Enhanced power ranking function and basketball postseason brackets prediction through an innovative non-linear regression approach

Thibault Collin, Fall 2022

Abstract

The aim of our research is to design an algorithm capable of modelling an optimal and non-linear power ranking function for professional teams from the *National Basketball Association* based on their regular season performance, with the final objective being to predict the outcome of the postseason tournament. The algorithm shall be made such that the correct equation be available and usable. To the best of our knowledge, the following techniques used to design the algorithm were not used in the past literature. The power ranking function given by the algorithm is to be found at the end of this paper, and the program in itself was made to be easily usable by anyone.

Introduction

Arguably the most popular and competitive professional basketball league in the world, the *National Basketball Association* is scheduled as a two steps tournament. The regular season first opposes thirty teams equally divided in two conferences (*Eastern* and *Western*). These teams face each other multiple times until the postseason kicks off, where only those with the best record compete for the championship in a bracket fashion. Out of those thirty teams, all represent a city in the *United States* with the exception of *Toronto*. The plethora of consistently recorded statistics from every game in the league for the past couple of decades opened up a general interest in finding what statistics allow to accurately predict the outcome of a game, and more generally a season.

We wanted to build an algorithm such that the outcome function is known and available, which is not possible using machine learning techniques. This explains why we stuck with statistics and simple regression analysis. The algorithm will be divided into two steps: a first function will randomly generate a structure given an input set of variables - another will perform a linear regression analysis, whose R^2 is to be computed. The final model retained will be the first one displaying a score above a predefined threshold.

This model comes with an equation, which we will coin as our *non-linear power ranking function*. It will be used to rank teams based on their performance. The lower the better, our model shall eventually display how far each team is predicted to go in the postseason. Come the end of the *2022-23* season, we shall publish the prediction given by the algorithm for the *2023* playoffs, therefore assessing the practical accuracy of our program.

The paper is built as the following. First, we will delve into an analysis of the most interesting variables to consider for our study, based on historical data and general basketball knowledge. We shall next explain how the random structure generator was developed using *Python*. The paper ends with the non-linear power ranking function constructed and given by the algorithm, followed by remarks and ideas for further lines of research. The entire program is freely available in the following *GitHub* repository: *thibaultcoo/ball-postseason-predictor*

Considered Statistics

Tracking and recording sports-related statistics started to appear during the early *1980s*, driven both by the need for the coaching staff to have a more detailed view of what the team produced, but also by the betting industry, whose actors saw an incredible opportunity for profit. Statistics are nowadays heavily scrutinized and analyzed, and the duality of their quantity and quality allows for very precise studies to be conducted.

In this study, we focus on basketball and its professional North-American league. This choice was fueled by several key factors, the first being the worldwide popularity of the sport and its numerous star players. Another was the variety and quantity of well-recorded advanced statistics and analyses on each game and season for a significant time period. The final and most important one was our profound interest and love for the game of basketball.

Played at the highest level, basketball requires the mastering of several different aspects of the game from both sides of the court. Lacking an equilibrium between attacking sharpness and defensive solidity almost certainly prevents a team from being a championship contender, although the *'17 Golden State Warriors* and the *'04 Detroit Pistons* could argue contrariwise.

In that sense, delving into what precisely makes a team efficient in those two sectors becomes an absolute necessity. The most valuable variables to consider will be the ones best displaying the dominance, efficiency and consistency of a team in a season. All data is to be extracted from [2] - going from *2015* to *2022*, unless mentioned otherwise. We noticed that going back

further in time significantly reduced the quality of our regressions, and the reasons for that are to be explained later in the paper.

ORtg	Offensive rating
DRtg	Defensive rating
FTr	Free-throw rate
Pace	Estimate number of possessions per game
eFG%	Effective field-goal percentage
TOV%	Turn-over percentage per 100 plays
ORB%	Offensive rebound percentage
FT/FGA	Free-throw per field goal attempt
AdveFG%	Opponent effective field-goal percentage
AdvTOV%	Opponent turn-over percentage per 100 plays
DRB%	Defensive rebound percentage
AdvFT/FGA	Opponent free-throw per field goal attempt

Table 1: *List of all the variables we shall consider*

Random Structure Generator

In this section, we shall explain clearly and concisely how can our algorithm randomly generate a new data structure from the input variables, leading then to a linear regression analysis [1] on this transformed set. From the initial set of considered variables given by the user, the program generates a transformed set comprised of several new variables. Each new variable is a random non-linear combination of variables from the initial set.

For each of these variables is firstly randomly decided how many initial variables to consider. On this restricted new set is then chosen two separate types of operations: the power applied to each individual variable, and the operations linking all the variables from this set. Between one and three variables are generated following such process. The final transformed data set is therefore composed of several new variables, each produced as a combination of raw input variables, and for which the structure is random and potentially nonlinear.

Results and Conclusions

Running our algorithm a significant amount of times provided several interesting equations, among which the most relevant is to be expressed below. Initially aiming for a model with an *adjusted* $R^2 \geqslant 62$ %, our method gave a result above the threshold in less than 20 minutes.

Proposition 1. *According to our algorithm and given the following set of inputs, the equation below is supposedly able to predict the outcome of the 2022 NBA postseason bracket, given the performance of each team in the regular season (the lower the better):*

$$\phi\text{: } ORtg \qquad \alpha\text{: } DRtg \qquad \rho\text{: } FTr \qquad \omega\text{: } Pace \qquad \gamma\text{: } eFG\% \qquad \beta\text{: } TOV\%$$
$$\xi\text{: } ORB\% \qquad \Omega\text{: } FT/FGA \qquad \epsilon\text{: } advFG\%$$
$$\zeta\text{: } advTOV\% \qquad \eta\text{: } DRB\% \qquad \upsilon\text{: } advFT/FGA \qquad \tau\text{: } Power\ Ranking$$

$$\tau = 3.17 - \frac{6}{10^4\sqrt{\upsilon}}\left(\left(\phi^3 - \alpha^3\right)\gamma^4 + \Omega - \sqrt{\eta}\right) + \frac{4\epsilon\zeta}{10^6}\left[\left(\frac{\sqrt{\rho}}{\beta} - \xi\right) + \eta^3 + \upsilon\right] - \left(\frac{\beta\gamma^4\sqrt{\epsilon\alpha\omega} - \upsilon}{10^2}\right)$$

Phoenix Suns	Champion
Boston Celtics	Runner-up
Utah Jazz	Western Conf Runner-up
Miami Heat	Eastern Conf Runner-up
Golden State Warriors	Western 2nd round
Memphis Grizzlies	Western 2nd round
Milwaukee Bucks	Eastern 2nd round
Philadelphia 76ers	Eastern 2nd round

Table 2: *Algorithm prediction for the 2022 season with data ranging from 2015 to 2021*

Structure code	4be52aaeab
R^2 threshold	62%
Iterations	214682
Time elapsed	1126 seconds

Table 3: *Specifics of this structure*

The algorithm was constructed such as to be as flexible as possible. This allowed for lots of experiments and variations, to try and explore other axes of research. One first important conclusion is that the coefficient of determination seems to reach a cap at around 60 % on the entire data set, while it was on some occasions able to reach slightly higher levels for shrunk sets. The growing difficulty to find a significant and consistent relationship between our variables throughout three decades of seasons might have been caused by the many drastic changes that gradually appeared in the way the game of basketball was played, as well as the different dynasties that have successively reigned in the league.

For example, the *Chicago Bulls* dominated their conference for an entire decade playing tough defense and being aggressive in the paint with the likes of *Rodman* and *Jordan*, a play style

in which AdvTOV% and DRB% arguably do play a more significant role than other statistics that could be considered. This is to be contrasted with the dynasty of the *Golden State Warriors* from the mid-to-late *2010s*, for which eFG% and TOV% play a significantly stronger role, considering their offensively lethal trio formed by *Curry*, *Thompson* and *Durant*.

On a technical side, the algorithm is able to extract 32 seasons worth of data in approximately 50 seconds, with on average 258 new random structures generated and analyzed every second. This is considered satisfying and not time consuming, although finding a solution may be really slow for a higher threshold. For example, no solution was found by the algorithm with a coefficient of determination above 65%, after more than 2,600,000 iterations.

Further Research

More than a tool designed to build a simple prediction one year ahead, the flexibility provided by the algorithm and its structure allows for a variety of interesting tests and simulations to be realized. Being able to extract a data set of any size ranging from any periods gives an opportunity to test the accuracy of our algorithm for a prediction made several decades ago.

In this algorithm, one is able to tweak approximately everything, ranging from the dates to the list of input variables, as well as the score threshold or even the probabilities of appearance for certain linking operations of the input variables.

We could try as many combinations as possible to evaluate the accuracy of a randomly generated structure. This is made possible by the storing feature that was implemented in our algorithm, which allows one to save a generated structure and to use it later on a potentially different data set. Indeed, we showed empirically that because of the heavy changes in the way the game was played and the changing statistics that were at stake, considering a data set too large worsened the predictive ability of our algorithm.

A vast amount of further research can also be realized with the considered input variables. Realizing further individual tests of significance for each variable might allow one to detect some that are very weak in the analysis, and their removal would potentially greatly increase the pace and accuracy of the program. On the other hand, adding some key and not yet considered variables could add some explanatory power to our regression.

References

1. Jeffrey M. Stanton (2001), *Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors*, Journal of Statistics Education.

2. https://www.basketball-reference.com/