The National Basketball Association (NBA) Prediction*

Author: Sanele Rudolph Mkhize[1]

School of Mathematics, Statistic, and Computer Science, University of Kwazulu-Natal Private Bag Box X54001, Durban 4000, South Africa
215018500@stu.ukzn.ac.za
https://ukzn.ac.za/

Abstract. Start

Keywords: Basketball · Machine Learning · Classification · Classification · Data mining · Features selection · Machine learning · NBA

1 Introduction

Basketball is one of the most popular sports in the United States and is also popular internationally. Early on, it was a simple gym exercise and slowly spread to high schools and colleges. The National Basketball Association (NBA) emerged in the early 1950s as the major governing body for the professional version of the game. Every year, the NBA hosts a tournament that brings together teams from all over North America. The tournament is known as a season. In a season, each team competes against others several times and records its wins and losses. In an elimination-style tournament called the Playoffs, the 16 teams with the most wins compete against each other after a number of games (in the hundreds). Four out of seven playoff matches are very competitive. Whoever wins the Playoffs wins the whole season.

This project is centred around determining outstanding players, detecting outliers of players, given two teams, determining who will win solely based on statistical analysis. In this project, we aim to help people who often play fantasy basketball, a virtual game in which players create teams based on real-life players and get rankings depending on the performance of their selected players in games.

1.1 Problem Statement

As stipulated in the introduction, the main aim of this project is to be able to predict game results given two teams, detect outlier detection in our data, and determine outstanding players.

Game Outcome Prediction: being able to predict accurately a winning team

* University of KwaZulu-Natal.

to a certain degree for a team can prove instrumental to companies usually betting companies in setting the odds for teams playing, in order to improve their profit margins. This can also help teams in assessing their performance against rival teams at least statistically.

Outstanding players: Being able to identify outstanding players or at least players who have a positive trajectory towards reaching that feet can be an instrumental tool for teams when doing player transfer as it is common in NBA, or buying new players in the market. Creating a model that can make a prediction or classify if a player is an outstanding player can be beneficial, or at least a first step in making a decision what is important to note is that machine learning models are not a defining tool to decision process but can help in getting a clear direction.

Player outlier detection: This is also an instrumental parameter in terms of making an informed decision about a player. This is very helpful in gauging a player two extremes between good and bad players.

2 Related Work

Even though machine learning has been used in different sports analytics, it still remains necessary to improve the performance of the models this intelligent technology can generate. The following section describes how machine learning can be used to predict basketball game results and previous research studies.

A study of artificial neural networks for predicting NBA game results (ANN) was developed by Loeffelholz et al. Some features were compiled taking the unprocessed data and used it as input for neural networks. Afterwards, various basketball sports experts were given an opportunity to make predictions in order to compare their decisions with the outcomes assigned by the ANN model. The ANN offered higher predictive decisions than the domain experts, i.e., 74.33%. The regression method was constructed based on a historic spread point to develop a method to forecast the spread point of the NFL (National Football League). This kind of research can be used to predict the game results. The NFL winners can be predicted by using the forecast method of the spread point [5].

Data from historical NBA games was used to create a predictive model based on Naïve Bayes Classifier by Cao. Approximately 65.82% of the test dataset was correctly classified [2]

3 Methods and Techniques

3.1 Data Preprocessing

Our work on the National Basketball Association (NBA) Prediction consisted of the following task, studying extensively the data that was provided in our University of KwaZulu-Natal learning site portal, the data was obtained in a form of a zip file which was also obtained from the basketball reference website. The dataset that was obtained spans from the 1970 to 2004 season the main reason why we truncated the data is that steals, blocks, and turnovers were not official NBA stats until the 1970s this was done to ensure that the is consistency in our model, remembering that our model depends on the garbage in and garbage out meaning that our model will be as good as the data that drives it.

3.2 Game Outcome Prediction

The idea implement game outcome prediction into high accuracy has a lot of implications in this day and age of companies especially the betting companies such as Betway, Hollywood bets, etc. Being able to predict a game outcome to a certain level of confidence is quite useful in certain odds for payouts, or if it is worth adding a term in the ticket. In implementing this kind of task we employed on using Linear regression.

Linear regression: Regularly, linear regression is the first classifier prefaced in machine learning. With least-squares analysis, one or more independent variables (input features) are modeled against another variable, the dependent (output). Linear weights are assigned to the input features as regression coefficients, which weight the input features linearly. Linearity does not mean that the classification boundary is a hyperplane, but rather, it means that the regression coefficients are linear in degree. Te following is an equation for linear regression:

$$Y = \sum_{x=1}^{n} w_i * x_i + b \tag{1}$$

What you need to note in the above is that i refers to the number of attribute in our vector representation, and b is just the intercept. To get a clear representation of what our model determined the values to be please find the following table: For our specific research project, we have a lot of features used to create our model, and these are cumulatively added to create a team's statistics and it contains both offensive and defensive statistics, the following is just a snippets of some of the features used all in all the about 30 features used from our data. We also used win to loss ratio as our target value.

- **o_fgm and d_fgm** *(Field of goals made)*: a field goal is a basket scored on any shot or tap other than a free throw, worth two or three points depending on the distance of the attempt from the basket [1].

Attributes values	Coefficient
o_fgm	-1.81982791e-01
o_fga	-2.50692817e-04
o_ftm	-9.17571855e-02
o_fta	5.10110362e-04
o_oreb	2.23839045e-03
o_dreb	3.12538103e-03
o_reb	-2.13574822e-03
o_asts	1.03818978e-03
o_pf	-3.66203820e-03
o_stl	5.29162338e-03
o_to	-1.82197098e-03
o_blk	2.36635999e-03
o_3pm	-8.65176007e-02
o_3pa	-6.91378765e-04
o_pts	1.26073997e-01
d_fgm	-6.32086824e-01
d_fga	5.49391224e-03
d_ftm	-3.20829173e-01
d_fta	6.19984023e-03
d_oreb	9.32646499e-04
d_dreb	-2.62519252e-03
d_reb	-2.02820330e-03
d_asts	-1.53541888e-03
d_pf	4.03530523e-03
d_stl	-5.47918926e-03
d_to	6.21048468e-04
d_blk	-8.97843185e-03
d_3pm	-3.16575878e-01
d_3pa	4.85875282e-04
d_pts	2.78317438e-01
intercepts	43.53436867092078

Table 1. Linear regression coefficients

- **o_fga and d_fga** *(Field of goals attempted)*: This is just the field of goal attempted, both the successful and unsuccessfully.
- **o_ftm and d_ftm** *(Free through made)*: This is just the 1 point goal made in the basket, usually awarded as result of what is called a technical foal in basketball.
- **o_fta and d_fta** *(Field of goals assist)*: This is just the pass that is made and leads to a team member scoring.
- **o_oreb and d_oreb** *(Offensive rebound and defensive)*: a rebound, sometimes colloquially referred to as a board, is a statistic awarded to a player who retrieves the ball after a missed field goal or free throw [3].

- **o_stl and d_stl** *(Offensive and defensive steals)*: a steal occurs when a defensive player legally causes a turnover by their positive, aggressive action [4]
- **win to loss ration**: Calculate the chances or percent that a term will win in general. Win to loss ratio is computed as follows:

$$win\_to\_loss\_ratio = \frac{win}{loss + win} \tag{2}$$

3.3 Outstanding players and Player outlier detection

The purpose of this subsection of our research is very vital in determining the outliers in the given dataset. The idea is to be able to get extremes that are the lower bound as well as the upper bound in our dataset. This will give us an insight into the player quality that we have in the pool of players. One of the applications of creating a model that can solve such problems can be used by NBA teams to assess the player market, analyze a player to cost ratio when deciding salaries of such players and try to extrapolate the potential of such players, keeping in mind that such models can be used as a blueprint in making such decisions or at least the first step.

We will be using unsupervised learning to try and understand the underlying structure of our attributes, and observe the extremes in the dataset, one of the reasons why we opted to use unsupervised learning is that no current statistics can be used to accurately predict outstanding players, since that depends on a decade that a player was playing at, over the multiple decades that the sport of basketball has existed, the has been certain rules that has been changed, this also ensures that we don't compare apple and oranges. So the main idea of using this unsupervised learning is to try and find the clusters in our learning data to get some kind of group characteristics. To tackle this problem we will be using K means clustering which is a very popular algorithm typically we don't know what we are looking for but we identify clusters. K is usually an odd number, depending on what K is we place points in our dataset also referred to as centroids, and the next step is to identify the distance from the data points to the centroids, usually, we will like to adjust the centroids to maximize the clusters. This process is iteratively done by adjusting the centroids and calculating the distances from the centroids to the data points. Usually, the best k can be computed by using what is called an elbow technique.

Usually, the performance of individual players depends on the occasion logical players that are playing in the basketball playoffs will try and play at their maximum best since this increases their chances of getting an NBA championship which is the most prized trophy in basketball usually players that have won an NBA championship have bragging rights, while during the regular season the are many games to play teams usually don't play at their maximum and "star" players tend to be not as active as much as the regular season leads to the west

and east final.

For clustering purpose we used a total of 17 features to describe each instance, which is just the career statics for each players. The following features we used for clear understanding:

- gp (games played): Which is the total amount of games played.
- minutes (minutes played): Which is the number of minute played in the basketball field.
- pts (points) : Which is the career points scored.
- dreb (defensive rebound) and oreb (offensive rebound): Which is the defensive rebounds and offensive rebounds.
- reb (rebound) : Is just the sum of the dreb and oreb.
- asts (Assist to turnover ration)
- stl (steals): The number of steals by a defensive player or team
- blk (blocks): The number of blocks by a defensive player or team
- turnover: turnover occurs when a team loses possession of the ball to the opposing team before a player takes a shot at their team's basket. This can result from a player getting the ball stolen, stepping out of bounds, having a pass intercepted, committing a violation (such as double dribble, traveling, shot clock violation, three-second violation or five-second violation), or committing an offensive foul (including personal, flagrant, and technical fouls).
- pf
- fga: the total number of field goals attempted, FGA = 2PtA + 3PtA
- fgm: the total number of field goals made, FGA = 2PtA + 3PtA
- fta: the total number of field goals assist, FGA = 2PtA + 3PtA

In order to visualize the data we have we projected the 16 dimensions into a 2 dimensions, using a method called Principle Component Analysis (PCA) is a method of dimensionality reduction. It has applications far beyond visualization, but it can also be applied here. It uses eigenvalues and eigenvectors to find new axes on which the data is most spread out. From these new axes, we can choose those with the most extreme spreading and project onto this plane, the following is a scatter plot that resulted:

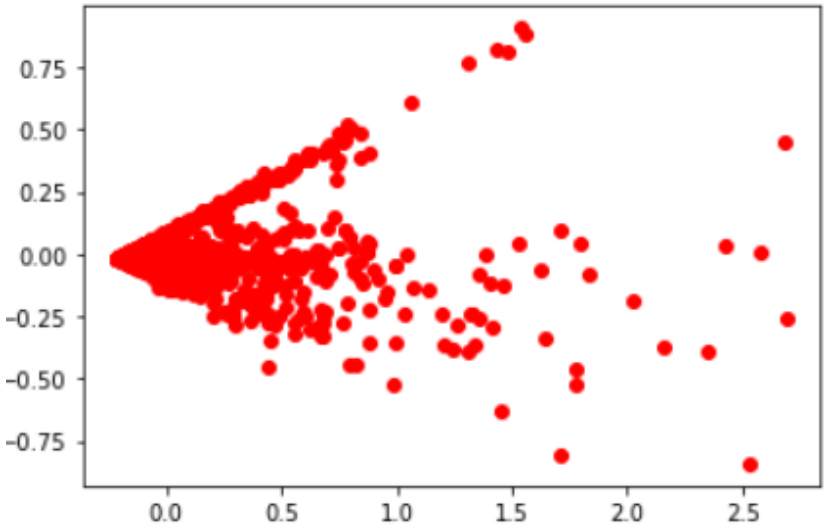


Fig. 1. Representation of the steps that are used in our model.

In our implementation we used k-nearest clustering to determine the outliers in our dataset, and the are marked with a red cycle, please find the following image that shows the three clusters:

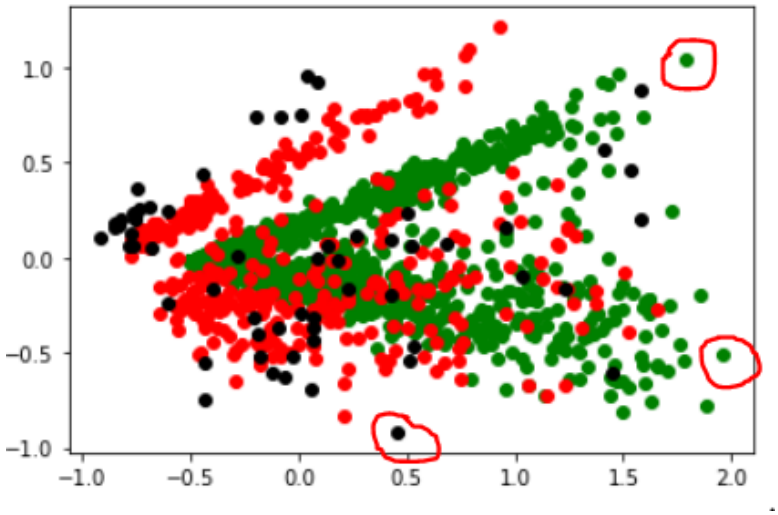


Fig. 2. Representation of the steps that are used in our model.

4 Results and Discussion

4.1 Winning team prediction model analysis

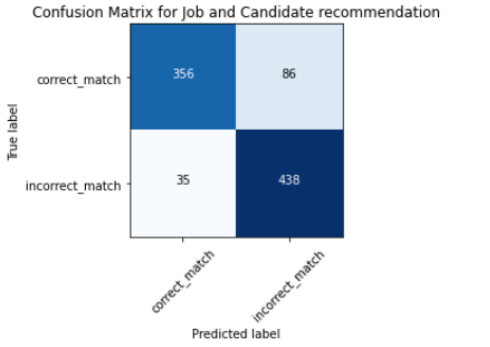
To test our predictive model (After using Linear regression) that should be able to determine the winner given two teams we used both the mean squared error and root mean squared error, using our ground truth win to loss ratio as the true y values and the predicted values from our model. We were able to get the following values respectively 15.518814431580513% and 3.9393926475512075%. The mean squared error is defined as follows:

$$rmse = \sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2} \tag{3}$$

The following is the mean square error formula that was used to test the effectiveness of our regression model:

$$mse = (\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2 \tag{4}$$

For the second method was implemented we used the support vector machines to determine if given two teams our model we will be able to detrmine the winning team. To test the accuracy of our model we used the confusion matrix. The number of instances that we tested are equal to 915 instances, and we correctly grouped 794 instances and incorrectly mis-classfied a total of 121 instances, the following is an image representation of our confusion matrix:



	precision	recall	f1-score	support
0.0	0.91	0.81	0.85	442
1.0	0.84	0.93	0.88	473
accuracy			0.87	915
macro avg	0.87	0.87	0.87	915
weighted avg	0.87	0.87	0.87	915

Fig. 3. Representation of the steps that are used in our model.

Support Vector Machines (SVM) rely on the training points closest to the boundary of the classification to build a classifier that maximizes margin. By ignoring the "easy" points that lie far from the decision boundary, SVM focuses on the points that really matter, namely the points that define the decision boundary.

4.2 Outstanding players and Player outlier detection model analysis

In order to measure the accuracy of our clustering model we can use what is called cross-tabulation but the issue with this clustering approach is that it requires that our players come pre-grouped into outstanding players and outlier players. So in order to validate the model we need a way to measure the quality of our clustering model with prior grouping, thus it needs to use only the samples and their cluster labels. So a good cluster has a tight cluster meaning that the sample is grouped together and not spread out, we can measure this by calculating the inertia, the idea is to measure the distance between the centroids of our cluster, so intuitively lower values of the inertia are better. Typically after training our model the is associated inertia. In our case the inertia was computed to be 803398841.4483262.

5 Conclusion

In the field of professional basketball, we used ML methods to predict game outcomes, identify outstanding players, and choose optimal player positions to

support betting, coaching, sponsorship, and other decisions. For game outcome prediction our best accuracy results was obtained by using the support vector machine in which the accuracy was determined to be 87% with the help of confusion matrix. But using clustering to determine the outliers wasn't quite fruit full due to the fact that the computed inertia was big.

References

1. Arseneault, P., Assaff, P.: Steve Nash. Heritage House Publishing Co (2006)
2. Cao, C.: Sports data mining technology used in basketball outcome prediction (2012)
3. Frazier, W., Sachare, A.: The Complete Idiot's Guide to Basketball. Penguin (2004)
4. Ivanković, Z., Racković, M., Markoski, B., Radosav, D., Ivković, M.: Appliance of neural networks in basketball scouting. Acta Polytechnica Hungarica **7**(4), 167–180 (2010)
5. Loeffelholz, B., Bednar, E., Bauer, K.W.: Predicting nba games using neural networks. Journal of Quantitative Analysis in Sports **5**(1) (2009)