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

Thomas Abraham
210639757
Hazar 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.

I. 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.

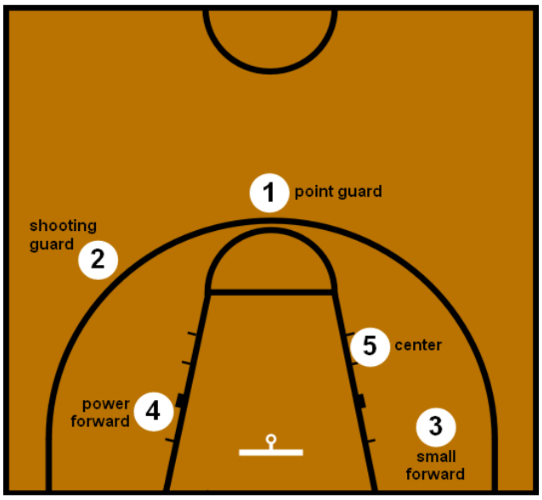

Figure 1.1. Traditional positions in basketball

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 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 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 evaluates the results of matches and values the strength of a team in comparison to other teams. Using this data, teams and coaches may then decide how to construct their teams, trade for players, rotate their roster, and draft future talent. [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.

II. 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.

A. *Elo Algorithm:*

According to a fundamental premise of the Elo model, a metric's treatment of variation must coincide with the decision being made for it to be useful for that decision [1]. Even though we can separate some player, season, and team variation through data analysis and division, all measures we do are affected by chance's variability. 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]." This is the rationale behind modelling each player's contribution as a regularly distributed random variable [4].

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 to create a team rating, the player ratings are added together. The chance of each team winning during the season is then determined by comparing team ratings pairwise [4]. 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 comparing them to actual information from prior 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.

B. *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 does [14].

C. *Logistic Regression:*

This 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].

III. METHODOLOGY

A. Elo Approach

The Elo algorithm was developed initially to provide a useful chess player ranking system. 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 is made to monitor each basketball player's performance and aggregate their ratings to provide a team score that can be used to the simulations [4].

1) 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 considers 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.

2) Player Strength

Let p be the variable used to represent the number of points an NBA player contributes every minute. P thus evaluates a player's strength. Each member of a team is initialized with a p value of 1000 (normalizing constant a is multiplied) to make it easier to understand the data. An appropriate value of a is obtained empirically [4].

$$1000 = a \times p \tag{3.1.1}$$

3) 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 $Team_i$ plays, $Team_j$ then the rating is updated as [3]:

$$R_{i_{new}} = R_{i_{old}} + K(S_{ij} - x_{ij}) \tag{3.1.2}$$

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.

4) Actual Score (S)

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

The definition of S_{ij} is depicted as:

$$S_{ij} = \begin{cases} 1, if\ Team_i\ beats\ Team_j \\ 0, if\ Team_j\ beats\ Team_i \end{cases} \quad [3]$$

5) Expected Score (x)

Variable x_{ij} is used to denote the expected outcome of a match between $Team_i$ and $Team_j$. 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 $Team_i$ wins against $Team_j$ is depicted as [4]:

$$P_r(i > j) = \frac{P_i}{P_i + P_j} \tag{3.1.3}$$

Where P_i and P_j are individual scores assigned to $Team_i$ and $Team_j$. The expression becomes a logistic function when an exponential score is considered [4]:

$$P_r(i > j) = \frac{e^{r_i}}{e^{r_i} + e^{r_j}} \tag{3.1.4}$$

Where r_i and r_j are the ratings of $Team_i$ and $Team_j$. The standard Elo algorithm is depicted as:

$$P_r(i > j) = \frac{1}{1 + 10^{\frac{r_j - r_i}{400}}} \tag{3.1.5}$$

Where 400 is the constant scale factor. Let x_{ij} be used to denote $P_r(i > j)$. Therefore x_{ij} is denoted as:

$$x_{ij} = \frac{1}{1 + 10^{\frac{r_j - r_i}{400}}} \tag{3.1.6}$$

A model is developed to predict a player's plus-minus score. Every time two teams compete; the individual player strengths are added up to provide a team's combined strength parameter [4]. The strength of $Team_i$ is depicted as:

$$m_i = \frac{\sum_{n=1}^{N} t_{in} p_{in}}{\sum_{n=1}^{N} t_{in}} \tag{3.1.7}$$

Where t_{in} denotes the minutes played by the nth player on the ith and p_{in} denotes the estimated strength of the nth player on $Team_i$. Furthermore, m_i denotes the average points scored per minute by $Team_i$ [4].

6) 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%.

7) Match Outcome

Utilizing (3.1.7), the overall team rating is obtained based on the strengths of the individual players on the team. Consider that the team ratings of $Team_i$ and $Team_j$ are denoted as r_i and r_j respectively. The probability that $Team_i$ wins a matchup against $Team_j$ is derived using (1.5.3). For the simulations being considered in these particular instances, a win is predicted if $Team_i$ has a higher overall rating than $Team_j$.

8) 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.

9) Algorithms

a) 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 $Team_i$ and $Team_j$ do

Calculate x_{ij} and x_{ji} which corresponds to the probability of $Team_i$ and $Team_j$ winning respectively

Update rating for $Team_i$ and $Team_j$ according to equation 3.1.2

end for

b) Predict Match Winner

Algorithm 2: Predict the winner of match between $Team_i$ and $Team_j$

Determine the functional strength of $Team_i$ based on (3.1.7)

Determine the functional strength of $Team_j$ based on (3.1.7)

It is anticipated that the team with the better strength of the two will prevail.

10) Datasets

a) 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.

b) 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. To extract the necessary statistics from online tables, a Python programme was created, and the results were saved locally as a CSV file. An example of the data being aggregated is depicted in Table 3.10.1 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 3.10.1 The Scraped Data being Collected

c) 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 (3.1.7). The winning team is the one with the highest strength.

B. 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.

1) 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

2) 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.

3) 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].

4) Scikit-Learn

Skearn is the name of the most efficient and dependable Python machine learning package (Skit-Learn). It provides several efficient tools for statistical modelling and machine learning, such as classification, regression, clustering, and dimensionality reduction, through a Python consistency interface. This library was mostly created in Python and is based on NumPy, SciPy, and Matplotlib [23].

5) 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.

```
In [14]: # Baseline Win % for Home Teams

         n_games = df['Home Win'].count()
         n_homewins = df['Home Win'].sum()
         win_percentage = n_homewins / n_games

         print('Home Team Win Rate : {0:.2f}%'.format(100 * win_percentage))

         Home Team Win Rate : 54.96%

In [15]: # Predicting the baseline for Home Teams with a classifier

         from sklearn.metrics import f1_score

         y_pred = [1] * len(y_true)
         print('F1 : {0:.4f}%'.format(f1_score(y_true, y_pred, pos_label = None, average = 'weighted') * 100))

         F1 : 38.9848%
```

Fig 2.4.1

6) 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.

```
In [19]: # Basic Decision Tree Classifier

         from sklearn.tree import DecisionTreeClassifier

         # Remove random_state to get non-replicable results

         dtc = DecisionTreeClassifier(random_state = 14)

In [20]: from sklearn.model_selection import cross_val_score

         # Use selected features as target
         X_previouswins = df[['Home Last Win', 'Visitor Last Win']].values

         # Decision Tree Classifier
         dtc = DecisionTreeClassifier(random_state = 14)
         scores = cross_val_score(dtc, X_previouswins, y_true, scoring = scorer)

         # Display Outcome
         print("Using The Last Result from Home and Visitor Teams : \n")
         print('F1 : {0:.4f}%'.format(np.mean(scores) * 100))

         Using The Last Result from Home and Visitor Teams :

         F1 : 38.9849%
```

Fig 2.5.1

C. Logistic Regression

In order to forecast the outcome of an NBA game, the model uses eight variables that were scraped from the league's website [8]. To guarantee that pace has no bearing on the forecasts, each stat is converted to per 100 possessions. For the sake of visibility, it is also possible to see predictions for a single day period along with a past set of dates. The factors being considered are:

1) Variables

a) 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. The NBA's home court advantage lends itself incredibly nicely to study. The impact is significant. The average point differential between home and away teams is usually around 3.5 points, and the home side typically wins roughly 60% of the games [25].

b) Win Percentage

Win percentage is obtained by multiplying the teams current win-loss record by a 100. The calculated metric is then utilized to calculate the effective probability of the outcome of a particular matchup.

c) Rebounds

An effective center can position themselves optimally either to retrieve a rebound or in order to box out an opponent which will then allow their teammate to obtain the rebound. The offensive and defensive rebounds gained and lost each possession directly influence the number of points score. Hence, making them a crucial metric.

d) Turnovers

A turnover cost your team the opportunity to obtain a shot at the hoop since it gives the opposition another possession, which they can turn into another shooting attempt. The team with more possessions will score more points if we assume that all the other specified elements are fairly equal. If a team has a lot of turnovers, it will be seen when it is near to a team that shoots at a higher % and has a low shooting percentage. Turnovers become a significant metric as a result.

e) Plus-Minus

As mentioned in an earlier section, the plus-minus rating shows how a team fared with a certain player on the floor. This measure is frequently employed for thorough player evaluation before a major matchup. A player's contribution to their team's success while they are on the court is shown by a good score, and the opposite is also true [19].

f) Offensive Rating

The number of points a player score for every 100 total individual possessions is known as individual offensive rating. Individual Total Possessions and Individual Points Produced serve as the fundamental building elements in the calculation of the Offensive Rating [21].

g) Defensive Rating

The amount of points a player concedes per 100 total individual possessions is known as their individual defensive rating. The idea of the individual Defensive Stop is the basis of the calculation of Defensive Rating. Stops include both an estimate of the number of forced turnovers and forced misses by the player that aren't recorded by steals and blocks, as well as instances of a player terminating an opponent's possession that are marked in the box score (blocks, steals, and defensive rebounds).

h) True Shooting Percentage

The shooting percentage, which calculates a player's shooting efficiency by adjusting it for three-pointers and free throws, is known as the true shooting percentage. True shooting percentage is calculated by dividing the number of field goals and free throws attempted by the total number of points scored.

2) Dataset

The required dataset was acquired from the NBA's website [8]. The factors mentioned above are collected from the 2018-2019, 2019-2020 and 2020-2021 seasons during the execution of the model. The data was obtained by scraping it directly from the website.

IV. Results

All the above-mentioned models were executed on data pertaining to the NBA 2018-2019 season and the following results were observed. The datasets required were obtained from the NBA's official website [8] and Basketball Reference [21].

A. Elo Algorithm

Team	Team Rating
Atlanta Hawks	988
Boston Celtics	1111
Brooklyn Nets	855
Charlotte Hornets	930
Chicago Bulls	1010
Cleveland Cavaliers	1008
Dallas Mavericks	931
Denver Nuggets	1042
Detroit Pistons	918
Golden State Warriors	1244
Houston Rockets	1097
Indiana Pacers	1019
Los Angeles Clippers	1108
Los Angeles Lakers	890
Memphis Grizzlies	956
Miami Heat	1071
Milwaukee Bucks	1031
Minnesota Timberwolves	915
New Orleans Pelicans	979
New York Knicks	879
Oklahoma City Thunder	1057
Orlando Magic	886
Philadelphia 76ers	846
Phoenix Suns	824
Portland Trailblazers	1072
Sacramento Kings	923
San Antonio Spurs	1131
Toronto Raptors	1100
Utah Jazz	1118
Washington Wizards	1068

Table 4.1.1 Elo ratings at the end of the regular season

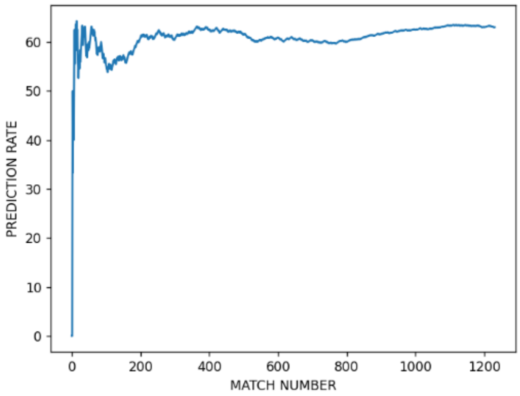

Fig 4.1.1 Graph depicting the prediction rate using the Elo algorithm

B. *Random Forest Classifier*

```
In [47]: # Random Forest Classifier
         rfc = RandomForestClassifier(random_state = 14)
         scores = cross_val_score(rfc, X_all, y_true, scoring = scorer)

         # Display Outcome
         print("F1 : {0:.4f}%".format(np.mean(scores) * 100))

         F1 : 60.3828%

In [48]: # Attempt to improve accuracy using more detailed features
         # Adjusting the parameters of the algorithm using GridSearchCV

         parameter_space = {
             "max_features": [2, 10, 50, 'auto'],
             "n_estimators": [50, 100, 200],
             "criterion": ["gini", "entropy"],
             "min_samples_leaf": [1, 2, 4, 6],
         }

         # Random Forest Classifier
         rfc = RandomForestClassifier(random_state = 14)
         grid = GridSearchCV(rfc, parameter_space, scoring = scorer)
         grid.fit(X_all, y_true)

         # Display Outcome
         print("F1 : {0:.4f}%".format(grid.best_score_ * 100))
         print(grid.best_estimator_)

         F1 : 63.0377%
         RandomForestClassifier(criterion='entropy', max_features=50, min_samples_leaf=2,
                                n_estimators=50, random_state=14)
```

Fig 4.2.1

C. *Logistic Regression*

D. *Final Analysis*

Algorithm/ Model	Correct Predictions	Total Number of Games	Accuracy Rate
Elo Algorithm	779	1230	63.33%
Random Forest Classifier	775	1230	63%
Logistic Regression			%

Table 4.4.1 Final Results for all three models

V. 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 when the main metric being considered is the plus-minus score. Individual player ratings are combined to obtain a team rating which is then used to predict the outcome of matches.

Table 4.4.1 indicates that the Elo algorithm performed relatively poorly when compared to the other two models. The most likely reason that the random forest and logistic regression approach outperformed the Elo approach is because of the number of variables or factors being considered in each model's analysis. The random forest approach considered multiple records from the league standings as well as the regular season results, the logistic regression approach considered eight unique factors during its execution, whereas the Elo algorithm

Overall, the hypothesis stated during the introduction of this paper appears incorrect as the logistic regression approach produced the most accurate result and not the Elo approach as originally anticipated.

VI. Future Work

A. *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.

B. *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.

C. *Logistic Regression*

The same elements affect each NBA player differently and have an impact on their game in different ways. Some players perform noticeably better at home or when playing their former squad. Some people don't do well in certain time zones or on the second night of a back-to-back. By virtue of this variance, each NBA club will have its own logistic regression model for determining whether they will prevail in a forthcoming contest. We could add more variables from additional datasets or increase the number of observations by incorporating more players in order to improve accuracy.

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 [Accessed June 2022].

[9] 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].

[10] 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]

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

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

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

Available at:

https://www.geeksforgeeks.org/decision tree/ [Accessed July 2022]

[13] Towards Data Science (2019). Understanding Random Forest [online]. Available at: https://towardsdatascience.com/understanding-random-forest-58381e0602d2 [Accessed July 2022]

[14] 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]

[15] IBM. What is logistic regression? [online]. Available at: https://www.ibm.com/uk-en/topics/logistic-regression [Accessed July 2022]

[16] 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]

[17] 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]

[18] 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].

[19] 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

[20] David, H.A., (1963). The method of paired comparisons (Vol. 12, p. 120). London.

[21] Basketball Refernce (2022). Expanded Standings [online]. Available at: https://www.basketball-reference.com/leagues/NBA_2020_standings.html [Accessed: August 2022]

[22] 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]

[23] Tutorials Point (2022). Scikit Learn – Introduction [online]. Available at: https://www.tutorialspoint.com/scikit_learn/scikit_learn_introduction.htm# [Accessed: August 2022]

[24] 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]

[25] 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

[26] NBA Stuffer (2022). NBA Stats [online]. Available at: https://www.nbastuffer.com/nba-stats/ [Accessed: August 2022]