MINI PROJECT REPORT

NBA Basketball Game Forecasting & Visualization

Submitted in partial fulfilment of the requirements for the degree of

Bachelor of Engineering

in

Computer Science and Business System

Submitted By

Arun Upreti	**101918049**	Arun
Shivam Munjal	**101918055**	Shivam
Omkar Arora	**101918070**	Omkar
Prince Garg	**101918075**	Prince

Under the supervision of:

Dr. Rupali Bhardwaj

Assistant Professor, TIET

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY

PATIALA – 147004

December 2021

Table of Contents

1. Introduction

Basketball is a team sport in which two teams, most commonly of five players each, opposing one another on a rectangular court, compete with the primary objective of shooting a basketball through the defender's hoop while preventing the opposing team from shooting through their own hoop.

No doubt, basketball is one of the most followed game on the planet and one name that surely strikes our mind when we talk about basketball is, NBA.

National Basketball Association (NBA) is a professional basketball league formed in the United States in 1949 by the merger of two rival organizations, the National Basketball League (founded 1937) and the Basketball Association of America (founded 1946). At present, there are 30 teams taking part and more than 4000 players have played for the biggest league. So, we decided to blend the craze for the game with some analysis models and visualization techniques to choose as a topic for our project.

2. Background

Professional basketball has looked very different over the past ten years thanks to the use of data analytics. It is almost as if these Analytics models have inputted the conventional NBA and outputted an entirely new game. A game where every team now has at least one data analyst. A game where every move is calculated with the aim to optimize for a winning strategy.

Data has become front and center for NBA games where almost every decision is now based on analytics. This data is being analysed by machine learning models to recommend the winning strategy accordingly and helps the team perform efficiently.

Teams have been using high-tech analytics mainly in three ways:

- Designing winning strategies
- Predicting and avoiding player injury
- Scouting

As basketball enthusiasts we follow this game often, and this motivated us in selecting this topic for our mini project. The knowledge about the game and the statistical measures used, really helped us in understanding the problem statement and building the machine learning model.

3. Objectives

- The goal of this project is to predict upcoming NBA game result based on team stats and ELO Ratings. On our research we found that the already built predicting models were using different features like FGA, FGM, 3ptPct, 3pA etc (other than Elo rating and recent team performances) have low accuracy. So we are looking to develop a model that used ELO ratings as well as other features to predict the game outcome as the win % depends on the individual team performance as well

- To compare the players more easily based on their past performances we will use different data visualizing techniques using tableau. By this we can sort or filter players and rank them according to their various features.

4. Methodology

Tools used:

- Google Colab
- Python
- Tableau Prep Builder - Data Cleaning and Pipelining
- Tableau Desktop - Data Analytics and Visualizations
- Microsoft Excel - Data Preperation

Libraries used:

- Scikit Learn
- Seaborn
- Matplotlib
- Pandas
- Numpy

Following steps were followed sequentially while developing the project.

4.1 Collection of Data

We downloaded the data of 30 teams and various NBA players of season from 2003-2004 to the 2020- 2021 from the website.

- Basketball Reference
- Synergysportstech
- NBA.com
- Kaggle.com

4.2 Cleaning of Dataset

We read the dataset and combined it into large dataframe which contained all the team stats and player stats for the past seasons.

As the dataset was very large and included many features, we had to study and clean the dataset before using it in the machine learning model.

We used Tableau Prep that provided various cleaning operations that were used to clean and shape the Player Stats dataset. We cleaned the data by applying cleaning operations such as filtering, adding, renaming, splitting, grouping, or removing fields. Cleaning up dirty data made it easier to combine and analyze our dataset.

We used Python libraries like Pandas and numpy to clean the team stats dataset like dropping unnecessary columns, removing null values, and improving column formatting/labels.

4.3 Feature Engineering

We need some reliable features to make predictions about future games. Our primary goal was to make all of the available data understandable. For example - Game-by-game rebounds of an entire team don't help us much unless we can use that data in a higher-level analysis that leads us to our ultimate goal – predicting wins and losses. To that end, we sought to create 2 different features which we would use in understanding how our teams progressed and regressed throughout each season. To obtain the results we created some features.

4.3.1 Elo Ratings

The first feature we will add is Elo ratings. The Elo rating system can be used to calculate the relative skill and quality of teams in a league. Each team will start with the same Elo rating and it will be updated with each game a team plays. Thus, for each game in the team stats data that we collected, we will have the Elo Rating of the home and away team. Elo ratings are a little more sophisticated than simple win loss percentage so hopefully it will help us achieve better model overall.

The exact formula is as follows:

The present Elo rating of a team is *R_i*, then the Elo rating after its next game is calculated using formula:

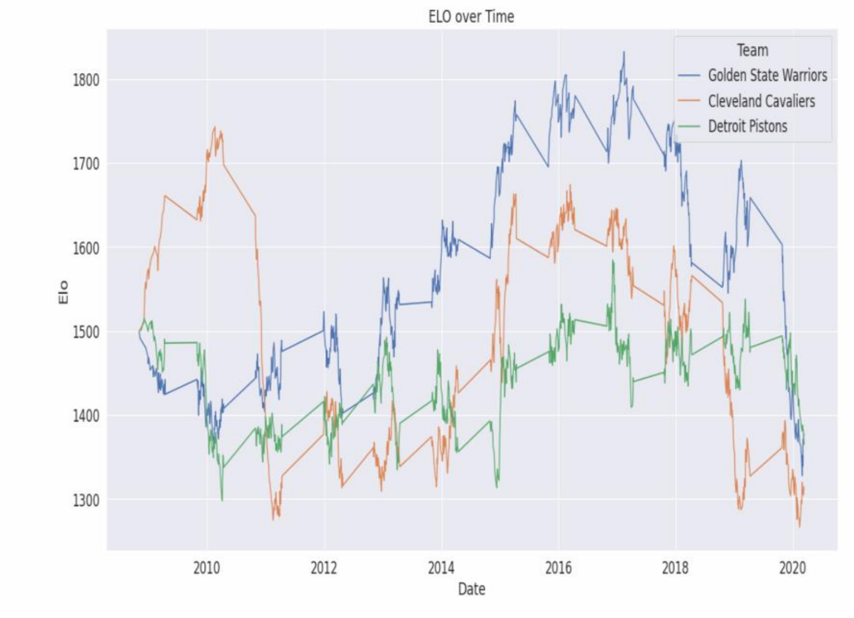
$$\mathbf{R_{i+1} = R_i + K * (S_{team} - E_{team})}$$

Here, state variable is S_{team} : 1 for the team wins, 0 for the team loses. The expected win probability of the team is E_{team}, which is represented as:

$$E_{team} = \frac{1}{1 + 10^{\frac{opp\_elo - team\_elo}{400}}}$$

k is a constant which changes continuously, which depends on both the margin of victory(Difference between points of winning and losing team) and difference in Elo ratings:

$$k = 20\frac{(MOV_{winner} + 3)^{0.8}}{7.5 + 0.006(elo\_difference_{winner})}$$



4.3.2 Recent Team Performance (Avg. stats over 10 most recent games)

How a team has been performing recently is likely a good indicator of how they will perform in an upcoming game. Thus, for every game we have in the team stats data that we will scraped, we will keep track of the rolling averages of each teams stats over their previous 10 games.

4.3.3 Recent Player Performance

Likewise, how a player has been performing in recent games will likely give us a good idea of how well that player will perform in an upcoming game. For each entry in the player stats data that we collected, we made visualizations on their stats. If we know how well a player will perform, we can translate that information into how well a team will pick players for a season.

4.4 Machine Learning Models Used

4.4.1 Logistic Regression Classifier Model - It is a supervised learning classification algorithm which is used to predict observations to a discrete set of classes. Practically, it is used to classify observations into different categories. Hence, its output is discrete in nature. It is one of the most simple, straightforward and versatile classification algorithms which is used to solve classification problems.it is the classification algorithm and it is used for supervised learning classification problems. Logistic Regression models categorizes predictions into binary results(1 for team win and 0 for team loss). As we are predicting match result, so this this model is helpful to us in prediction.

4.4.2 Random Forest Classifier Model - Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. We used this classifier efficiently for both regressions and classifications., this Classifier built a proper decision tree to predict results from the available data of team.

To get more accurate results we did the hyperparameter tuning of the above models, firstly by Random Search CV and then by Grid Search CV

5. Observations and Findings

We trained our features using Logistic Regression Classifier And Random Forest Classifier with Randomized search CV and Grid Search CV and the following were the accuracies obtained

Accuracy of models:

- logistic regression classifier –65.8%
- logistic regression classifier with hyperparameter tuning using Random Search CV-66.8%
- logistic regression classifier with hyperparameter tuning using Grid Search CV-66.8%
- Random Forest classifier –65.8%
- Random Forest classifier with hyperparameter tuning using Random Search CV-66.8%
- Random Forest classifier with hyperparameter tuning using Grid Search CV-65.4%
  After compiling the program for Logistic Regression and Random Forest Classifier on the available team data, we get accuracy of **65.4%–66.8%** for win prediction. The Highest testing accuracy is **66.8%** and it is achieved by using **Random Forest Classifier Model with hyperparameter tuning using Random Search CV**.
- By building visualization on **tableau**, we are able to sort and filter players according to their various features like number of matches played, accuracy of goals etc.

6. Limitations

We developed the models to predict the wins and losses on basis of the team stats ignoring the individual player performance. Our model is biased towards some factors like player injuries, unavailability, newer drafts and roster. We are also assuming that the ELO ratings of the players will gradually increase or decrease. No sudden changes will occur over time, whereas in the real-world scenario, there are chances of sudden change in the ELO due to player transfer, retirement, age factor etc.

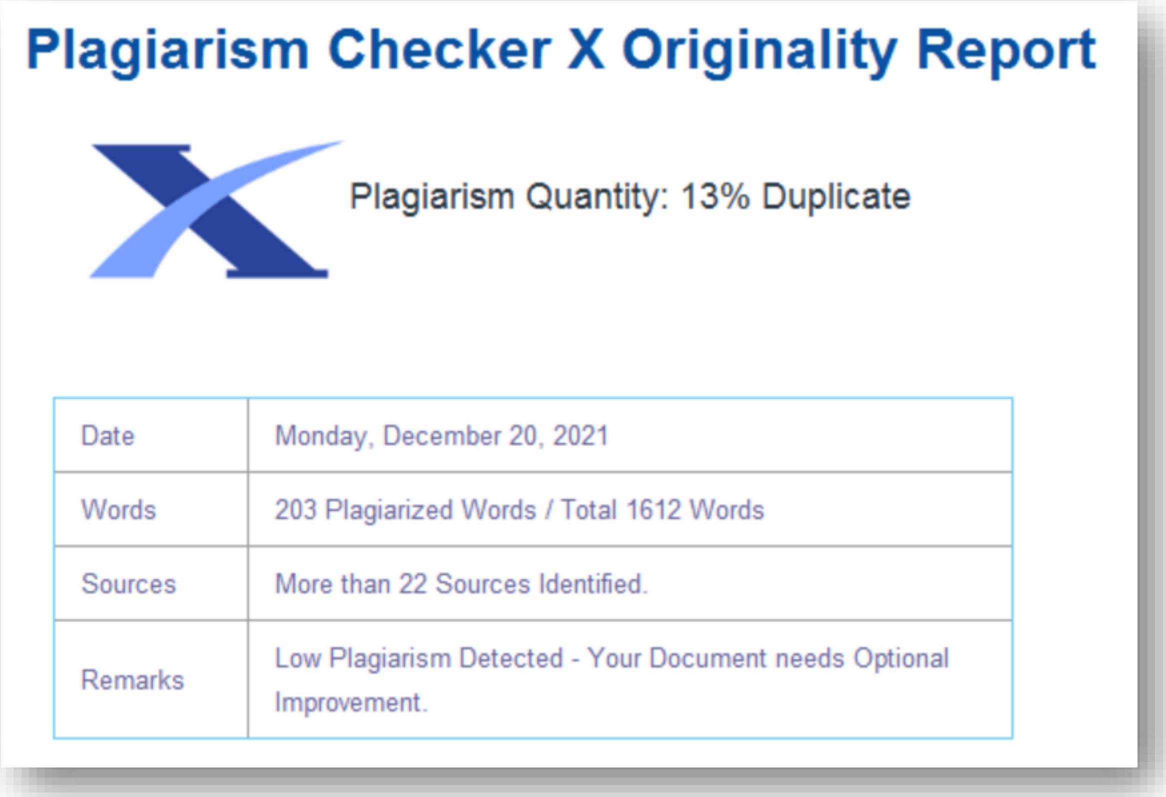
7. Conclusions and Future Work

The goal was to see if it was possible to predict game result based on just previous game result. Both models predicting the result achieved good results given that the team scores vary at around. The results were affected by various factors. Team rosters/coaches change between seasons so the team's scoring is not consistent from season to season. Our model works on assumption that players and team staff do not varies much between consecutive seasons. For future work, given more features like players rating, to train the models, the results could be improved.

8. Bibliography

- https://www.basketball-reference.com/
- https://www.synergysportstech.com/Synergy
- https://www.nba.com/
- https://www.kaggle.com/
- https://towardsdatascience.com/understanding-random-forest-58381e0602d2
- https://www.researchgate.net/publication/308516324_A_Rating_System_For_Gaelic_Foot
- https://www.coursehero.com/file/52858842/footballdocx/
- https://homepages.cae.wisc.edu/~ece539/fall13/project/AmorimTorres_rpt.pdf

9. Plagiarism Report



Plagiarism Checker X Originality Report

Plagiarism Quantity: 13% Duplicate

Date	Monday, December 20, 2021
Words	203 Plagiarized Words / Total 1612 Words
Sources	More than 22 Sources Identified.
Remarks	Low Plagiarism Detected - Your Document needs Optional Improvement.