defensive play of a player. They are represented by: blocks, defensive rebounds and steals.
- Offensive performance: indicators that show the offensive play. They are represented by: Offensive personal fouls and rebounds.

- Scoring: all indicators that have a relationship with the scoring and points: the attempts of field goals, attempts goals successful on the field, attempts goals from three points on the ground, the attempts of goals successful three-point field, attempts goals from two points on the ground, successful goals attempts goals from two points on the ground, free throws, the successful free throws.
- Play-making: It gives information on the participation of a player in building the game on the field. It is represented by: the assists, turnovers, number of games played and minutes spent on the field. The performance of a given team is represented by the sum of performance indicator values of all its players.

We worked on two types of indicators to best meet our needs in terms of players' performance description. The first type consists of basic statistics (rebound, interception points...) as they give a summary of the performance during a particular season. The second type allows a more detailed overview of players' performances (a zoom) according to a particular event (match or possession).

4 Data Mining Tasks for Basket Ball Analysis

In this section we will present the different DM tasks that we used for basket ball analysis.

4.1 Clustering

Clustering is a descriptive task of DM. It consists on partitioning the data into homogenous clusters using similarity measures. The objective of applying clustering on a set of NBA players is to form groups of players statistically homogenous as they differ in their playing style. In this study, the data used for clustering represents the 10 NBA seasons from 2005–06 until 2013–14.

The clustering allows to have an overview on Player's Skills and to compare their efficiency according to the partition they belong to. As each player plays in a particular position, clustering is also used to compare players of the same position according to their performance. It also provides decision makers with elements for future analysis that enable them to identify potential undervalued players or even to propose appropriate salaries to the players.

Moreover, the use of clustering could be extended for comparative purposes. This is done by applying clustering after the prediction in order to compare the salaries of the players based on their future statistics. This is useful when multiple players meet the need of the team. In this case, the one whose value increases will have an advantage over others, whether for performance or salary. We used K-means algorithm for its simplicity and efficiency on large dataset.

4.2 Classification

Classification is a predictive task of DM. It consists on affecting new objects to existing groups. In our case, we apply classification on the results of clustering.

It will determine to which cluster, a new player can belong according to his features. Hence, this classification allows to have an overview on the players competences depending on the cluster to which he belongs and the players therein. This classification is interesting when acquiring a new player specifically when one wants to substitute a player against a new one, having the same characteristics or to fill a gap in the team. Indeed, once the clusters have been defined using a sample of players according to their skills, the classification will determine to which cluster a new player will belong according to its descriptive features.

We have chosen to use Naïve Bayes algorithm (see [10] for more details). It is a supervised classification algorithm based on the Bayes theorem with strong features' independence hypothesis. Its principle is as follows: suppose, we have K clusters, and we went to know cluster of a new entries X. X is affected to a cluster C_k if:

$$P(C_k/X) = P(C_k)P(X/C_k) = max_{j=1,\ldots,K} P(C_j)P(X/C_j). \quad (1)$$

Where, $P(C_j)$ is the apriori probability of the cluster C_j and $P(X/C_j)$ the probability of X in cluster C_j. The advantage of this algorithm is that it requires relatively little training data.

4.3 Prediction

Basketball is a very competitive sport and a great business. As a result, prediction becomes important. Indeed, NBA statistics have particular relevance for team managers and coaches. When acquiring a new player or signing and renewing a contract, it is possible to project the player history in order to predict future trends of his performance and salary, and make the right decisions. This is done by studying and analysing the existing relationships in past occurrences of that player's performances. In this perspective, the multiple linear regression and exponential smoothing on a set of historical data retrieved from the database according to specific needs are used. We can use the results of predictive analysis to evaluate a given player and predict future values of his performance and salary. Moreover, we can predict team's costs.

Performance Prediction. Several metrics and measures are used to evaluate the performance of a player in the NBA. Some represents performance detailed on several axes (defensive, offensive, shooting rate), some reflect the player's contribution to the victory of his team and others are more general and global. We are interested, in this study by predicting the PER (Player Efficiency Rating). It is a performance measure that summarizes the performance of a player during a season in a single number. This metric reflects individual performance of a player in several sports. It was proposed by [7]. In Basketball, Hollinger's formula takes into account many variables that represent the positive achievements (the points scored on the ground (FG), defensive rebounds (DRB) ...), the negative effects (personal faults (FP), team faults (FT) ...) and adjust them according to games' time and games' rhythm. The value of the average PER

of the league is set to 15.00; it allows to compare the performances of players over the seasons. This task allows to predicts the future value of *PER* based on a set of performance indicators selected. To do so, we use linear regression model to predict this performance to evaluate players' performance indicators and selecting those that have a significant effect on the *PER* and predict their values.

Salaries Prediction. Players' salaries through the seasons are considered as time series. Each observation is associated to a particular season. The income of player at time t is a function of previous incomes. The future value of the player's salary can be predicted in short term, using exponential smoothing. This method is used when the series (T observations) are not seasonal.

5 Tests and Results

In this section we will show through different scenarios, how the different tasks of data mining, cited above, can help managers make decisions on acquiring efficient players for basket ball teams. Indeed, we have as objective to group players according to their performance, evaluate and predict player performance and predicting the salaries of players and franchises costs.

5.1 Clustering

We present two scenarios of clustering. In the first one, the players are grouped according to their performances and in the second one, we focus on the move of players from cluster to another across the seasons.

Scenario 1. We will present in what follows the result of a clustering which purpose is to have a comprehensive view of the talents distribution, either in the league or in a team. It also provides an overview of the skills of the players according to the cluster they belong to compared to their opponent. We applied K-means, with $k = 6$, on features of 337 players of season 2012–2013. The used feature were: Minutes played (*MP*), Field Goals (*FG*), Field Goals Attempt (FGA), basket of threes Points (*3P*), basket of threes Points Attempt (*3PA*), basket of two Points (*2P*), basket of two points attempt (2PA), Free Throw (*FT*), Free Throw Attempt (*FTA*), Offensive ReBound (*ORB*), Defensive ReBound (*DRB*), ASsisTs (*AST*), Interception (*STL*), Block (*BLK*), Turnover (*TOV*), Points (*PTS*), Player Efficiency Rating (*PER*). Table 1 represents this clustering by giving the average of each feature for all clusters We distinguish some clusters from others. Indeed cluster1 contains imposing statistics. It is a superstar cluster as the best NBA players belongs there. It contains 25 players from a total of 337. It includes players who have proven their talent by playing exceptional season and winning awards such as: Most Valuable Player (best player of the regular season) and NBA All-Defensive Second Team (best defensive player of

the regular season). They participated in the NBA, All Star and won the NBA championship with their team. Among them, we have: Lebron James, Stephen Curry, Kobe Bryant, Kevin Durant ... According to the results of the Table 1, one can see that cluster 4 includes excellent rebounders who scored many two points baskets. However, cluster 2 is the one with lower statistics particularly in defence (STL, BLK...). Cluster 3 also displays statistics that are far from impressive, particularly in defence. It includes players like Louis Amundson and Joel Anthony who are considered by the site Bleacherreport.com as "50 Most Worthless Players in the NBA for the 2012–13 Season". We can also find Rodrigue Beaubois, a player from whom one expected a lot but he did not have great performances. The distribution of the players is given by the following Fig. 1. Almost half of the sample players are in clusters 2 and 3 with low to medium skills.

Table 1. Clustering results of 337 players on season 2012–2013.

Clus	MP	FG	FGA	3P	3PA	2P	2PA	FT	FTA	ORB	DRB	AST	STL	BLK	TOV	PTS	PER
Clus0	1580	219	419	42	117	177	374	89	119	70	192	131	48	30	83	570	13.17
Clus1	2766	528	1167	119	326	408	840	300	368	64	296	428	112	31	213	1477	14.99
Clus2	389	46	112	12	36	34	76	17	25	16	44	27	11	6	19	124	8.75
Clus3	840	106	241	21	61	58	179	37	54	47	111	53	25	20	40	271	12.35
Clus4	2192	382	751	3	13	378	738	173	248	190	423	126	61	90	126	940	16.28
Clus5	1999	312	723	104	277	207	445	140	172	47	188	209	67	21	115	870	14.74

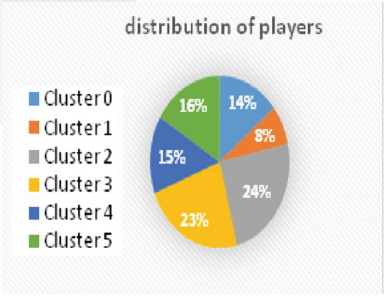

Fig. 1. Distribution of players.

Scenario 2. For this clustering, the goal was to observe a variation in players' performances from a season to another according to their productivity and substitution of players and talented players in the team. For this purpose, we have chosen to work on Phoenix Sun players for the two seasons 2011–12 and 2012–2013. We tested the clustering only on players that were part of the team during the two seasons. K-means with $K = 3$ was used. The results are given

Table 2. Results of K-means (K = 3).

Clusters	FG	FGA	3P	3PA	2P	2PA	FT	FTA	Pts
Cluster 0	420	863	30	98	389	765	169	269	1066
Cluster 1	78	178	7	26	71	152	38	51	202
Cluster 2	257	591	56	163	201	427	87	116	660

Table 3. Repartition of the players in the clusters according to the two seasons 2011–2012, 2012–2013.

Cluster 0	Cluster 1	Cluster 2
G. Dragic	J. Childress	M. Beasley
M. Gortat (2011–2012)	D. Garrett Robin	S. Brown [x2]
L. Scola	L. Kendall	J. Dudley [x2]
	M. Ronnie Price	C. Frye 11.12
		M. Gortat [2012–13]
		W. Johnson
		M. Morris [x2]

in Tables 2 and 3. We note that for the two selected seasons, the players Markieff Morris, Shannon Brown and Jared Dudley remain in the same cluster (2), which groups efficient players with the highest number of 3 points baskets scored. Cluster 0 is the one with best performances. The player Marcin Gortat whose 2011–12 season statistics allowed him to be in cluster 0, has shown a performance decrease during 2012–2013 and moved to cluster 2. His only improvement is seen in the three points baskets scored, which is a characteristic of cluster 2.

5.2 Classification

To test the classification, we used, first a clustering of 2012–13 season statistics. Some players have been excluded from this clustering to be used as test sample in this section. These players are: Isaiah Thomas, Klay Thompson and John Wall. The raison of choosing these three player is a paper in Bleacherreport.com published in June 2012, intituled "*15 NBA players who will be the stars of 2015*" and in which the three players are included. We wanted to check if with our classification, we would obtain the same prediction. For this purpose a Naive Bayes algorithm is used on the statistics of the three players in 2014–2015 season, the results confirmed Bleacherreport.com ranking. Indeed, Klay Thompson and John Wall were classified in the ≪cluster 1≫ containing NBA superstars (See Table 4). On the other hand, Isaiah Thomas, is assigned to cluster 5 for the same season: a cluster grouping players with a great number of marked points and very good defensive statistics. However, it should be noted that during the 2013–14 season, his excellent statistics affected him to cluster 1. The site

Bleacherreport.com also quoted some players as being the worst in the NBA, among them, Brian Cook. He averaged 5.7 points and 2.7 rebounds in less than 14 min per game and spent a lot of time on the bench during his career. The last season for Cook is 2011–12; so we took the statistics of this last season to predict to which cluster he will be assigned. The algorithm classified him in cluster2. A very natural result because Cook spent only 276 min on the playground which generated very weak statistics in defence as in offense (Table 6). The average salary of Cook is 1,955,569; an average salary that matches perfectly to the cluster to which he belongs, as the average salary of cluster 2 is 1,774,548 (Table 5).

Table 4. Thompson and Wall statistics in cluster1.

Clusters	MP	FG	FGA	3P	3PA	2P	2PA	FT	FTA	ORB	DRB	AST	STL	BLK	TOV	PTS	PER
K. Thompson	2455	602	1299	239	545	363	754	225	256	27	220	87	60	149	122	1668	?
J. Wall	2837	519	1166	65	217	454	949	284	362	36	330	792	138	45	304	1387	?

Table 5. Isaiah Thomas statistics in cluster 5.

Cluster	MP	FG	FGA	3P	3PA	2P	2PA	FT	FTA	ORB	DRB	AST	STL	BLK	TOV	PTS	PER
I.Thomas13-14	1726	335	797	129	346	206	451	302	348	33	120	284	57	5	143	1101	22.3
I.Thomas14-15	2497	496	1096	127	364	369	732	346	406	47	163	545	93	8	213	1465	20.5

Table 6. Isaiah Thomas statistics in cluster 2.

Cluster	MP	FG	FGA	3P	3PA	2P	2PA	FT	FTA	ORB	DRB	AST	STL	BLK	TOV	PTS	PER
Brian Cook	276	31	98	10	50	21	48	9	10	10	53	10	6	5	11	81	10.4

5.3 Performance Prediction

We have selected players who have played several seasons and apply predictions on their statistics.

1- To predict Performance of the players, we used the following multiple regression model:

$$\begin{aligned} PER = {} & 21.436 - 0.0426G - 0.0068MP + 0.0419a_3FG - 0.0118FGA \\ & + 0.0215FT - 0.00467FTA + 0.0183ORB + 0.00433DRB \\ & + 0.0135AST + 0.0222STL + 0.0171BLK - 0.0196TOV - 0.0113PF. \end{aligned}$$

Multiple linear regression tests conducted on R software showed that all variables are statistically significant with p-value under $\alpha = 5\%$.

Moreover, we also used previous performance in different seasons to predict players' performance in a given season by Holt Winters algorithm (exponential smoothing algorithm).

2- In the following example, we chose to predict the performance of Darius Miller who played eight NBA seasons. We used statistics of the first four seasons as learning data and the rest to test a prediction to compare the predict values to the real ones. The results of prediction of PER are displayed in the Table 7 and represented in Fig. 2. Figure 2 shows that the prediction obtained across the season by the exponential smoothing are closer than the regression model to the true values. Note that, prediction value by regression are obtained by firstly predict by Holt Winters method the values of indicators that affect PER, than compute this latter.

Table 7. Results of performances prediction.

Seasons	2006–07	2007–08	2008–09	2009–10	2010–11	2011–12	2012–13	2013–14
	Learning values				Predicted values			
True values	14.2	11.8	10.8	14.4	12.4	15.3	16	14.1
Pred values: reg	14.2	11.8	10.8	14.4	16.8	18.49	20.79	23.09
Pred values: exp smooth	14.2	11.8	10.8	14.4	15.57	14.69	16.85	12.99

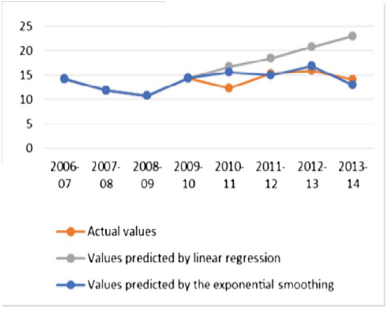

Fig. 2. Comparison between linear regression and exponential smoothing in performance prediction.

3- In other example, we tested performances of the NBA player, Michael Jordan. We focused on prediction of its interception and the number of points scored using simple exponential smoothing. We recall that Michael Jordan is the first player in NBA history who has scored 200 interceptions in a season. The results are displayed in Figs. 3 and 4.

From this two figures we see clearly that the obtained predicted value matches the true values of interceptions or Fields goal.

4- The prediction of players' salaries was also tested using multiple regression model and significant explanatory variables are selected. These variables are also been predicted using exponential smoothing since they occurred each seasons.

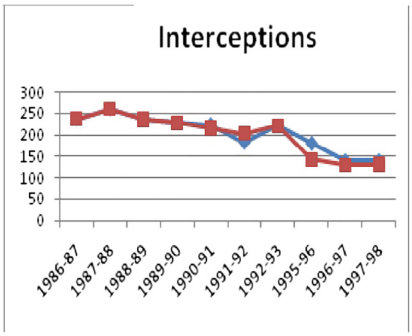

Fig. 3. Predicted and true values comparison of Michael Jordan interceptions across the seasons.

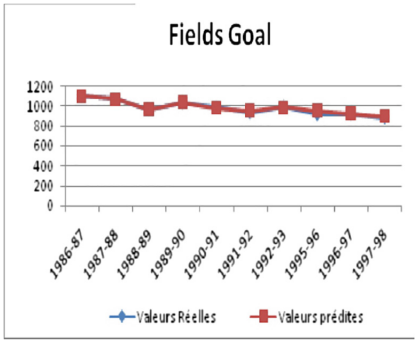

Fig. 4. Predicted Fields goal across the seasons.

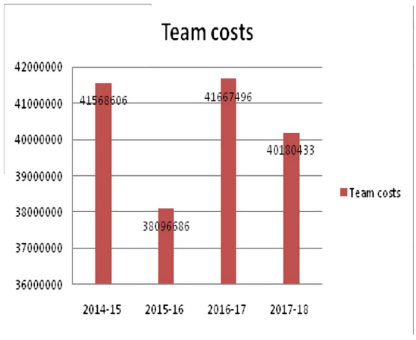

Fig. 5. Histogram of predicted cost of Phoenix Suns team.

Hence the salaries of the players of Phoenix Suns team are predicted using the following estimated regression model:

$$Sal = 15122431 - 7099MP + 38047FG - 11575FGA + 18615FT - 14558FTA + 22853DRB + 9222AST - 21840STL - 28384PF.$$

According of the predicted salaries of the players of Phoenix Suns team, we obtained the predicted team franchise costs as shown in Fig. 5.

6 Conclusions

In this work, we aim to assist NBA sports decision makers in the process of acquisition of players by exploiting and applying the data mining techniques as k_means, Naive Bayes algorithm, linear regression and time series on large amounts of data. These data are the statistics of ten seasons in the NBA (from 2005–06 to 2014–15). The results we have reached, allow decision makers to identify their needs, determine the players who respond to this need and to select the one or ones that work best for them depending on their current and future performance, taking into account their cost.

References

1. Ayer, R.: Big 2s and big 3s: analyzing how a team's best players complement each other. In: MIT Sloan Sports Analytics Conference, Boston, MA, USA (2012)
2. Beckler, M., Hongfei, W., Papamichael, M.: NBA oracle. Zuletzt besucht am 17, 2008–2009 (2013)
3. Cao, C.: Sports data mining technology used in basketball outcome prediction. Dissertation, Dublin Institute of Technology (2012)
4. Oliver, D.: Basket Ball on a Paper. Rules and Tools for Performance Analysis. Potomac Books Inc., Lincoln (2004). 392 pages
5. Casals, M., Martinez, J.A.: Modelling player performance in basketball through mixed models. Int. J. Perform. Anal. Sport **13**, 64–82 (2013)
6. Hoffman, L., Joseph, M.: A multivariate statistical analysis of the NBA (2003). http://www.units.miamioh.edu/sumsri/sumj/2003/NBAstats.pdf
7. Hollinger, J.: Pro Basketball Forecast: Paperback, 1900 (2005)
8. Li, H.: True value in the NBA: an analysis of on-court performance and its effects on revenues. Undergraduate Honor Thesis, University of California, Berkeley (2011)
9. Maheswaran, R., Chang, Y.-H., Henehan, A., Danesis, S.: Deconstructing the rebound with optical tracking data. In: MIT Sloan Sports Analytics Conference 2012, Boston, MA, USA (2012)
10. Mitchell, T.M.: Machine Learning, Chap. 6. McGraw-Hill Science, New York (1997)
11. Wheeler, K.: Predicting NBA player performance (2012). cs229.stanford.edu/proj2012/Wheeler-PredictingNBA