NBA Data Analysis and Prediction

Milestone: Data Collection and Processing

Group 11

Yuyang Xiao

617-834-6463 xiao.yuya@northeastern.edu

Hsien Chih Wang

413-801-6469 wang.hsien@northeastern.edu

Submission Date: _____02/28/2022_____

1. Problem Setting

The NBA is the biggest and the most well-known basketball league in the world. Since basketball is a sport of scoring, the statistical data of all players in the league was carefully tracked by the NBA and saved on their website, which provides an opportunity for analysts to study about the players and the sport itself. In the NBA, there are some major awards for players and teams, such as MVP, DPOY, ROY, SMOY and the championship. Thus, in this project, we want to build supervised learning models on the NBA data and finally predict the award winners, future players' stats and more.

2. Problem Definition

a. For the MVP, DPOY and ROY awards, is there a pattern that is based on stats?
b. Who is going to win the MVP, DPOY, ROY and SMOY award this year?
c. How should a team perform to succeed in the playoffs?
d. Can a machine predict a player's stat of next year based on the current stats?

3. Data Sources

https://github.com/gmalim/NBA_analysis

4. Data Description

The data contains the following:

a. Players' basic stats (points, rebounds, assists etc.) from 1999-2000 season to 2019-2020 season.
b. Players' advanced stats from 1999-2000 season to 2019-2020 season.
c. Separated basic stats for all-star players (1999-2000 season to 2019-2020 season).
d. Separated basic stats for rookies (1999-2000 season to 2019-2020 season).
e. MVP, DPOY, ROY voting results from 1999-2000 season to 2019-2020 season.
f. SMOY voting results for 2018-2019 season to 2019-2020 season.
g. NBA team stats from 1999-2000 season to 2019-2020 season

Each dataset contains the data of every player involved in that category, e.g., every player in the league for player basic stats and every player who got at least a score in MVP voting for MVP voting result dataset, to name a few.

5. Data Collection and Preprocessing

For this part, we performed the following steps on the data:

a. Read the data into different dictionaries, separated by the award, e.g., MVP data is stored in MVP_dict, where the keys are the year to the season and the values are the dataframes.
b. Combine the advanced stats of the players onto the data stored in step a.
c. Drop the duplicated columns in the advanced dataset and the award datasets.
d. Drop the irrelevant columns that we think does not represent the performance of a player, such as team names, player names and position.
e. Fill the NA values with 0, since the NA values appears only on the three-point percentage column because the player has not taken a three-point shot in the whole season.
f. Combine the data from each year for each award and put the 'share' column on the last position, because the score share will be our target variable for future analysis.
g. Output the big datasets for each award as csv files.