Machine Learning and Data Mining project: Predicting NBA playoff results

Silvio Angelo Baratto Roldan

Course of AA 2021-2022 - Data Science and Scientific Computing

Abstract

Basketball is one of the most popular sports in the US and around the world and many sponsors invest large amounts of money for it in teams and players. The NBA playoffs are part of the NBA-related league. This competition consists of four rounds between eight Eastern Conference and Western Conference teams. The playoffs are made up in the eights, quarter-finals, semifinals and final of the Conference. Finally, the winners of each final of their conference clash with the other winner in the NBA Finals. Each round consists of 7 games, so it takes 4 games to win to move on to the next round.

1 Problem statement

The goal of this project is to create a model [1] that simulates the complete playoff scoreboard and predicts the winner of each game and round. The model is composed by three parts: training the model, generating feature values and simulation.

2 Assessment and performance indexes

To select the best classifier to train the model and to use it in the simulations three different metrics have been taken in consideration:

- `Accuracy`: intuitive performance metric and very useful one when the dataset is balanced or slightly balanced.
- `AUC`: area under ROC that measure the classifier performance.
- `MSE`: measures of the averages of the squares of the errors that is, the average squared difference between the estimated values and the actual value.

3 Proposed solution

For the purpose of simulation high accuracy and low error were the goals when selecting the right model. At this purpose three different classifier have been taken in consideration. In the table below [1] is show the Pros e Cons between them

Model	interpretable	Accuracy	Training speed	Prediction speed
SVM	No	Higher	Fast	Fast
Random Forest	No	Higher	Slow	Slow
Naive Bayes	Moderate	Lower	Fast	Fast

Table 1: Pros e Cons between SVM, Random Forest and Naive Bayes

4 Experimental evaluation

4.1 Data

The complete dataset was composed by 5 .csv files:

- `games.csv` : all games from 2004 season to last update with the date, teams and some details like number of points, etc.
- `games_details.csv` : details of games dataset, all statistics of players for a given game.
- `players.csv` : players details (name).
- `ranking.csv` : ranking of NBA given a day.
- `teams.csv` : all teams of NBA.

To train the model, only the data present in the games and teams dataset were used. The variables used within the `games.csv` dataset are: field goal percentage, field goal three point percentage, free throws made, rebounds, assists, steals and blocks from January 2004 untill march 2021 both for the home and visitor teams.

4.2 Procedure

The model consists of three parts:

- **ML model selection**: Initially, the data is analyzed and trained with a regression model to predict the probability of winning.

- **Generate feature values**: To simulate the playoffs, feature data were artificially generated with the use of different statistical distributions. For this purpose, the most recent data in `games.csv` were used to generate new data, as they are more representative.
- **Simulate playoff**: Using the generated data, multiple game simulations are done. The winner of each game is predicted using the trained model. The final simulation result is the average of the simulation results.

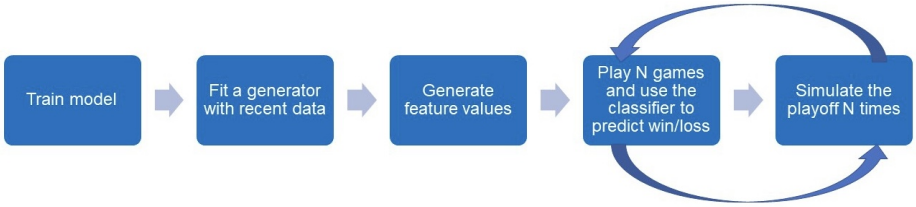

Figure 1: Flow diagram

4.3 Results and discussion

In the project, the selection of the right classification model is central to the correct forecast of the scoreboard. In the section [3] the benefits of the three models chosen for the simulations have been listed. The table [2] shows the results:

	SVM	Random Forest	Naive Bayes
Accuracy	0.71	0.84	0.83
AUC	0.71	0.83	0.82
MSE	0.28	0.16	0.16
Time	1min, 15sec	31min, 27sec	51 sec

Table 2: Metrics results

As you can see from table [2] Random Forest and Naive Bayes result the best models for both accuracy and average error. From table [1] we expected Naive Bayes perform worse than SVM but both models are sensitive to parameter optimization, without tuning the hyperparametres SVM performs better than Naive Bayes. Random Forest result the most accurate classifier. Despite this, in the game simulation the execution times for 5000 iterations were extremely higher for Random Forest compared to Naive Bayes and SVM. However, for the correct forecast of the scoreboard, the waiting time necessary for the simulation is not considered an essential metric, as the model predicts the result before the event of the playoffs and therefore does not require a forecast rate in real time.

In conclusion, Random Forest is the model that is considered most appropriate to use to predict the final scoreboard.

4.3.1 Model output

The figure below shows the output achieved after 5000 simulations of the playoffs of the current season 2021-2022, using Random Forest as a classifier. The program assigns a probability to all teams for each round. The team that receives the highest probability is the one that according to the model will advance to the next round. In the current season at the time of writing the report, the semifinals still have to be played (Heat - Celtics: 3-3, Warriors passed to the final), as can be seen with the exception of the Suns-Mavericks and Bucks-Celtics matches, the rest was correctly predicted. The reason for the mistake can be admitted to the fact that Suns and Bucks played below expectations.

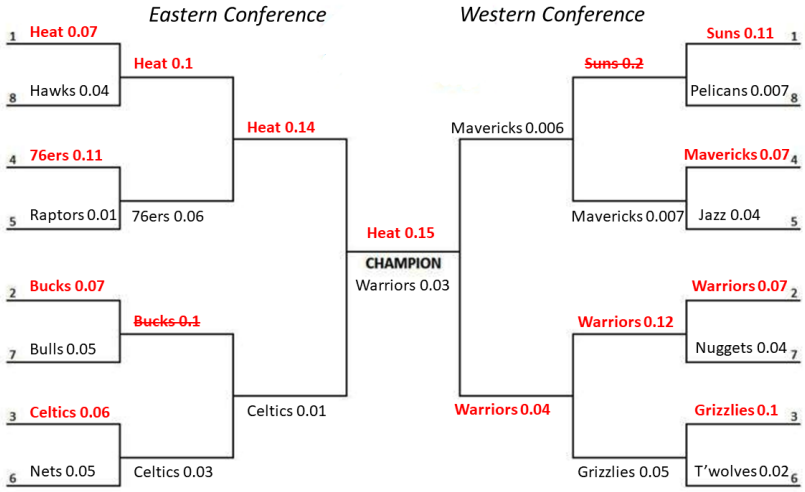

Figure 2: playoff season 2021-2022

5 Conclusion

The result in the figure 2 predicts the result very well based on the statistics of the matches. These predictions could be further improved by adding player statistics, injuries, or playing strategies used in previous matches.