MSc Project - Reflective Essay

Project Title:	Predicting the final seeds of National Basketball Association teams, a Machine Learning approach
Student Name:	Thomas Abraham
Student Number:	210639757
Supervisor Name:	Hazar Emre Tez
Programme of Study:	MSc Computer Science (FT)

This research sought to analyse the widely available National Basketball Association (NBA) statistics and attempt to predict the final seeds or ranking of each team at the end of the regular season. The main algorithm being considered here is the Elo algorithm which is a popular rating system. It assesses the outcomes of games and values a team's strength in relation to other teams. Teams and coaches can use this information to make decisions about how to put their teams together, trade for players, rotate their roster, and select future prospects. The Elo system does not use an absolute metric to assess performance. It can be deduced based on wins and losses. As you can see in the paragraphs below, I first provide a brief explanation of the rationale behind the chosen strategy before going into detail about the project's advantages and disadvantages in relation to earlier research, as well as potential areas for future study and its practical difficulties and solutions.

Approach:

The data chosen for the required analysis was obtained from the NBA’s official website and Basketball Reference. For the execution of both the Elo algorithm and the Logistic Regression model, the data was scraped directly from the site and stored locally as CSV files. The Random Forest model was executed on Jupyter Notebook and data obtained was exported to the required format directly from Basketball Reference.

The genuine value of a player is not exactly quantifiable in a dynamic sport like basketball where teams and players have different playing styles, hence it cannot be evaluated or analysed. Therefore, we rely on the sport's observable measurements, such points scored, offensive and defensive rebounds, assists, turnovers, and so forth. The plus-minus score is the main statistic taken into account by the Elo approach. The plus-minus rating shows how a team performed collectively when a certain player was on the court. The advanced adjusted plus-minus metric used by the NBA takes into account a player's marginal impact on the score per 100 possessions in comparison to the league average player. Unadjusted plus-minus is typically favoured to adjusted plus-minus because the latter heavily influences each player's score based on his on-court teammates' talent and abilities (Ghimire S. et al.,2020). 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. Each individual players plus-minus score is aggregated to provide a team rating which was then used in the required simulations. One of the major reasons for selecting the Elo algorithm is that it can account for margin of victory. Teams will receive more rating points for victories and fewer rating points for defeats, but they will gain or lose more rating points depending on the margin of victory, i.e., a lopsided win or loss has a higher impact on the team's total rating. This was done by giving each game a multiplier and dividing it by the team's expected margin of victory.

The Random Forest Classifier was selected as a comparative model due to its adaptability, simplicity, and suitability for both classification and regression problems. A decision tree is a structure that resembles a flowchart, in which each node represents a test on an attribute, each branch a possible conclusion, and each terminal node a class label. Each decision tree in this model learns from a random sample of data points that are drawn without replacement, giving it the option for random sampling of data. As a result, the likelihood of overfitting is reduced, and overall predicted accuracy is increased. This methodology is based on the core tenet that "a large number of relatively uncorrelated models working as a group will outperform any of the individual constituent models." The practical objective during the implementation of this approach was to create a predictive model that can foresee if the home team in a particular matchup will win. This was done by taking the teams home win-loss record, road win-loss record and points scored per match. Additional features such as winning streaks and previous match wins were added directly on Jupyter Notebook to further increase accuracy. The data for the 2018-2019, 2019-2020 and 2020-2021 NBA season were exported directly from Basketball Reference. A baseline of home team victories was selected and the python libraries Pandas and Scikit-Learn were utilized for implementation and analysis. The ultimate aim was to obtain an accuracy score higher than the chosen baseline.

The likelihood of an event occurring is estimated by the logistic regression model. The main justification for choosing this model is because it just uses the supplied data to generate a forecast and ignores all other outside influences. Consequently, all bias is taken out of the equation. The approach is an effective choice since it usually yields good accuracy for straightforward data sets and performs well when the dataset can be linearly separated. Eight factors were chosen based on their ability to influence game wins and losses. The factors selected were home court, previous win percentage, rebounds, turnovers, plus-minus, offensive rating, defensive rating, and true shooting percentage. These eight variables were utilised to forecast game outcomes after being directly scraped from the NBA's official website. To guarantee that pace has no bearing on the forecasts, each stat is converted to per 100 possessions.

Finally, all three models were run over past NBA seasons, and they were compared based on overall prediction accuracy. The main aim was to identify an effective machine learning algorithm for NBA match prediction while also taking into account the various features that the models are utilizing and identifying which of these variables is best suited for analysis.

Contribution and Further Work:

The likelihood that a club will succeed before the postseason depends heavily on accurate projections about its seed. Ideally, being able to predict a first-round opponent will allow a team's coaches and trainers to prepare players and create offensive and defensive plays specifically for those opponents. The ability to compare the results of all three models will provide players the chance to get ready for any possible scenario going into the postseason. For this reason, the primary contribution of this project was to identify the most accurate model which can be used for predictive analysis of an NBA season.

This research aimed to develop a model that can precisely forecast each of the fifteen seeds in both conferences for a season by building on earlier work (M. Richard Einstein Doss, 2018) where a modified version of the Elo algorithm was constructed where the plus-minus metric was given the utmost importance. Using all available traditional team statistics, the Elo ratings are calculated and then used to order the teams and thereby identify their individual seed. Then, team ratings are compared pairwise to calculate the likelihood of each team winning during the season. These wins and losses are combined and assigned to each team as a season record, this record is then used to determine the team’s seed. The seed is ultimately used to determine round one playoff matchups.

According to an analysis of the test findings, additional adjustments to the Elo algorithm may lead to a better understanding of each player's capabilities and injury susceptibility. The front office management may benefit from this when it comes to player trades and contract extensions. If time permitted, a more accurate and consistent prediction might be obtained by including more criteria such as offensive rating, defensive rating, and true shooting percentage in the Elo rating calculation. Incorporating match odds from an external source could provide a new dimension to this research, thereby giving it an additional purpose.

In order to increase accuracy, more variables from more datasets may be included if there was more time. Incorporating advanced statistics such as defensive rating, offensive rating and true shooting percentage may generate standings for a season with greater accuracy. Additionally, the models specified can be tested on various leagues such as the National Collegiate Athletic Association and the National Basketball League to ensure its usability.

Limitations and Practical Challenges:

The main constraints of this work were time and computational resources, which made it more difficult to complete follow-up analyses to expand on the above contributions. Additional challenges faced during the execution of the predictive analysis involved the drawbacks of the individual models which will be elaborated on below. Apart from those, recent changes to the python NBA API caused errors during code execution and consequentially required debugging of the code related to all three models.

The Elo algorithm's biggest observable flaw is the possibility of two teams having equal results but different ratings because the ratings are determined as a shift from the present ranking. Practically speaking, it functions as it should because the vast majority of teams advance extremely slowly. The approach, meanwhile, may appear unfair to teams that advance quickly after a poor start.

The Random Forest model has the observable flaw of being difficult to interpret. However, it does not offer total visibility into the coefficients; it only shows feature importance. Large datasets require a lot of work, and the user has limited influence over the model's behaviour.

The presumption of linearity between the independent and dependent variables is the main downside of logistic regression. Furthermore, as logistic regression can only be used to estimate discrete functions, it is confined to discrete number sets. The number of observations should always be higher than the number of features to prevent overfitting, which hinders the model from correctly predicting new data since it is unable to distinguish between noise and essential information.

Legal, Social and Ethical Challenges:

There are no known legal, social, or ethical challenges as all required metrics were obtained either from the National Basketball Association’s official website or Basketball-Reference. This data is available freely to the public as is stated in the Terms of Use for both the NBA [8] and Basketball-Reference [9].

References:

[1] 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

[2] 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].

[3] Geeks for Geeks (2022). Decision Tree [online]. Available at: https://www.geeksforgeeks.org/decision tree/ [Accessed July 2022]

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

[5] 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

[6] 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]

[7] 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]

[8] National Basketball Association (2021). Terms of Use [online]. Available at: https://www.nba.com/termsofuse [Accessed August 2022]

[9] Sports Reference (2020). Sports Reference Terms of Use [online]. Available at: https://www.sports-reference.com/termsofuse.html [Accessed August 2022]