Predicting the final seeds of National Basketball Association teams, an Elo based approach

Thomas Abraham  
210639757Hazar Emre Tez  
MSc Computer Science, QMUL

*Abstract*—The introduction of statistical analysis into the National Basketball Association has radically changed the way basketball shot selection and plays are thought about. The Elo algorithm is a popular ranking algorithm used to rank players in a competitive setting. The Elo algorithm is often associated with chess as it is used extensively by FIDE (the governing body of international chess) to determine world rankings of chess players. For the model being considered here, the Elo rating will depend on the players +/- for each game. The model will include only regular season games as they are defining factor when it comes to a team’s final seed. Teams will gain points after winning matches and lose points post defeat. Finally, teams will be seeded from one to fifteen in each conference based on their Elo scores. The generated seeds will be validated by creating the model for a previous season and cross-checking its accuracy with real world results. This model will then be compared with a Random Forest Classifier and Logistic Regression approach and a distinction will be made based on efficiency and accuracy.

Keywords—plays, Elo algorithm, playoffs, seeded, Random Forest Classifier, Logistic Regression.

# Introduction

The National Basketball Association (NBA) is a men’s professional basketball league in North America, composed of thirty teams divided into western and eastern conferences. Each NBA team has a maximum of fifteen players, out of which thirteen are allowed to be active in each game. Players on a basketball court position themselves in five locations as shown in Figure 1.1. Each of these positions require distinct abilities and physical attributes.

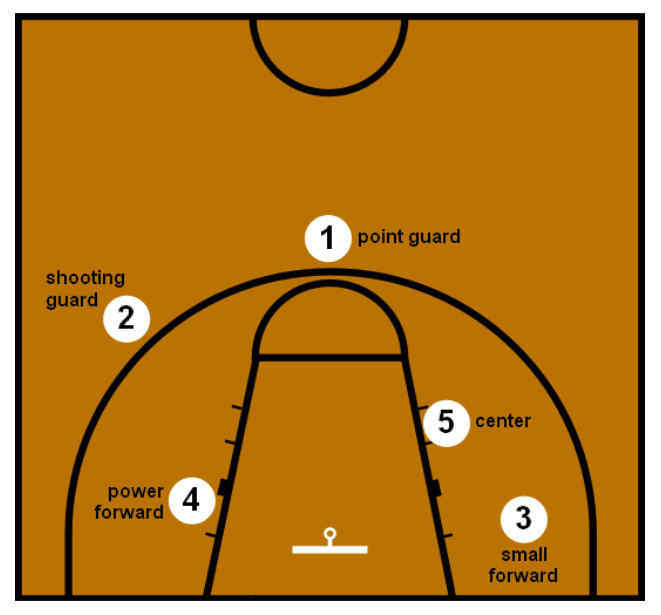


Figure 1.1. Traditional positions in basketball

At the highest level of professional basketball, every play is executed by a team is carefully devised by a team of coaches. Basic plays such as a switch or a give-and-go are widely used and can be situational. However, particular plays such as those centred around off ball movement in order to ensure that a particular player loses their defender are drawn up after taking into consideration the player as well as the defender’s ability. The ultimate goal of the offense is to shoot the ball, the strategy lies in devising plays to create good shot opportunities. Defensive plays on the other hand are often drawn up to restrict the movement and ability of one particular player or a set of players. These types of plays are often drawn up after watching a particular player over multiple games and analysing their offensive and defensive capabilities.

General managers such as Darryl Morey set out to prove that data driven decisions would result in a competitive edge. Previously implemented conventional methods used to simulate gameplay and deduce plays have ignored that in a sport such as basketball the dynamics of movement and cohesiveness are unique from line-up to line-up and do not depend solely on individual offensive and defensive ability.

The Elo rating system is popular and widely used, this is mainly because it is elegant yet simple in its execution. A rating system analyses the outcomes of matches and assigns a value to the strength of a team relative to others [3]. This information then allows teams and coaches to make beneficial long-term decisions regarding team formation, trades, player rotations and drafts [4]. The performance in the Elo system is not measured in absolute terms. It is inferred from wins and losses.

The Random Forest Classifier is a flexible and intuitive algorithm that can be used for both classification and regression tasks and tends to produce an accurate result. It is widely used due to its simplicity and effectiveness. This algorithm uses the decision tree classifier method but instead of creating a single tree, it creates multiple. This gives it the opportunity of random sampling of data as each individual tree learns from a random sample of data points which are drawn without replacement. This then minimizes the possibility of over-fitting and improves overall predictive accuracy of the model. This is because the final predictions will be made by computing the mean of the predictions of each individual tree. Using this particular model should also give us the ability to extract feature importance which tells us which variable or variables contribute the final result the most. One foreseeable drawback is that this model may be difficult to interpret in comparison to a single decision tree because of the complexity involved with combining multiple trees. Finally, the results obtained from both models will be compared for accuracy and efficiency [10].

The logistic regression model is used for classification and prediction analysis. It estimates the probability of an event occurring such as a win or a loss in this case. Since the result generated is a probability, the dependant variable is bound between 0 and 1 [15]. This model only uses the specified data to make a prediction without taking any other factors into account. Thus, removing all bias from the equation.

Based on the analysis of the possible advantages and drawbacks of all three approaches it can be assumed that the Elo based approach will produce the most accurate result.

# Literature Review

Accurate predictions of a team’s seed greatly influence a team’s chances of winning prior to the playoffs. The ability to discern a potential round one matchup at the least will give coaching staff the opportunity to condition players and design game plans specifically suited to their opponents. The Elo model when run through the course of a season should analyse various strengths and weaknesses when it comes to likely matchups. Comparing the outcomes of both the Elo model and the Random Forest model will allow coaches and trainers to prepare for every likely outcome.

## Elo Algorithm:

A base understanding of the Elo model indicates that for a metric to be useful for a particular decision, its treatment of variation needs to match up with the decision that is being made [1]. While we can isolate some player, season, and team variation by analysing and dividing the data, all measurements that we take are confounded with the randomness of chance. The main feature of the Elo rating system is that performance is not measured in absolute terms but is derived from wins and losses against other players with varying ratings [4]. In other words, player ratings depend on both their performance and the ratings of their opponents [11]. This ensures that an effective comparison can be made with other Machine Learning models.

From a general understanding of most sports, it can be discerned that better statistics do not always infer a win. In most cases an athlete will perform at their average level throughout their career. Deviations from this do occur, large deviations occur less frequently than smaller ones. Hence, it was assumed that “the many performances of an individual will be normally distributed, when evaluated on an appropriate scale [2].” It is for this reason that the contribution of each player is modelled as a normally distributed random variable.

Another reason for the selection of the Elo model is that it can account for the margin of victory. Teams will gain rating points after wins and lose the same after losses, but they also gain or lose more points based on margin on victory i.e., a blowout win, or loss is more consequential to a team’s rating. This can be implemented by assigning a multiplier to each match and dividing it by the team’s probable margin of victory if they win the game [9].

Building on previous work [4], this project sets out to create a model that can accurately predict each of the fifteen seeds in both conferences for a season. An Elo based approach is employed to obtain probable wins in a head-to-head matchup between two teams based on player ratings. The individual player ratings are combined to obtain a team rating. Team ratings are then compared pairwise to obtain the probability of a win by each of the teams during the season. The probable wins are then cumulated, and a seed is calculated for each team. Each of the fifteen teams in each conference are then ordered based on their win to loss ratio. The rating system is validated by running them over real-world data from previous NBA seasons.

The largest noticeable drawback of the Elo algorithm is that two teams can have identical results but end up with different ratings because the ratings are calculated as a change to the current rating. In a practical sense it works as it is supposed to because a vast majority of teams improve at a very slow pace. However, the system can be seen as an unfair to teams that improve rapidly from a low starting point.

## Random Forest Classifier:

Classification constitutes a large portion of machine learning. The ability to precisely classify observation is extremely important when it comes to making accurate predictions. Individual decision trees are combined to make a random forest. A decision tree is a flowchart-like structure, where each node denotes a test on an attribute, each branch represents an outcome or result, and each terminal node holds a class label [12]. The varying number of decision trees in the model operate as an ensemble. Each of the individual trees in the random forest then reports a class prediction and the class with the most words are taken as the model’s prediction. The fundamental principle behind this approach is “A large number of relatively uncorrelated models operating as a group will outperform any of the individual constituent models [13].” The randomness associated with generating the individual trees minimises the possibility for over-fitting and improves the overall accuracy of the model. This is primarily because the final prediction is deciphered by calculating the mean of the predictions of each individual tree, thus following the above-mentioned principle.

An observable drawback of this model is that it is not easily interpretable. It provides feature importance, but it does not provide complete visibility into the coefficients. It is also computationally intensive for large datasets and the user has very little control over what the model actually does [14].

## Logistic Regression:

This particular model allows its user to estimate the probability of a categorical response based on predictor variables. These responses are traditionally binary values but can even be categorical if required [16]. Logistic regression is an ideal choice because it tends to produce good accuracy for simple data sets and performs well when the dataset is linearly separable, and it can interpret model coefficients as a measure of feature significance.

The main disadvantage of logistic regression is the presupposition of linearity between the dependant and independent variables. Furthermore, logistic regression is bound to discrete number sets as it can only be used to predict discrete functions. The number of observations should always be greater than the number of features, otherwise, it can lead to overfitting i.e., the model won’t be able to make accurate predictions about new data because it cannot distinguish between noise and essential data [17].

# Methodology

## Elo Approach

The original purpose of the Elo algorithm was to develop a viable rating system for chess players. As the popularity of the algorithm increased, analysts and statisticians began modifying the algorithm so it could be applied to various other sports. In the simulations being considered here, players alone are not given individual attention when it comes to win prediction. Instead, the team is considered as a single entity.

The true value of a player is not specifically quantifiable and therefore, cannot be measured and analysed. Hence, we depend on the observable metrics of the sport such as points scored, rebounds, assists and so on. The primary statistic being considered in the algorithm is the plus-minus score. The algorithm as a whole is designed to track the performance of individual basketball players and combine their ratings in order to obtain a team score which can then be used in the simulations.

### Plus-Minus Score (+/-)

The Plus-Minus score reflects how a team performed while a particular player was on the court. The introduction of the adjusted plus-minus score redefined the understanding of player value. The league wide statistic of adjusted plus-minus takes into account a player’s marginal effect on the score per 100 possessions as compared to a league average player. This metric is widely used for comprehensive player analysis prior to a crucial matchup. Adjusted plus-minus is preferred over unadjusted plus-minus because in the latter each players rating is heavily influenced by the play of his on-court teammates [19]. A positive score indicates that a player has a beneficial effect on their team’s performance when they are on the court and vice-versa.

### Player Strength

Let the variable used to denote an NBA player’s point contribution per minute be . Therefore, assesses the strength of a player. In order to simplify the interpretation of the data, each player in a team is initialized with a value of 1000 (normalizing constant is multiplied). An appropriate value of is obtained empirically [4].

(1.2.1)

### Estimate

In the algorithm, it is assumed that each team's actual strength is derived from a normally distributed random variable, with the team's actual strength being represented by the mean. A team that maintains the same lineup every game should perform at the same strength. Due to this reason a normal distribution is chosen. The rating of a team is updated continuously based on observed wins and losses. If plays, then the rating is updated as [3]:

(1.3.1)

Where *R* refers to the rating, *K* refers to the K factor, *S* refers to the actual score and *x* refers to the expected score.

### Actual Score (S)

The actual score being considered refers to the victory/defeat information acquired after a match.

The definition of is depicted as:

[3]

### Expected Score (x)

Variable is used to denote the expected outcome of a match between and . When two players are matched up with each other, the overall performance of the players is modelled as a normal random variable [20].

The probability that wins against is depicted as [4]:

(1.5.1)

Where and are individual scores assigned to and . When considering an exponential score, the expression reduces to a logistic function:

(1.5.2)

Where and are the ratings of and . The standard Elo algorithm is depicted as:

(1.5.3)

Where 400 is the constant scale factor. Let be used to denote . Therefore is denoted as:

(1.5.4)

In order to estimate the plus-minus score of a player, a model is created. Whenever two teams have a matchup, the individual player strengths are aggregated to obtain a combined strength parameter for the team [4]. The strength of is depicted as:

(1.5.5)

Where denotes the minutes played by the th player on the th and denotes the estimated strength of the th player on . Furthermore, denotes the average points scored per minute by [4].

### K Factor (K)

The K-factor determines how quickly the rating reacts to new game results [9]. A high K value allows the estimate to adapt quickly, however if K is set to high it will result in the large variations in the estimate. On the other hand, if the K value is set too low then the estimate will take too long to recognize important changes. The K factor being selected here depends on the total number of minutes being played by each individual player.

There are still multiple cases where the algorithm is too slow to catch up to major trades or signings like when Lebron James was signed by the Lakers or when Kevin Durant left the Golden State Warriors. Furthermore, a bad start to the season could result in extremely low team rating, however the team may go on to finish the season with a win rate of greater than 50%.

### Match Outcome

Utilizing (1.5.5), the overall team rating is obtained based on the strengths of the individual players on the team. Consider that the team ratings of and are denoted as and respectively. The probability that wins a matchup against is derived using (1.5.3). For the simulations being considered in these particular instances, a win is predicted if has a higher overall rating than .

### Seed Outcome

A seed in basketball represents a number which correlates to a team’s ranking. In the NBA seeds are determined based on a teams win record or win rate. The team which finishes the season with the best record is awarded the first seed, the second-best team gets the next seed and so on. In each conference i.e., Eastern and Western, the top 8 seeds advance to the playoffs and the higher seeds are awarded home court advantage. The match results correlated above are recorded and teams are ordered based on their win rate.

### Algorithms

#### Update Team Rating [4]

|  |
| --- |
| **Algorithm 1**: Update Team Ratings according to Elo Algorithm |
| Initialize all team ratings to a 1000.  **for all** matches between two teams and do  Compute and which corresponds to the probability of and winning respectively  Update rating for and according to equation 1.3.1  **end for** |

#### Predict Match Winner

|  |
| --- |
| **Algorithm 2**: Predict the winner of match between and |
| Determine the functional strength of based on (1.5.5)  Determine the functional strength of based on (1.5.5)  The team that has the greater strength of the two is expected to win |

### Datasets

#### Data:

The required datasets are obtained by scraping data from the NBA website and Basketball Reference. The data is obtained from the 2018-2019, 2019-2020 and 2020-2021 regular seasons. Both the player box score and the team box score are recorded.

#### Data Scraping

Data or web scraping refers to the process of importing information from a web page, typically written in HTML or XHTML, into a locally saved spreadsheet. A Python program was written to extract the required statistics from online tables and is stored locally as a CSV file. An example of the data being aggregated is depicted in Table 1.10.1 below and is stored locally.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TEAM | DATE | MATCHUP | W/L | MIN | PTS | +/- |
| GSW | 10/17/2017 | GSW vs. HOU | L | 241 | 121 | -1 |
| HOU | 10/17/2017 | HOU @ GSW | W | 239 | 122 | 1 |
| BOS | 10/17/2017 | BOS @ CLE | L | 241 | 99 | -3 |
| CLE | 10/17/2017 | CLE vs. BOS | W | 239 | 102 | 3 |
| ATL | 10/18/2017 | ATL @ DAL | W | 241 | 117 | 6 |
| DET | 10/18/2017 | DET vs. CHA | W | 239 | 102 | 12 |
| NOP | 10/18/2017 | NOP @ MEM | L | 240 | 91 | -12 |
| SAC | 10/18/2017 | SAC vs. HOU | L | 242 | 100 | -5 |
| HOU | 10/18/2017 | HOU @ SAC | W | 240 | 105 | 5 |

Table 1.10.1 The Scraped Data being Collected

#### Metric Calculation

Here, the plus-minus score is calculated using the same mathematical formula as in (1.2.1). The two teams' respective effective strengths are computed using (1.5.5). The winning team is the one with the highest strength.

## Random Forest Approach

The objective is to create predictive models that can predict if the home team will win an NBA regular season basketball game. After that, the models' performance will be assessed. The 2017–18, 2018–19, and 2020–21 NBA season's data were used.

### Dataset

The required dataset was obtained from Basketball reference [21] and the NBA website [8] and stored locally as a CSV file. The source is extensive and consists of vast statistics related to player and team statistics over a 40-year period. The chosen dataset consists of regular season standings and regular season results for 2018-2019, 2019-2020 and 2020-2021. An excerpt of the dataset being considered is displayed in Table 2.1.1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Date | Start (ET) | Visitor/Neutral | PTS | Home/Neutral | PTS |
| Tue, Oct 22, 2019 | 8:00p | New Orleans Pelicans | 122 | Toronto Raptors | 130 |
| Tue, Oct 22, 2019 | 10:30p | Los Angeles Lakers | 102 | Los Angeles Clippers | 112 |
| Wed, Oct 23, 2019 | 7:00p | Chicago Bulls | 125 | Charlotte Hornets | 126 |
| Wed, Oct 23, 2019 | 7:00p | Detroit Pistons | 119 | Indiana Pacers | 110 |

Table 2.1.1

### Data Cleaning

Data cleaning is the process of eliminating or changing data that is inaccurate, lacking, unnecessary, duplicated, or formatted incorrectly in order to prepare it for analysis [24]. For this particular case, games that ended after regulation time were left blank in the overtime column, they value “RT” was added to these rows.

### Pandas

The most often used open-source Python library for data science, data analysis, and machine learning activities is called Pandas. It is constructed on top of NumPy, a different package that supports multi-dimensional arrays. Pandas is one of the most widely used data wrangling tools, and it normally comes with every Python distribution. Pandas integrates nicely with many other data science modules in the Python ecosystem [22].

### Scikit-Learn

The most effective and reliable Python machine learning library is called Skearn (Skit-Learn). Through a Python consistency interface, it offers a variety of effective tools for statistical modelling and machine learning, including classification, regression, clustering, and dimensionality reduction. This library is based on NumPy, SciPy, and Matplotlib and was written primarily in Python [23].

### Baseline

A random prediction of all games results in rough accuracy of 50% correct predictions. A more accurate baseline in sports is home win percentage. In most sports the home team has a higher chance of winning a match as is depicted in Fig 2.4.1. In order for the model to be practical it has to have an accuracy rate greater than the baseline.

Fig 2.4.1

### Basic Classification with Decsion Tree

Two additional features are added which are checks to see if the home team and the visitor team won their last game. The addition of these features improves the F1 score slightly as we can see in Fig 2.5.1.



Fig 2.5.1

### --

## Logistic Regression

In order to forecast the outcome of an NBA game, the model uses eight variables that were taken from the league's website [8] and basketball reference [21]. To guarantee that pace has no bearing on the forecasts, each stat is converted to per 100 possessions. The factors being considered are:

### Variables:

#### Home Team

As noted above, the most accurate baseline in most sports is home team win rate as home court advantage plays a major role in match outcomes. [25].

#### Win Percentage

#### Rebounds

#### Turnovers

#### Plus-Minus

#### Offensive Rating

#### Defensive Rating

### True Shooting Percentage

# Results

## Elo Algorithm

## Random Forest Classifier

## Logistic Regression

# Discussion

The Elo algorithm is a widely used rating system due to its simplicity and the fact that it offers relatively high prediction accuracy. The main reason it was selected is because the algorithm considers the whole team as a fundamental unit. In the models being discussed above, a modified version of the Elo algorithm is used where a team’s performance is modelled using the +/- metric of individual players. Individual player ratings are combined to obtain a team rating which is then used to predict the outcome of matches.

# Future Work

## Elo Algorithm

Analysis of test results indicate that further modifications to the algorithm may result in further understanding of individual player strengths and susceptibility to injuries. This may be able to aid front office management when it comes to player transfers and contract extensions.

## Random Forest Regression

The analysis above can be improved by incorporating match odds from an external source. This should allow a user to analyze the outcome and place accurate bets. The result can be further improved by incorporating per-player data such as offensive and defensive ratings.

## Logistic Regression

##### References

1. Franks, Alexander M., D’Amour, Alexander, Cervone, Daniel and Bornn, Luke. “Meta-analytics: tools for understanding the statistical properties of sports metrics” Journal of Quantitative Analysis in Sports, vol. 12, no. 4, 2016, pp. 151-165. DOI: <https://doi.org/10.1515/jqas-2016-0098>.
2. Elo, Arpad E. “The Rating of Chessplayers, Past and Present. New York: Arco Pub, ” 1978. Print.
3. Hao Ji, Erich O’Saben, Adam Boudion, and Yaohang Li. “March Madness Prediction : A Matrix Completion Approach” 2015.
4. Marveldoss, Richard Einstein Doss. “An Elo-Based Approach to Model Team Players and Predict the Outcome of Games, “ August 2018 unpublished. URI: <https://hdl.handle.net/1969.1/173956>.
5. Song Yan, Siyuan Meng, Qiwei Liu, Jing Li. “Design and Implementation of NBA Playoff Prediction Method Based on Elo Algorithm and Graph Database, ” November 2019. ISSN 23275219. DOI: <https://doi.org/10.4236/jcc.2019.711004>.
6. Raquel YS Aoki, Renato M Assuncao, Pedro OS Vaz de MElo. “Luck is Hard to Beat: The Difficulty of Sports Prediction” June 2017. DOI: <https://doi.org/10.48550/arXiv.1706.02447>.
7. Keshri, Suraj Kumar. “Essays in Basketball Analytics, ” September 2019, unpublished. DOI: <https://doi.org/10.7916/d8-1ghx-zy51>.
8. National Basketball Association (2022).Player Box Score Search [online].

Availabile at: [https://www.nba.com/stats/search/player-game/?CF=PTS\*gt\*40](https://www.nba.com/stats/search/player-game/?CF=PTS*gt*40) [Accessed June 2022].

1. Five Thirty Eight (2015). How We Calculate NBA Elo Ratings [online].

Available at: <https://fivethirtyeight.com/features/how-we-calculate-nba-elo-ratings/> [Accessed June 2022].

1. Towards Data Science (2021). Introduction to Random Forest Classifiers from sklearn [online].

Available at: <https://towardsdatascience.com/introduction-to-random-forest-classifiers-9a3b8d8d3fa7> [Accessed July 2022]

1. Cantors Paradise (2019). The Mathematics of Elo Ratings [online].

Available at: <https://www.cantorsparadise.com/the-mathematics-of-elo-ratings-b6bfc9ca1dba> [Accessed July 2022]

1. Geeks for Geeks (2022). Decision Tree [online].

Available at:

[https://www.geeksforgeeks.org/decision tree/](https://www.geeksforgeeks.org/decision%20tree/) [Accessed July 2022]

1. Towards Data Science (2019). Understanding Random Forest [online]. Available at: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2> [Accessed July 2022]
2. Data Driven Investor (2020). Random Forest: Pros and Cons [online]. Available at: <https://medium.datadriveninvestor.com/random-forest-pros-and-cons-c1c42fb64f04> [Accessed July 2022]
3. IBM. What is logistic regression? [online]. Available at: <https://www.ibm.com/uk-en/topics/logistic-regression> [Accessed July 2022]
4. Towards Data Science (2021). How to Predict NBA Double-Doubles [online]. Available at: <https://towardsdatascience.com/how-to-predict-nba-double-doubles-f4c30be08ca0> [Accessed July 2022]
5. Geeks for Geeks (2020). Advantages and Disadvantages of Logistic Regression [online]. Available at: <https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/> [Accessed July 2022]
6. Campbell, Z. (2020). Development of a logistic regression model to predict the outcome of NBA games. DOI: <https://doi.org/10.15786/13701274.v3> [Accessed: August 2022].
7. Ghimire S, Ehrlich JA, Sanders SD (2020) Measuring individual worker output in a complementary team setting: Does regularized adjusted plus minus isolate individual NBA player contributions?. DOI: <https://doi.org/10.1371/journal.pone.0237920>
8. David, H.A., (1963). The method of paired comparisons (Vol. 12, p. 120). London.
9. Basketball Refernce (2022). Expanded Standings [online]. Available at: <https://www.basketball-reference.com/leagues/NBA_2020_standings.html> [Accessed: August 2022]
10. Active State (2021). What Is Pandas in Python? Everything You Need to Know [online]. Available at: <https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/> [Accessed: August 2022]
11. Tutorials Point (2022). Scikit Learn – Introduction [online]. Available at: [https://www.tutorialspoint.com/scikit\_learn/scikit\_learn\_introduction.htm#](https://www.tutorialspoint.com/scikit_learn/scikit_learn_introduction.htm) [Accessed: August 2022]
12. Obviously.ai (2022). Data Cleaning: The Most Important Step in Machine Learning [online]. Available at: <https://www.obviously.ai/post/data-cleaning-in-machine-learning> [Accessed: July 2022]
13. Jones Marshall B, 2007. "Home Advantage in the NBA as a Game-Long Process," Journal of Quantitative Analysis in Sports, De Gruyter, vol. 3(4), pages 1-16, October. Available at: <https://ideas.repec.org/a/bpj/jqsprt/v3y2007i4n2.html>
14. G