Machine Learning Techniques with Reduced Dimension NBA Data

Adam Williams

December 12, 2017

Abstract

This paper presents the results of utilizing predictive algorithms with reduced dimensionality NBA player data. Advanced statistical analysis has become a dominant trend in sports, and basketball is no exception. Principle component analysis, support vector machines, and K-means clustering are all applied to a dataset drawn from the aggregate season statistics for the 2013-2014 NBA season. The first section will provide background on common NBA statistical analysis. The following section will describe the manner in which the data was organized and the algorithms were applied, while the third section will present the results. The final section then provides conclusions to the work and possible future approaches.

1 Background

Bill James, considered the father of sabermetrics and thus the advanced sports analytic movement, began his first works on the statistical analysis of baseball while working night shifts as a security guard at a pork and beans cannery. His insight into the statistics beyond the well known ”counting stats” eventually led to the adoption and integration of sizable advanced analytics departments for every major baseball team. Though baseball lends obviously to the use f advanced statistics, with its defined plays and clearly defined situational scenarios, basketball has proven to be an equally rich statistical field. The advent of cutting edge computer vision techniques has led to the installation of player tracking cameras in every arena in the NBA, and several companies have been founded that purport to provide the advanced in-depth analysis of the vast amounts of data collected [1]. The amount of data and the interconnectedness of basketball's statistics provide and excellent opportunity for the use of machine learning for analysis. There has been some previous work on the prediction of the winner of games as well as on the use of intelligent methods to find situational NBA statistics [6],[7]. Using ANN's, head to head win predictions have had success rates as high as 64%.

The initial proposal for this work detailed three aims, reproduced in the list below.

Proposed Aims:

1. **Reduce dimensionality of player statistics utilizing Principle Component Analysis (PCA).** [2] PCA can be applied to the data in order to compress the n-dimensional statistics pertaining to a player's per game offensive, defensive, or overall contribution to their team's success to a k-dimensional representation, where $k < n$.

2. **Build a Support Vector Machine(SVM) that takes the reduced dimensional player statistics to predict the winner of head to head matchups.** As a maximum margin, nonlinear classifier an SVM is an excellent option to use as a predictive tool [3].

3. **Sort player contribution to team success using reduced dimension statistics against team wins, in order to identity contributions of players with ”intangible” skills.** Implementing a k-means clustering algorithm to group players by performance and team success will identify players who help their team in ways that do not show in team box scores [4].

Summarized, the first is to apply principle component analysis in order to derive a reduced order statistic from a range of 21 statistical categories commonly recorded for every player. The second is to use these reduced order statistics to build a support vector machine that can be used to predict the winner in games between two teams. Finally, the last aim is to use K-means clustering in order to identify players whose contributions to winning teams are not shown in their statistical contributions.

In basketball, there have been a number of attempts to create an ”overall” statistic that quantifies a players contribution on the court. Some examples include a Win-shares system similar to that used in baseball that attempts to quantify the number of extra wins a player has contributed to the team compared to the league average, real plus-minus(RPM) which is a statistic that measures a league adjusted positive or negative value for how many points a player's presence on the court contributes, and player efficiency rating(PER), an advanced statistic developed by John Hollinger, an ESPN statistician. PER adjusts a players contribution for league averages as well as weighting the various statistics in order to represent their impact on the game [5]. Of the three, PER is the one that is most commonly referred to in sports programs. It will be used as a metric for the PCA decomposition, in order to demonstrate its validity as well as its effectiveness.

2 Research Design and Datasets

The machine learning algorithms used in this paper were implemented in a Python 3.6 environment, utilizing the scikit-learn package [8]. The NBA statistical data was collected via a Python web scraper collecting the total season statistics for the NBA seasons 2013-2014 and 2014-2015 from basketball-reference.com [9]. While previously it was proposed to aggregate box scores, implementation and organizational challenges to implementing randomly sampled iterations from a dynamically aggregating database necessitated the use of season-long statistics.

3 Implementation and Results

3.1 Aim 1

The standard 21 counting stats for an NBA player are shown in the table below. Key abbreviations include FG for field goal made, attempted and percentage. A field goal is any shot taken on the court. The abbreviation 3P stands for 3-point shot, while 2P stands for 2-point shot and FT is free throws. In addition, RB means rebound, either offensive, defensive, or total. Finally, AST represents assist, STL is steals, BLK is blocks, TOV is turnovers, PF is personal fouls committed, and PT is points.

It is easily seen that these statistics are highly implicit, with the shot percentages being combinations of attempts and makes or the points scored being a linear combination of the types of shots made. Thus these values are an excellent target for dimensional reduction through PCA, with low loss of variability. Rather than purposefully divide between defensive and offensive stats as previously stated, PCA was instead run twice. The first run combined the season statistics

FGM	FGA	FG %	3PM
3PA	3P%	2PM	2PA
2P %	eFG %	FTM	FTA
FT %	ORB	DRB	TRB
AST	STL	BLK	TOV
PF	PTS		

Table 1: NBA Counting Stats

from 2013-2014 into a two dimensional variable space. The results, labeled with the names of the corresponding players, can be seen in Fig. 1.

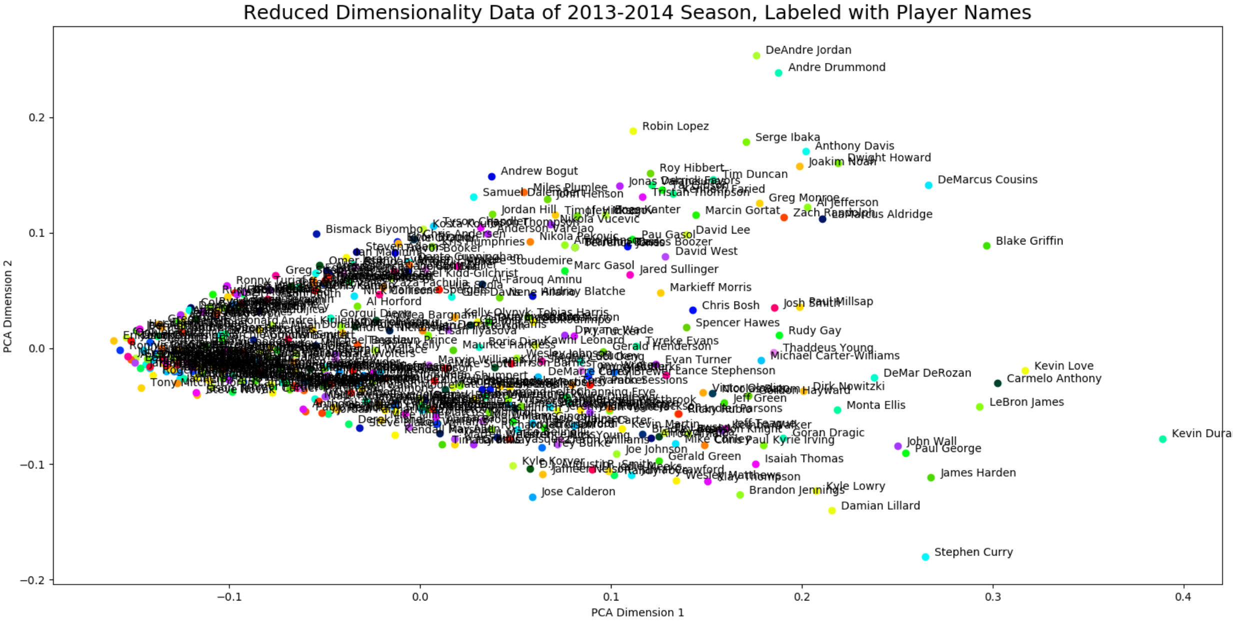

Figure 1: Two Dimensional Reduced Order NBA Player Data

While the sheer number of labeled player names makes this data somewhat difficult to interpret, the outliers are instructional as to the meaning of the reduced dimensional data. For instance, in the positive y-direction, DeAndre Jordan and Andre Drummond are at the extreme. Both players are specialists who put up huge numbers in the rebounding and shot blocking department, while shooting a high 2P percentage but poorly from 3 and at the free throw line. Moving counterclockwise, the positive x-direction leads to players such as Kevin Love and Blake Griffen. These are big men who both rebound and shoot well, while not necessarily providing the defensive impact the previous players do. Continuing, the lower right quadrant includes perennial All-Stars and MVP contenders such as Lebron James, Kevin Durant, and Carmelo Anthony. Finally, the furthest outlier in the negative y-direction is Steph Curry. who in 2013-2014 was beginning to perfect the otherworldly 3-point accuracy that fueled his two MVP awards in 2014-2015 and 2015-2016, and additionally contributing high steal numbers but few blocks or rebounds.

To check the validity of using reduced dimensionality data as a metric for player performance, the second iteration of PCA was run, this time decomposing the data to a single dimension. This single dimension value was then plotted against the player's recorded PER for the 2013-2014 season.

The Pearson's correlation coefficient between the two statistical measures was calculated and found to be 0.735, indicating a strong correlation between the two. Differences most likely arise from two sources: the first being the calibration of PER to represent 15 as a league average value, the second being the impartiality of the PCA decomposition to "negative" stats such as turnovers or low FG percentage. This lack of intended negative bias may cause such statistics to be represented in a different portion of the variable space than might be considered normal for something that is labeled as negative.

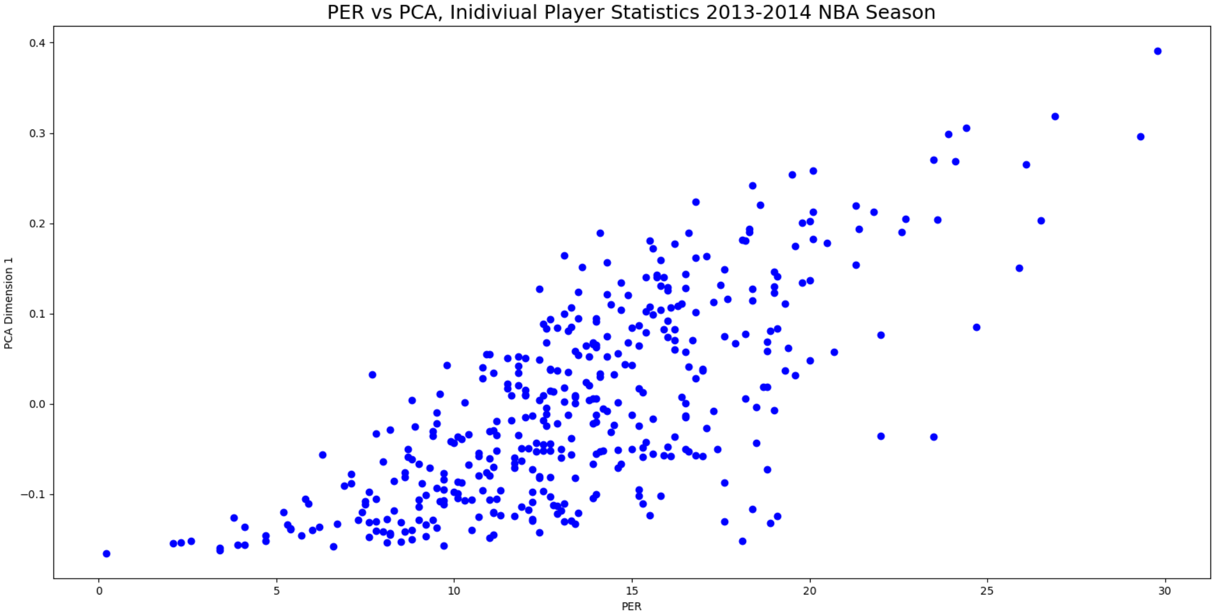

Figure 2: Relationship between PER and Single Dimensionality NBA Data

3.2 Aim 2

The use of season-aggregated data helps to simplify the organization hurdles of randomly sampling a database that varies in time due to the progression of an NBA season, so as not to include data that may have occurred in the very game that is being predicted. However, this then gives rise to a second problem, namely how to utilize a prior seasons data in an SVM model. Rather than utilize data collected from the very season it was collected from, and thus developing an implicit and flawed model, the player data was instead attributed to their corresponding team in the 2014-2015 season. In order to provide feature vectors of uniform size, the 2 dimensional player score found in the previous aim was summed for the players on each team, to take into account variations in roster size.

The feature vectors were organized such that the home team was the first feature set, then the away team. If the home team won, in the training data this was classified as a 1, while home team losses were classified as a 0. The resultant dataset thus contained n = 1230 head-to-head matchups, each containing a 4 -dimensional feature vector of the team's total PCA Dimension 1 and Dimension 2 values as well as a binary classifier indicating a home win or not. In order to train the SVM, 10-fold cross validation was utilized. In each iteration, the test data was randomly sampled from the duration of the season. The remaining matchups were then used as the cross validatoin set. In order to account for nonlinearities in relative team capabilities as well as provide a smooth classification boundary, a Gaussian RBF kernel was utilized. The radial basis function takes the form:

$$K(x_i, x_j) = e^{\gamma||x_i - x_j||^2} \qquad \text{for} \quad \gamma > 0$$

While the SVM was trained utilizing a 4-dimensional feature vector, a 2-D slice of the vector space was taken and plotted to illustrate a portion of the decision boundary, shown in Fig. 3 The data is depicted such that wins are represented in red, while losses are represented in blue. The cross validation data points follow the same color scheme, however these points are indicated by a white outline to each point. The decision boundary for the predicted win and loss section are depicted by the red and blue background, respectively, and the decision boundaries are shown in black.

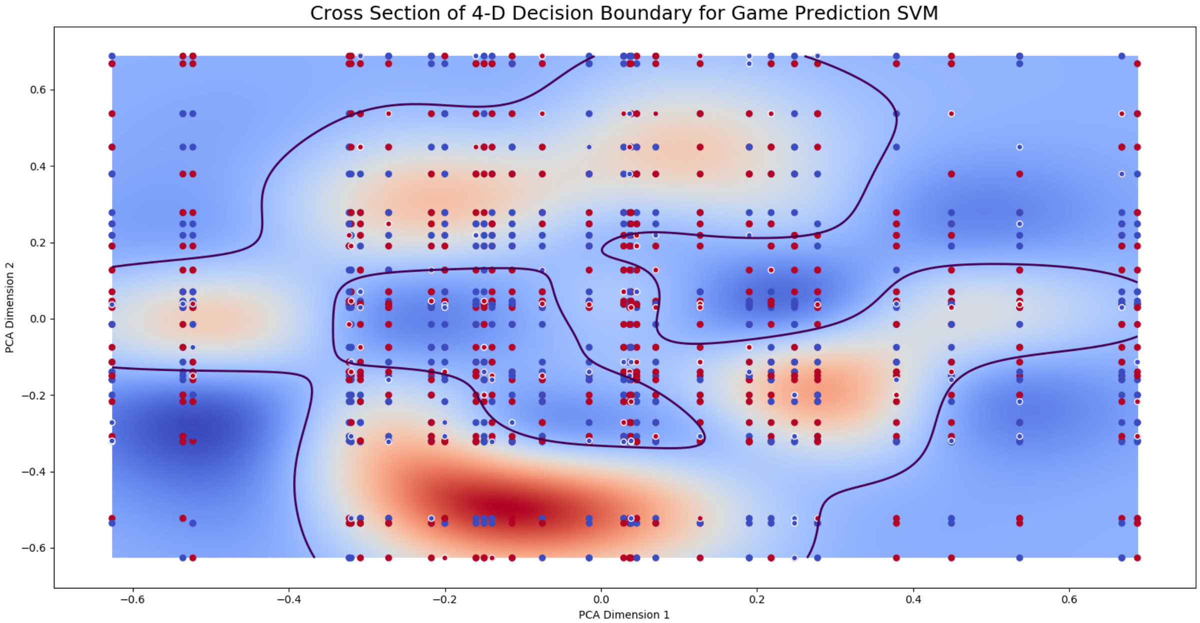

Figure 3: SVM Decision Boundary Visualization

Unfortunately, but not unsurprisingly, the use of data from player performance is not especially useful as a predictive tool, beyond a percentage point or two. This is understandable, as this implementation fails to take into account player aging or performance degradation due to injury, team momentum(winning or losing streaks), or coaching changes. Additionally, it does not account for the player fatigue, a hot topic in the NBA, as the league is notorious for schedule games on back to back games. Research has shown that teams playing the second game of a back to back tend to lose more often, especially when it comes at the end of an extended sequence of away games [10].

n = 123	Win	Loss	
Predicted Win	32.33	34.28	66.61
Predicted Loss	29.66	26.4	56.06
	61.99	60.32	

Table 2: Confusion Table for Predictive SVM

The predictive ability of the SVM is shown in in the confusion matrix presented in Table 2, collected by averaging the confusion matrices of 100 randomly initialized training iterations.

3.3 Aim 3

Upon inspection of the initial 2-dimensional PCA data, the clear delineations between different ”classes” of players in the variable decomposition lent itself very well to the use of unsupervised clustering algorithms. Choosing the number of centroids to be 15, K-means clustering was run with 20 random initializations in order to provide a robust grouping. The results can be seen in Fig. 4. In 4 the decision boundaries for the various clusters are illustrated in color with the corresponding data points shown as black boxes within, and the centroids represented as white X’s.

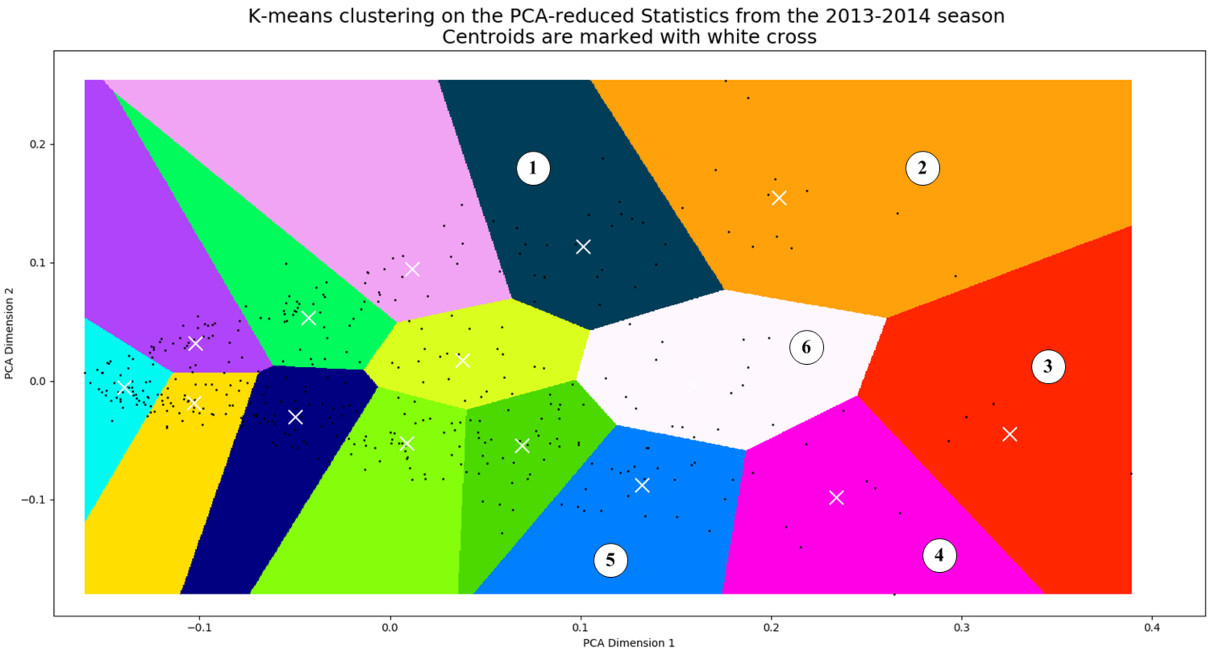

Figure 4: K-means Statistical Grouping

In the interest of brevity, only a selection of six of the more interesting player classifications will be discussed. Those groups are labeled 1-6 in 4. Group 1 represents average level starting forwards and centers such as Andrew Bogut, Nikola Vucevic, and Miles Plumlee. These players are decent rebounders and shot blockers, but do not stellar scorers from close or long range. Group 2 contains the top forward and centers in the league, including top rebounders such as Andre Drummond, Demarcus Cousins and Dwight Howard, as well as excellent shot blockers such as DeAndre Jordan and Anthony Davis. These players are considered the best at their positions. The third group is an extremely intriguing one, as it contains perennial MVP candidates such as Kevin Durant and Lebron James. In fact, Kevin Durant is the farthest outlier at the far positive x edge of the decision boundary, and in fact won the NBA Most Valuable Player award in the analyzed season. In group 4, guards are represented who excel from the 3-point line as well as dish out high numbers of assists, such as Steph Curry and James Harden. The fifth group contains the average level guards, such as Jameer Nelson, Kyle Korver, and Jose Calderon. Finally, group 5 includes versatile ”positionless” players, those considered to be between the traditional guard/forward distinction, such as Rudy Gay, Evan Turner, and Tyreke Evans.

The originally stated aim was to utilize the statistical results of the PCA decomposition in concert with team win totals and minutes played to find players who contribute to winning teams while not necessarily putting up high visible statistical contributions. This type of analysis is a highly desirable line of inquiry, as the scouting departments of many NBA teams spend large amounts of time attempting to uncover these distinctions. In an attempt to identify such players, the 2-dimensional PCA data was plotted with an Win-Minutes metric for each player as the z-axis. The goal was to capture the contributions of a player whose minutes successfully translate to wins for their team. K-means clustering was then performed on the resultant dataset with k=12 and 20 initializations, the results of which are illustrated below in Fig. 5.

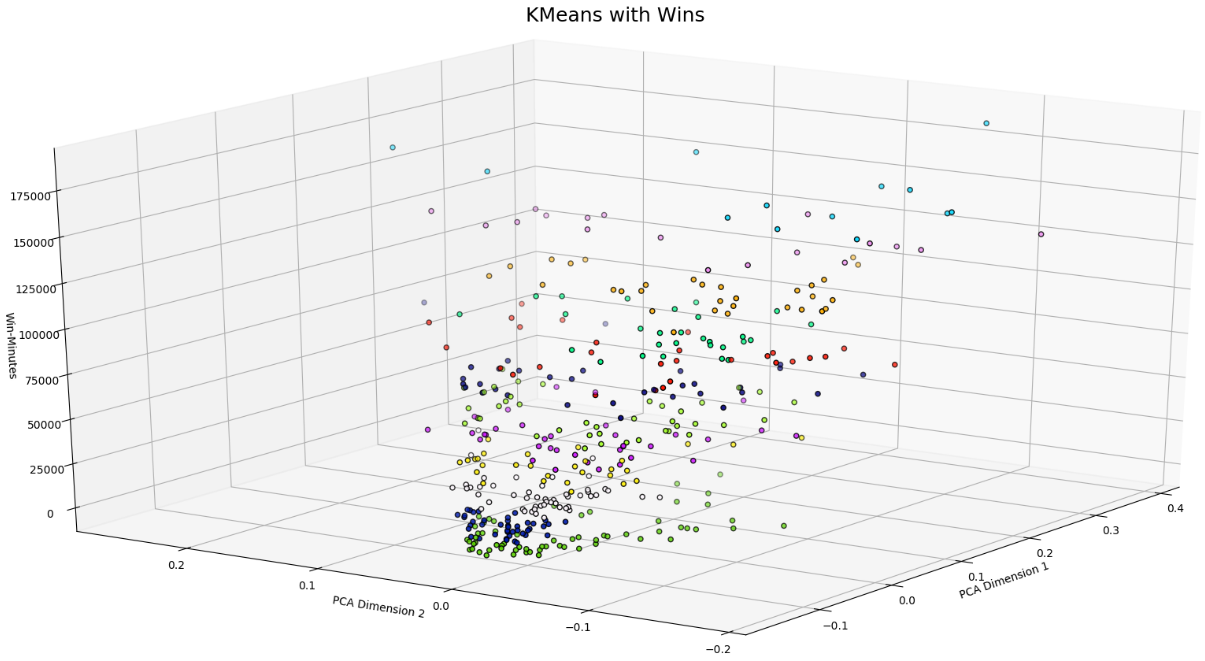

Figure 5: K-Means Clustering for Reduced Dimensionality Data vs. Win-Minutes

Unsurprisingly, several players with high statistical contributions were immediately visible in the high Win-Minutes contribution groupings(teal with a dark outline), such as Lebron James and Kevin Durant. However, within that group of high Win-Minutes there are also several players that at the time were less recognized for their contributions, including Chandler Parsons, Lance Stephenson, Nic Batum, and Wesley Matthews. Tellingly, all four players received large raises in the two years following the charted season in the form of new contracts during free agency with new teams. If such contributions had been caught slightly earlier by each player's original team, perhaps the players could have been retained for less money than each got in free agency.

4 Conclusion

In summary, machine learning can provide valuable inferences on NBA statistical data. While the ability of previous season was proven to be an unsuccessful metric for the prediction of games, studies have found that rather than individual player statistics, the most important predictors for winning include team Win%, offensive efficiency, avg. fouls and avg. steals [11]

References

[1] STATS LLC. *STATS SportVu*. URL: https://www.stats.com/basketball/ (visited on 2017).

[2] Svante Wold, Kim Esbensen, and Paul Geladi. "Principal component analysis". In: *Chemometrics and Intelligent Laboratory Systems* 2.1-3 (Aug. 1987), pp. 37–52. ISSN: 01697439. DOI: 10.1016/0169-7439(87)80084-9. arXiv: 1011.1669.

[3] Corinna Cortes and Vladimir Vapnik. "Support-Vector Networks". In: *Machine Learning* 20.3 (1995), pp. 273–297. ISSN: 15730565. DOI: 10.1023/A:1022627411411. arXiv: 1011.1669v3.

[4] Source Journal et al. "Algorithm AS 136 : A K-Means Clustering Algorithm Author (s): J . A . Hartigan and M . A . Wong Published by : Wiley for the Royal Statistical Society Stable URL : http://www.jstor.org/stable/2346830". In: 28.1 (2017), pp. 100–108. ISSN: 10510761. DOI: 10.1890/11-0206.1.

[5] Basketball Reference.com. *Calculating PER*. URL: https://www.basketball-reference.com/about/per.html (visited on 2017).

[6] Afroza Sultana et al. "Incremental discovery of prominent situational facts". In: *2014 IEEE 30th International Conference on Data Engineering*. IEEE, Mar. 2014, pp. 112–123. ISBN: 978-1-4799-2555-1. DOI: 10.1109/ICDE.2014.6816644. arXiv: arXiv:1311.4529v2. URL: http://ieeexplore.ieee.org/document/6816644/.

[7] Albrecht Zimmermann, Sruthi Moorthy, and Zifan Shi. "Predicting college basketball match outcomes using machine learning techniques: some results and lessons learned". In: (Oct. 2013). arXiv: 1310.3607. URL: http://arxiv.org/abs/1310.3607.

[8] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[9] Basketball-Reference.com. *Basketball-Reference.com*. URL: https://www.basketball-reference.com/ (visited on 10/15/2017).

[10] Baxter Holmes. *The NBA schedule turns teams into the sleepwalking dead*. URL: http://www.espn.com/nba/story/_/id/17790282/the-nba-grueling-schedule-cause-loss.

[11] Kazimierz Mikolajec, Adam Maszczyk, and Tomasz Zajac. "Game Indicators Determining Sports Performance in the NBA". In: *J Hum Kinet* 37 (July 2013). 24146715[pmid], pp. 145–151. ISSN: 1640-5544. DOI: 10.2478/hukin-2013-0035. URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3796832/.