MSc Project Proposal
Predicting the Outcome of the National Basketball League Games

Oyeleye Mojubaoluwa
MSc Date Science

April 2020

1 PROJECT GOAL

Over the span of the National Basketball Association (NBA) regular season (a full season usually will run from October to April barring delays), at least one game is played on most days with a few exceptions e.g. Christmas Eve.

The goal of this project is to design an algorithm that takes in the up to date NBA regular season match results and predicts what teams are going to win games the next day games will be played. This algorithm will be based on the methods discussed in Section 3.

Predicting a winner when competing teams are of comparable strength is difficult and this is the case in the NBA. As a result, this project will go a step beyond Algebraic methods to compare teams (such as Massey's and Colley's rating methods), by accounting for fatigue.

2 INTRODUCTION

Sports has become a part of many lives around the world. From participating athletes to ever-expanding teams put together to assist the athletes' performance, to fans investing in teams or athletes they support, everyone involved wants one question answered. Is my team going to win?

As the NBA popularity and revenue expands, more technology is used in athlete monitoring, and as a result, more data is being collated. This allows for deeper exploration into how best to determine which team wins a game.

This project aims to answer the question posed above, determining which of the two teams involved in a regular-season game is going to win. There are only two possible outcomes for each of the teams involved in an NBA game, a win or a loss. This makes the problem a binary classification problem.

This process will involve: (1) Gathering preseason and regular season match data for NBA games, (2) Generating features from basic match data, (3) Selecting features, through data exploration and statistical analysis, which will best point to the team

more likely to win a game. (4) Building a predictive model that will be trained on the selected features.

The main challenge presented by this project is the difficulty in capturing all factors that affect the outcome of games. A games' outcome depends on the participating athletes and coaching staff. The athletes and coaches are also affected by many other factors, most of which cannot be quantified. This could lead to sudden performance changes mid-game which are too complex to represent as data.

Even if the coaches and players output could be quantified, another important factor in determining the relative team strength is team chemistry. This is a measure of how well everyone involved gets along together or how well they trust each other. Attempts have been made to quantify team chemistry by a metric called 'prior shared success.' This metric factored in a team's past success in predicting the outcome of games. Across different sports including the NBA, this inclusion was found to improve the accuracy of predicted games by between 2 and 7 per cent [1].

Player trades, [2] buyouts [3], as well as coaching changes, should be also be considered as these factors affect both the strength of the team and the team chemistry.

3 PROBLEM DESCRIPTION AND METHODOLOGY

The binary classification problem will answer the question; will the home team win? The answer will be a Boolean value with 1 representing yes and 0 representing no.

To solve the problem, I intend to spend quality time focusing on three overlapping areas:

3.1 GENERATING FEATURES

For each game, features will be generated by looking at data from previous games. There are three major kinds of features I intend to capture:

3.1.1 RELATIVE STRENGTH OF TEAMS

This will include capturing basic features such as win percentage of teams, home and away records as well as more complex features that take into account the strength of opponents. This will include rating methods such as Massey's, Colley's as well as other rating methods outlined in the book [4]. These rating methods are discussed in more details in section 4.

To generate these features, I intend to investigate the effect of looking only at n games where:

$$\boldsymbol{few\ previous\ games\ \leq n\ \leq multiple\ previous\ seasons}$$

I also intend to investigate the effect of home games and away games in rating teams. The disparity in home and away records of teams could be significant enough that a home team looks like a completely different team away from home. For example, as at the all-star break in the 2019/2020 season, the Philadelphia 76ers have won 34 out

of 55 games played. This gives them a winning percentage of 62%. However, they have a home winning percentage of 93% and an away winning record of 47%. This will be investigated further when calculating the relative strength of teams.

3.1.2 FATIGUE FACTOR

The features will try to capture the element of fatigue and how it affects a team's performance. The exact value of team load depends on each players load and this, in turn, depends on many factors that cannot all be captured including the amount of sleep, commitments outside basketball, sleep quality. The idea is to try and approximate the load value of each team based only on the games played. This will include investigating basic features e.g. a Boolean value to show whether or not a team is playing a back-to-back game (2 games on consecutive nights), and more complex features which include both the acute and chronic versions of the exponential weighted moving averages (EWMA) given by:

$$\boldsymbol{EWMA_{today} = LOAD_{today} \times \lambda_a + \left((1 - \lambda_a) \times EWMA_{yesterday}\right)}$$

Where λ_a is a number between 0 and 1 that attributes a decay rate to the load value. This decay rate is calculated as

$$\boldsymbol{\lambda_a = \frac{2}{N + 1}}$$

In this equation, N is the chosen time-decay constant, or in this scenario the time usually given to the acute (7 days) and the chronic (28 days) periods [5].

The distance between previous and current playing venues will also be investigated to see how much it contributes to team fatigue.

3.1.3 KEY INJURIES

Due to the size of the NBA playing surface and the number of players (5) from each team playing, each player's impact on the game is larger than in most other sports. The most elite athletes in the NBA are called superstars and their presence on the court has a big impact on winning [6]. A good example is LeBron James. LeBron James has left a team to join another three times. The team he leaves lost an average of 30 more games after his departure while the team he joins wins an average of 11 more games after he joins. The first 2 times he joined a new team, his new team won 11 and 20 additional games. However, his most recent move to the Los Angeles Lakers saw the team win only 2 games more than the previous season. This could be attributed to the fact that for the first time in his career, he missed 27 games due to an injury. Before his injury, his team was on pace to win 48 games. They ended up with 37 wins [7].

The features will try to capture how much a team is missing in terms of strength because one or more player(s) are missing.

3.2 SELECTING FEATURES

After features have been generated, it will be important to select only those that will best answer the question - will the home team win?

Features that try to capture the relative strength of teams will be reduced in terms of dimension. I intend to investigate two strategies;

1. Reducing the generated features to two features, one representing the strength of the home team and one representing the strength of the away team.
2. Reducing the generated features to one feature that represents the strength of the home team relative to their opponent.

The process of reducing the number of dimensions will involve investigating using traditional dimensional reduction techniques such as PCA, as well as using ensemble techniques to combine features.

The feature or features that measure the relative strength of teams will be the most important. The strategy is to try and improve model performance by adding in features measuring fatigue factor and key injuries.

3.3 MODEL SELECTION

This process will involve selecting a model that best answers the question. To begin with, I intend to build a simple logistic regression model with little hyper-parameter tuning. This model will serve as a benchmark in the model selection process.

Accuracy will be used as the standard measure to compare the performance of models.

I also intend to design and build a multilayer perceptron. This will include hyper-parameter tuning to minimise the classification error rate.

Lastly, I intend to investigate the use of recurrent neural networks in predicting. This would involve designing and building a recurrent neural network using long short-term memory (LSTM) architecture and comparing the performance of this neural network to the performance of the other models.

All models will be benchmarked against predicting based on the winning percentage of the competing teams i.e. selecting the team with the higher winning percentage to win.

4 RANKING METHODS [4]

4.1 MASSEY'S METHOD

The ratings of teams are obtained by solving the system $\bar{\boldsymbol{M}}\boldsymbol{r} = \bar{\boldsymbol{p}}$ to obtain the general Massey rating vector ***r***. To begin, two matrices ***T*** and ***P***, as well as two vectors ***f*** and ***a***, are created.

For ***n*** teams in the league:

T is an $\boldsymbol{n} \times \boldsymbol{n}$ diagonal matrix containing information about the total games played. $\boldsymbol{T_{ii}}$ is equal to the total number of games played by team ***i*** during the season.

$\boldsymbol{P}$ is an $\boldsymbol{n} \times \boldsymbol{n}$ off-diagonal matrix containing pair-wise matchup information. $\boldsymbol{P_{ij}}$ is equal to the number of pair-wise matchups between teams ***i*** and ***j***.

$\boldsymbol{f}$ is an $\boldsymbol{n} \times \boldsymbol{1}$ cumulative points-for vector. $\boldsymbol{f_i}$ is equal to the total points scored by team ***i*** during the season.

$\boldsymbol{a}$ is an $\boldsymbol{n} \times \boldsymbol{1}$ cumulative points-against vector. $\boldsymbol{a_i}$ is equal to the total points scored against team ***i*** during the season.

Next, the matrix ***M*** and vector ***p*** are created as;

$$\boldsymbol{M = T - P}$$

$$\boldsymbol{p = f - a}$$

The rows of ***M*** are linearly dependent, so $\boldsymbol{Mr = p}$ does not have a unique solution. Massey's workaround is to replace any row in ***M*** with a row of all ones and the corresponding entry of ***p*** with a zero. The new system with the row adjustment is denoted

$$\bar{\boldsymbol{M}}\boldsymbol{r} = \bar{\boldsymbol{p}}$$

4.2 COLLEY'S METHOD

The ratings of teams are obtained by solving the system $\boldsymbol{Cr = b}$ to obtain the Colley rating vector ***r***.

$\boldsymbol{C}$ is an $\boldsymbol{n} \times \boldsymbol{n}$ real symmetric positive definite matrix called the Colley matrix defined by;

$$\boldsymbol{C_{ij}} = \begin{cases} \boldsymbol{2 + t_i} & \boldsymbol{i = j}, \\ \boldsymbol{-n_{ij}} & \boldsymbol{i \neq j}. \end{cases}$$

For ***n*** teams in the league:

$\boldsymbol{t_i}$ is the total number of games played by team ***i***

$\boldsymbol{-n_{ij}}$ is the number of times team ***i*** and ***j*** faced each other

The right-hand side vector $\boldsymbol{b}$ is defined by;

$$\boldsymbol{b_i} = \boldsymbol{1} + \frac{\boldsymbol{1}}{\boldsymbol{2}}(\boldsymbol{w_i - l_i})$$

Where:

$\boldsymbol{w_i}$ is the total number of wins accumulated by team ***i***

$\boldsymbol{l_i}$ is the total number of losses accumulated by team ***i***

Other rating methods including Markov method and the offence-defence method, discussed in [4] will be investigated.

5 PROJECT STRUCTURE, SOFTWARE AND DATA

The algorithm will be built in Python using methods and functions from the pandas, NumPy, Matplotlib, requests and scikit-learn libraries. The requests library will be used to get data, the pandas and NumPy libraries will be used to generate features, the Matplotlib library for visualisation and the scikit-learn library to build machine learning models for classification. Figure 1 shows the proposed data flow and interactions between the different parts of the algorithm.

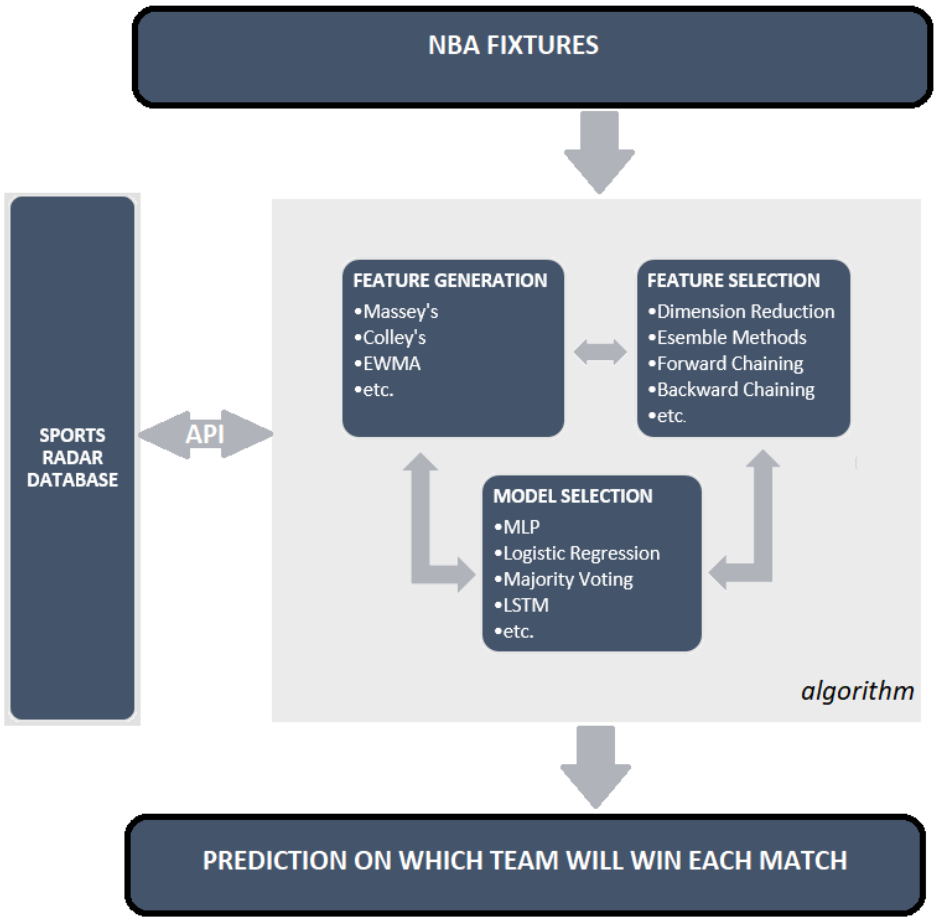

FIGURE 1: THE FIGURE SHOWS THE FLOW OF DATA AND THE INTERACTIONS BETWEEN PARTS OF THE PROPOSED ALGORITHM

The project will be divided into two main components, to be executed in sequence:

5.1 DATA GATHERING

The main data source for this project is Sports radar which is the NBA's official provider of real-time NBA league statistics [8]. The data will be accessed via an API connection using the requests library in python.

The features of interest are:

- Home Team
- Away Team
- Home Team Score
- Away Team Score
- Date
- List of unavailable players from the home team
- List of unavailable players from the away team
- Match venue

The rows of data will include all preseason and regular season NBA games between the 1999/2000 season and the 2019/2020 season.

Data showing the geographic coordinates of the match venues will also be necessary to investigate the effect of distance travelled on match outcome.

5.2 MODEL BUILDING

Once all the data discussed in section 3.1 is accessed, the feature generation, feature selection and model building discussed in section 2 will begin.

6 TIMELINE (PRELIMINARY)

For the execution of the project, the timeline is:

1. April 2020: complete the literature review and understanding ranking methods.
2. May 2020: complete the data gathering process.
3. June 2020: complete the creating of features from raw data.
4. July 2020: complete the selection of features and building of the algorithm.
5. August 2020: optimization and hyper-parameter tuning of the model.

REFERENCES

[1] N. University, “In team sports, chemistry matters: Sports analytics analysis reveals that past shared success among team members improves odds of future wins,” ScienceDaily, 4 December 2018. [Online]. Available: https://www.sciencedaily.com/releases/2018/12/181204095355.htm. [Accessed 28 March 2020].

[2] F. N. B. A. Staff and NBA.com, “NBA Trade Deadline Explained,” NBA.com, 6 February 2020. [Online]. Available: https://www.nba.com/trade-deadline-explained. [Accessed 28 March 2020].

[3] S. Quinn, “2020 NBA buyout market: How midseason free agency works, and who could be available after the trade deadline,” CBSSports.com, 26 February 2020. [Online]. Available: https://www.cbssports.com/nba/news/2020-nba-

buyout-market-how-midseason-free-agency-works-and-who-could-be-available-after-the-trade-deadline/. [Accessed 28 March 2020].

[4] A. N. Langville and C. D. Meyer, Whos #1?: the science of rating and ranking, Princeton, NJ: Princeton University Press, 2012.

[5] I. Foster, P. J. Byrne, J. A. Moody and P. A. Fitzpatrick, "Monitoring Training Load Using the Acute: Chronic Workload," *ARC Journal of Research in Sports Medicine,* vol. 3, no. 1, pp. 22-28, 2018.

[6] S. K. Deshpande and J. T. Shane, "Estimating an NBA Player's Impact on His Team's Chances of Winning," *Journal of Quantitative Analysis in Sports,* vol. 12, no. 2, 2016.

[7] "Basketball Statistics and History," Basketball, [Online]. Available: http://www.basketball-reference.com. [Accessed 29 February 2020].

[8] "National Basketball Association," Sportradar, [Online]. Available: https://www.sportradar.com/about-us/exclusive-data-partnerships/national-basketball-league/#article. [Accessed 28 March 2020].

[9] J. Bois, "13 points, 33 seconds: The night Tracy McGrady was a basketball god," SBNation.com, 27 August 2013. [Online]. Available: https://www.sbnation.com/2013/8/27/4661546/tracy-mcgrady-13-points-33-seconds. [Accessed 29 February 2020].

[10] "How the NBA Schedule is Made," 12th Anniversary, [Online]. Available: https://www.nbastuffer.com/analytics101/how-the-nba-schedule-is-made/. [Accessed 21 March 2020].

[11] Hierarchystructure, "NBA Basketball Team Hierarchy," Hierarchy Structure, 4 December 2017. [Online]. Available: https://www.hierarchystructure.com/nba-team-hierarchy/. [Accessed 21 March 2020].