Basketball win prediction using machine learning

Utsav Lal
ualal@ncsu.edu
North Carolina State University
Raleigh, NC, USA

Swaraj Kaondal
skaonda@ncsu.edu
North Carolina State University
Raleigh, NC, USA

ABSTRACT

We suggest a method for forecasting the outcome of basketball matches between two teams composed of randomly selected players, employing machine learning algorithms. Our plan involves leveraging the CMU NBA dataset along with match-by-match NBA data obtained from Kaggle. Through analyzing this data, our objective is to uncover compelling statistics and ultimately present our discoveries in a comprehensive final project report by the semester's end.

1 DATA DESCRIPTION

1.1 Introduction

We have access to data spanning from 1946 to 2004 for both the NBA and ABA basketball leagues. This dataset encompasses various attributes of players across three different types of basketball matches within these leagues, which are as follows:

- Regular Season: The regular season in basketball consists of a series of scheduled games where teams compete to accumulate wins and secure a spot in the playoffs.
- Playoffs: The playoffs are a series of elimination games following the regular season, where the top teams from each conference compete in a tournament-style format to determine the league champion.
- All-Star Game: The All-Star game is an exhibition match held midway through the season, featuring the league's top players as selected by fan, player, and media voting, providing a showcase of talent and entertainment for fans.

1.2 Data Source and Description

1.2.1 NBA Statistics data, Carnegie Mellon University. [2]

(1) Players table - The master list of all players.
(2) Player regular season table: The stats of each player for each season.
(3) Player regular season career table: Cumulative stats of each player during their career.
(4) Player playoffs table: The stats of players who have played in playoffs
(5) Player playoffs career table: The cumulative stats of players who have played in playoffs.
(6) Player allstar table: The stats of players for all star matches
(7) Teams table: Master list of all teams in NBA and ABA
(8) Team season table: The stats of whole team over a season for each season
(9) Draft table: The list of players drafted
(10) Coaches season table: The stats of coaches for a season for all season
(11) Coaches career table: The stats of coaches aggregated for all season in their career.

1.2.2 NBA Dataset, Kaggle. [4] The dataset includes information on individual games spanning from 1946 to 2023. The data is organized at the team level and includes attributes such as attempted and made goals, assists, steals, and blocks. Each player, team, and coach is assigned a unique ID that is shared across all entities.

1.3 Metadata

1.3.1 Player. We possess data for 3,572 players spanning 56 seasons. Each player is characterized by the following attributes: ID, Year, Name, Team, League, Games Played, Total Minutes Played, Points, Defensive Rebounds, Offensive Rebounds, Total Rebounds, Assists, Steals, Blocks, Turnovers, Personal Fouls, Field Goals Attempted, Field Goals Made, Free Throws Attempted, Free Throws Made, 3-Pointers Attempted, and 3-Pointers Made.

1.3.2 Team. We possess data for 96 teams. Each team is characterized by the same attributes as a player, with the inclusion of three additional attributes: Pace, Matches Lost, and Matches Won.

1.3.3 Coach. We possess data for 254 coaches. Each coach is described by the following attributes: ID, Team, Name, Year, Season Wins, Season Losses, Playoff Wins, and Playoff Losses.

1.3.4 Draft. We possess 8,583 draft records. Each draft entry is identified by the draft year, round, selection, team, player name, ID, college drafted from, and league name.

2 PROJECT IDEA

The main objective of the project is to ascertain the winning team when given two teams as input. A team is constituted by any combination of five players who have participated in either the NBA or ABA. Our aim is to evaluate various player attributes and their past performances to produce accurate predictions. To enhance prediction precision, we intend to utilize machine learning algorithms.

Furthermore, we will explore an alternative approach involving the creation of a novel artificial neural network to forecast game outcomes. This neural network will take combined player and team attributes as input and generate predictions as output.

By analyzing the outcomes from both models, we aim to extract interesting statistics and insights about the data and the training models.

3 SOFTWARE IMPLEMENTATION

3.1 Data Pre-processing

Due to the lack of match-specific details in the NBA Statistics data, we aim to merge the Kaggle dataset, which includes match outcomes, with CMU's dataset, containing player statistics for those matches. The objective is to develop a model capable of establishing a relationship between a set of player statistics and the likelihood of winning by training on historical match data. Initially, the primary

challenge lies in preprocessing the data to ensure its compatibility with the model. This includes aligning the Kaggle dataset with CMU's dataset, as they have varying naming conventions and attributes. Ensuring data authenticity involves verifying values from both datasets through cross-referencing for validation purposes.

3.2 Feature Selection

Since each table encompasses numerous attributes concerning teams, players, and coaches, a vital aspect involves selecting the suitable features for training the model. To achieve this, we will utilize methodologies such as Pearson Correlation and Principal Component Analysis to identify the most pertinent features.

3.3 Modelling

In the end, the selected features will be used to train machine learning models that can accurately predict match outcomes between two teams. Initially, we will use a Naive Bayes classifier as a starting point for prediction. Following that, our ultimate model will comprise two types of classifiers: one relying on decision trees and the other utilizing an artificial neural network.

4 RESEARCH REQUIRED

Several existing studies discuss the application of artificial neural networks and decision trees for prediction purposes. Hassan et al. (2020) [3] explore the use of artificial neural networks to predict matches during the 2018 soccer world cup, while Tang et al. (2018) [5] utilize decision trees for forecasting football matches using data from the CSFAL 2015, 2016, and 2017 seasons. Furthermore, Bruins Sports Analytics [1] discusses methods for forecasting NBA champions using neural networks. All of these sources serve as valuable readings to initiate our project.

5 WORK DIVISION

The project workload will be distributed as follows.

5.0.1 *Tasks for Swaraj.*

- Read at least 2 research paper centered around decision trees
- Cleaning of dataset for example removing NaN values and filtering out irrelevant or garbage data values
- Finding out interesting statistics (mean, median, standard deviation, etc) about the dataset
- Working on decision tree implementation

5.0.2 *Tasks for Utsav.*

- Read at least 2 research paper centered around artificial neural trees
- Joining the NBA and CMU dataset
- Work on feature selection using correlation analysis and dimensionality reduction techniques
- Working on artificial neural network implementation

5.0.3 Common Tasks. Building the baseline Naive Bayes model using pair programming. This will prove as a good catching point before dwelling into advanced model creation.

6 MIDTERM MILESTONE

By the midpoint of our project, our objectives encompass completing the following tasks:

- Finalizing data pre-processing, which includes completing the merging and cleaning of tables.
- Identifying intriguing statistics within the dataset.
- Implementing feature selection algorithms to refine our dataset.
- Constructing a baseline Naive Bayes model to serve as the foundation for our subsequent algorithms.

Additionally, we aim to delve into exploratory data analysis to gain deeper insights into the dataset's characteristics and uncover any potential patterns or trends. Moreover, we will begin laying the groundwork for our decision tree and artificial neural network models, outlining the steps needed for their implementation in the latter stages of the project.

REFERENCES

[1] Bruin Sports Analytics. [n. d.]. Decoding the Game: Forecasting NBA Champions with Neural Network Algorithms — bruinsportsanalytics.com. https://www.bruinsportsanalytics.com/post/nba-champs-neural-network. [Accessed 03-03-2024].
[2] School of Computer Science Carnegie Mellon University. [n. d.]. *NBA statistics data.* http://www.cs.cmu.edu/~awm/10701/project/databasebasketball2.0.zip
[3] Amr Hassan, Abdel-Rahman Akl, Ibrahim Hassan, and Caroline Sunderland. 2020. Predicting wins, losses and attributes' sensitivities in the soccer world cup 2018 using neural network analysis. *Sensors* 20, 11 (2020), 3213.
[4] Kaggle. [n. d.]. *NBA dataset.* https://www.kaggle.com/datasets/nathanlauga/nba-games
[5] Xiaohu Tang, Zhifeng Liu, Taizhao Li, Wenbin Wu, and Zhenhua Wei. 2018. The application of decision tree in the prediction of winning team. In *2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS).* IEEE, 239–242.