

NBA game result prediction and analysis

Course title : Big Data Analytics & Data Mining

Name: Haoran HU

Yuxuan CUI

Contact information : [haoran.hu@efrei.net](mailto:haoran.hu@efrei.net)

1. **Introduction/Motivation**

I'm an NBA fanatic, and every goal, steal or reversal of a whistle in the game can make me enthusiastic. Apart from watching the exciting game, we are also curious about the outcome of the game. Therefore, in this project, I will show how to use the past statistics of NBA games, determine the combat effectiveness of each team, and predict the outcome of a game.

We will predict the results of each game from April to June 2018-2019 based on the NBA October-March 2018-2019 game statistics.

1. **Data and Data Analysis**

We will use data from Basketball Reference.com

In this website, you can see the basic statistics of different players, teams, seasons and league games, such as scores, fouls, wins and losses, etc.

Here is the dataset I used:

1. Team Per Game Stats: average performance statistics per team per game.(T.csv)
2. Opponent Per Game Stats: The average per-game statistics of the opponents encountered. The statistics included are the same as those in Team Per Game Stats, but they only represent the statistics of the opponents of the team.(O.csv)
3. Miscellaneous Stats: Comprehensive Statistics(M.csv)
4. Match Results Stats: Game data (schedule and results) for each game of the NBA from 2018 to 2019 (October to March).(2018-2019\_result.csv)
5. Match Schedule Stats: Schedule information for each game of the NBA from 2018 to 2019 (April to June).

We need to preprocess the dataset. So I wrote Sport\_Analystic1.ipynb to delete the columns I don't need and clean the data so that they become the dataset I need for machine learning training

And then we used the **elo scoring algorithm** to analyze the results of the game. This method is used to get the elo score of each team. Therefore, the feature vector we will use to represent the data of a certain game (if A and B are playing): [A team Elo score, A team's T , O and M table data, B team Elo score, B team T, O and M table data]

1. **Architecture of Proposed Solution**

Data analytics architecture

Step1: Find the datasets I need from the website and save them in csv format.

Step2: Import data from csv format into python and store it in dataframe format.

Step3:

Preprocess the data. Look for errors in the data and delete unwanted columns.

Step4: Perform data analysis. We use the elo scoring algorithm to analyze the dataset of match results. This method is used to get the elo score of each team. Then, the feature vector we will use to represent the data of a certain game (if A and B are playing): [Elo score of team A, T, O and M statistics of team A, Elo score of team B, team B T, O and M table statistics]

Step5: We build a regression model using Logistic Regression method using sklearn. Calculate training accuracy using 10% cross validation. Finally, the trained model is used to make predictions from the game data from April to June. Use the model to judge a new match and return the probability of its victory.

1. **Evaluation/Results**

We build a regression model using Logistic Regression method using sklearn. Calculate training accuracy using 10% cross validation. Finally, the trained model is used to make predictions from the game data from April to June. Use the model to judge a new match and return the probability of its victory.

Datasets for training model:

1.T.csv

(Performance statistics per team per game)

电脑屏幕的截图

描述已自动生成

2.O.csv

(Statistics of opponents encountered per game)

电脑屏幕的截图

描述已自动生成

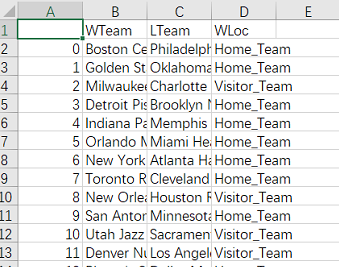
3.M.csv

(Comprehensive statistics)

**电脑萤幕画面

描述已自动生成**

4. 2018-2019\_result(new).csv



Datasets for testing :

2018-2019\_schedule(new).csv



Predicted result :



\*The code to implement these functions is in Sport\_Analystic2.ipynb

1. **Conclusions**

In this project, we use some statistics of Basketball-reference.com to calculate the Elo socre of each NBA team, and use these basic statistics to evaluate the past game situation of each team, and according to the international ranking method Elo Score scores the team's current battle level, and finally combines the characteristics of these different teams to determine which team can take the advantage in a match. However, in our prediction results, unlike in the past, we did not give absolute positive or negative points, but given the probability that the team with the better chance of winning can win the other side. Of course here, the amount of data we use to evaluate the performance of a team is too small (only the data from October 2018 to March 2019). If you want more accurate and systematic judgment, of course you are interested More years and more comprehensive data can be obtained from various statistical data websites. Combine different regression and decision machine learning models to build a more comprehensive model with higher prediction accuracy.

**6.References**