**MSc Project - Reflective Essay**

|  |  |
| --- | --- |
| **Project Title:** | Predicting the final seeds of National Basketball Association teams, an Elo based 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. In the conclusion, I take a broad look at the project's success in the context of my technological growth.

**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 Logistic Regression

**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.

**Limitations and practical challenges:**

**Technical Development:**

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

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