Can machine-learning methods predict the outcome of an NBA game?

Nachi Marc Lieder

March 2018

Abstract

Game outcome prediction is a well known problem to predict , yet there are aspects, features and advantages that aren't taken into account in most modules. In this paper, we investigate various features and aspects of a prediction outcome for a team vs team , this includes basic known features such as home court advantage , number of days between games and winning streaks as well as less conventional features such as player similarity scores , gambling odds , team specifications. The analysis is based on descriptive statistics and while it is mainly restricted to the National Basketball Association , it can be projected towards other sports as well.

Introduction

The NBA (National Basketball Association) is the professional basketball league played in North America (USA + Canada) , and is probably considered to be the most prestigious basketball league in world .
The league is formed from 30 teams , split into 2 conferences. Each conference is split into 3 divisions spread equally.
The popularity of the NBA across the world is one of the highest among all sports.

Each team in a regular season plays 82 games , which totals 1230 games per season across the league. In addition there are the playoffs where the number of games is not defined. (maximum 28 , minimum 0)

The main objective of this research is to create a more accurate predictor of the outcome between 2 given teams at a given time, based on different machine learning methods and selected features.
In order to understand the added value of the predictor , one benchmark that will be used are the betting odds of each team to win , which will define which team is favorite to win according to the better odds.
Another benchmark that will be tested against is the strategy of picking the team with the current better record as the winner of the given game.

This paper will prove that the combination of conventional and non-conventional features together will supply a much better predictor.

The Data

The data that was used in order to complete this research comes from several main point of views:
The players aspect - personal info of box scores per game, which contain the data such as points , rebounds , assists , minutes played , ect.
Team box scores - aggregated team stats by quarters.
Team vs. Player - team shooting distributions with/without every player on the floor.
Personal attributes per player.
Betting odds straight forward per game.

The data that was extracted goes back historically from 2014 .

Prediction Method

Intuitively the preferable machine learning and classification method would be the Logistic Regression test due to the format of the data and the format of the result being binary (win/loss). To ensure this I decided to run in addition to this test , the following tests:

- Linear Regression
- Passive Aggressive Classifier
- Perceptron
- Logistic Regression CV
- SGDClassifier
- Logistic Regression

Feature Selection

The model based itself under the assumption that a random guess (“coin-flip”) will provide an accuracy rate of 50%. There are several conventional features that are included in the model which were proven to improve the accuracy of the predictor.
To start off the predictor I added the features of whether a team is at home or away.
The outcome (Y variable) in this model is set to be 1/0 if the home team won that game.
The test was performed splitting the data into training sets and test sets of 80% - 20%.

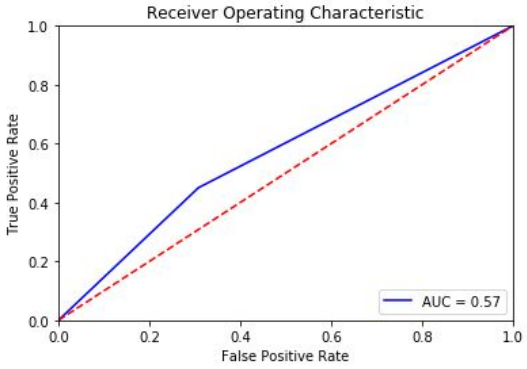
This set pertains observations through the dates 2014-01-02 until 2018-01-03.
There is a correlation of 16.7% between a team’s win and whether the team played at home.

Running the Logistic regression on the data sets , there is an accuracy score of about 57% . (Average over several random states)
The results featured in table 1 match the literature as well (Entine and Small 2008 , Tim B. Swartz, Adriano Arce 2014).

Table 1:

	Precision	Recall	F1-score
0	0.58	0.59	0.59
1	0.59	0.58	0.58
avg / total	0.59	0.59	0.59

Figure 2:



This insight gives a solid benchmark that will be comparable towards finding a better predictor and understanding towards outcomes of NBA games.

Going forward with the predictor several additional features that were taken under consideration were:

(a) How many days of rest were there between games
(b) Is the team on a current winning streak
(c) Average points per quarter prior to that given game.

One hypothesis was that given two teams , where one rested and the other played the day before , the chances of the rested team to win are higher than given both teams rested one day. The theory is that during this rest period , the players recover from the game before , and prepare better for the upcoming game.

In order to eliminate long term breaks between games such as end of seasons , all star breaks , any number of days rest that is larger than 5 was removed from the data set.

The distribution of the games according to the number of days rest between each game is described in figure (3). As expected , the league over the past several years has set the schedule so that most teams will have at least one day rest - mostly to prevent injuries and fatigue.

In figure (4) it is easy to see the incline in the winning average as the number of days between games grows.

Figure (3):

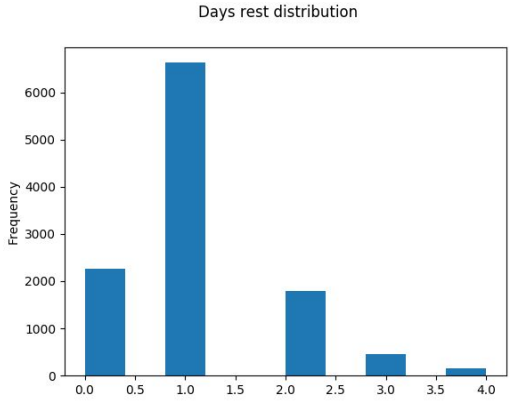
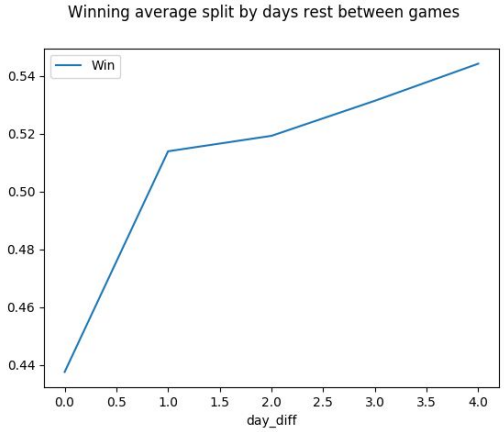


Figure (4):



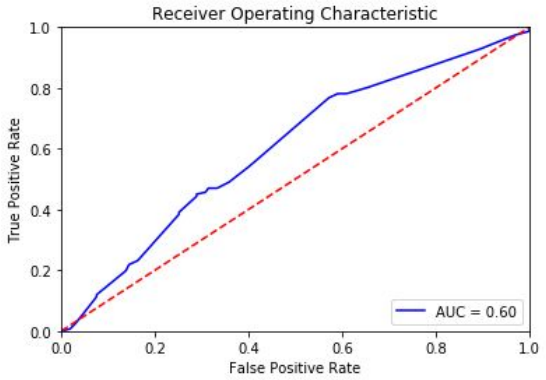
We suspect that a team that rested 4 days is not necessarily twice as likely to win as opposed to a team that rested 2 days, therefore an adjustment to the number of days ("day_diff") is made applying "log" to the number of days.

We see that when applying the adjustment of log to the number of days , the confidence level of the independent model goes up slightly and lowers the significance of an extra days rest slightly.
When adding the feature per home team and away team , the accuracy grows slightly . specifically in false negatives. [Table (5)]

Table (5):

	Precision	Recall	F1-score
0	0.56	0.75	0.64
1	0.59	0.38	0.47
avg / total	0.58	0.57	0.56

This ROC curve is very encouraging. In Figure (6) we can see that there is a distinctive improvement with the addition of the new feature , although 0.6 is still a poor result and requires more improvement.



One issue here is that there probably are teams that perform better at home and others that perform better away. In addition to that ,there are probably teams that matchup better against certain teams or perform differently against them. In order to test this hypothesis we are going to create an indication per game which team is at home and which is away .
Running the test using this feature only , we get to an accuracy score of about 65% , with an AUC score of 0.68 .(Figure 7) This can indicate that standing alone , this feature can explain a lot about the outcome of a certain game. On explanation to this might be that teams play differently when playing against teams from the same division. Another reason might be that some teams feel comfortable and used to playing against other teams , which might affect the home team's advantage.

New Metric calculation - "Similar Player"

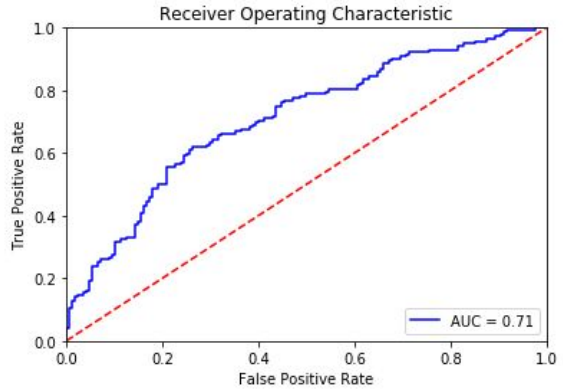
I wanted to add a metric which contributes the expected performance of the home team when they play against any specific team per player.
The way I created this metric is by understanding how similar players performed vs the opposing team. By understanding if similar players overperformed or underperformed I can understand and project towards the specific home team player , and expect a higher/lower performance accordingly. This can take in account mismatches of height or speed as well as defenses that the opposing team plays by which can neutralize or emphasize a specific role. (Lieder 2018)

In order to build this metric, we take the history per year 5 years back ,and look at the mean, sum and standard deviation aggregations of each player per category. This finds the players that develop in their career in the same manner. I run a process that finds the optimal number of clusters of players that perform the same way, using a scoring function that takes in account the silhouette score of the cluster and the log of the number of clusters that the model is splitting by. After receiving the most well split number of clusters given the former scoring restraint , I find the cluster that the player I am seeking (The given home team player) is in .
Per player in this cluster , I am interested in how well they play versus the upcoming away team , or in other words , do these players tend to over perform or underperform when playing against the given opposing team.
The method and metric that was used here was how many points per minute in average does the cluster give against that specific opposing team. This will give an idea of how well the player that belongs to this cluster will perform.
I run this method over all the players in the home team roster , and normalize this to a 240 minute game to get the projected number of expected points.
The output of the model is the projected number of points that the home team will have , and the projected number of points that the away team will have.
I calculate the odds of the home team to win using the formula

"Similar Player" Metric = $\frac{proj\_home}{proj\_home+proj\_away}$

Adding this metric into the model , we reach an improvement in the projection.

Figure (7)



The accuracy score of this model is about 68% after running from several random states.

	Precision	Recall	F1-score
0	0.85	0.21	0.33
1	0.51	0.96	0.66
avg / total	0.69	0.55	0.48

One interesting note here is that the recall for positive results is very high.

Using these controls to predict , it is easy to see that there is a distinct improvement in the accuracy.

In addition to these controlled parameters, I decided to include the betting odds of a given given game , prior to the game. The metric was calculated in a way that the team with the lower odds (without a spread) is favorable to win. Example , If Washington played Golden State , and the odds were 1.2 and 5.8 accordingly, I calculated the confidence level of Washington to be 1/1.2 = 0.83 and Golden State 1 / 5.8 = 0.17 . this metric will represent the confidence level and probability of each team to win.
The probabilities range between [0,1].

This as itself is a reasonable strategy (Pope & Peel, 1989) and represents the expert analysts opinions regarding the outcome. By adding this feature into the model we get similar results to the previous results with the AUC , though the accuracy score has gone down. Obviously the random state of the model needs to be taken into consideration , though it is safe to say that the contribution of this feature in the model. A future approach

can be a prior combination of the analysts prediction and the “Similar Player” prediction , and apply to the current model.

An approach to the use of this model would be to use this as a confidence score , where only if the model reaches an absolute prediction of >X% , then the outcome is defined , otherwise we can leave the prediction as “uncertain”.

Conclusions

I have provided a model using several features that touch many different aspects of the game using machine learning methods. This is a good start towards a more efficient and robust model . I moved from a truly random model to a basic prediction and further to a more unconventional method. The accuracy score of close to 70% . There can be several additional features that can be taken under consideration in the future , such as pre game rosters , injuries , player ranking , different streaks . Another approach would be to enhance the model using Zifan Shi et al recommendations . In addition , the parameters can be adjusted to provide even better results. Since there are many aspects which cannot be predicted in a basketball game , this score should be considered very high and satisfying.

References

(1) *Entine, O.A. and Small, D.S. (2008). The role of rest in the NBA home-court advantage. Journal of Quantitative Analysis in Sports, 4(2), Article 6.*

(2) *Lieder (2018). Common Denominator Between NBA Players.*

(3) *Pope & Peel (1989) .Information, Prices and Efficiency in a Fixed-Odds Betting Market*

(4) *Shi, Zifan et al. "Predicting NCAAB match outcomes using ML techniques - some results and lessons learned." (2013)*

(5) *Ashman, T., Bowman, R., & Lambrinos, J. (2010). The role of fatigue in NBA wagering markets: the surprising "home disadvantage situation."*

(6) *Harville, D. A., & Smith, M. H. (1994). The home-court advantage: how large is it, and does it vary from team to team? The American Statistician*

(7) *Bruce (2015) . A Scalable Framework for NBA Player and Team Comparisons Using Player Tracking Data*

(8) *Alagappan, Muthu (2012). "From 5 to 13. Redefining the Positions in Basketball". 2012 MIT Sloan Sports Analytics Conference.*