FAIKR-Mod3 NBA Project

Francesco Cavaleri, Giacomo Piergentili
Master's Degree in Artificial Intelligence, University of Bologna
{ francesco.cavaleri2, giacomo.piergentili2 }@studio.unibo.it

November 24, 2023

Abstract

In this mini-project, we utilized Bayesian networks to model the collective distribution of various performance indicators in basketball players. The primary goal was to uncover the probabilistic relationships among these indicators and identify the key factors influencing a player's winning percentage.

Upon analyzing the Bayesian network, several intriguing dependence relationships came to light. These findings not only validated existing knowledge, primarily derived from experience, but also provided a scientific backing to some well-known results. Ultimately, we simulated various scenarios to gain insights into the principal determinants contributing to an increase in the number of games won by a player.

Introduction

Domain

Sports analytics has experienced rapid growth both for teams and individual players in recent years. Statistics tools can be particularly useful to assess performances and define efficient gaming strategies. In this report, we focus on basketball, a sports pioneer in using analytic tools, and data from the National Basketball Association (NBA). In our work took inspiration from (D'Urso, De Giovanni, and Vitale 2023), trying to recreate a similar bayesian network with the various datasets we were able to retrieve from multiple sports statistics site and establish if the correlations described in the paper, between vaious statistics, can be considered a good way to explain the value of a player.

Aim

The purpose of this project is to implement part of the Bayesian network described in (D'Urso, De Giovanni, and Vitale 2023), compare it with the networks generated from our data and evaluate their performances to choose the best one in our case study.

Method

We used the pgmpy library for Bayesian Network modeling and then defined several Bayesian Network structures: one is based on statistical dependencies between variables, the remaining ones are obtained from the data. The choice of using more than one network was dictated by the problem that we had to face: the amount of available data wasn't enormous. In fact the resulting networks were relatively big for how much data we have. For that reason we tried to create different models and check which one performs better. The networks that developed from the data are obtained using a search methods called HillClimbSearch with different scoring methods (BicScore, BDeuScore, K2Score). The bayesian networks were then fitted with the training dataset (2021/2022 NBA Season). The trained models were tested with the test dataset (2022/2023 NBA Season) and the performance of the models were evaluated using the Mean Absolute Error (MAE) and Confusion Matrix. After comparing the results of the models we have found that, the network which provide us the best results, are the one made by experts. In the end we formulated some queries to explore the relationship between variables and visualized them.

Results

From our tests we saw that the best model for our purpose is the one created by the expert, because it captures relationships between nodes that, our models, created from the datasets, can't. Also, through some queries, we were able to analyze in a more detailed way the behaviour of certain statistics.

Model

In (D'Urso, De Giovanni, and Vitale 2023) we can find a network made by experts in the field: this network is able to capture some details that are difficult to obtain solely from the data. What we ended up using is a lighter version of this network. The other networks were created starting from the data and using built in algorithms in pgmpy. These algorithms manage the problem of finding the optimal DAG by using a search algorithm with heuristics (Hill Climbing). We generated three networks, each with a different estimator (BicScore, BDeuScore, K2Score) and the resulting networks show that: the BicScore estimator has difficulties capturing the dependencies between the nodes, BDeuScore estimator found a reasonable amount of dependencies and K2Score estimator found too many dependencies. For the last reason we decided to approximate the network made using the K2Score estimator with the Chow-Liu algorithm (it construct a tree that approximates the optimal structure), this allow us to fed the new simpler network with the amout of data that we have.

At the end of our results the network made by experts performs better than the other tested networks.

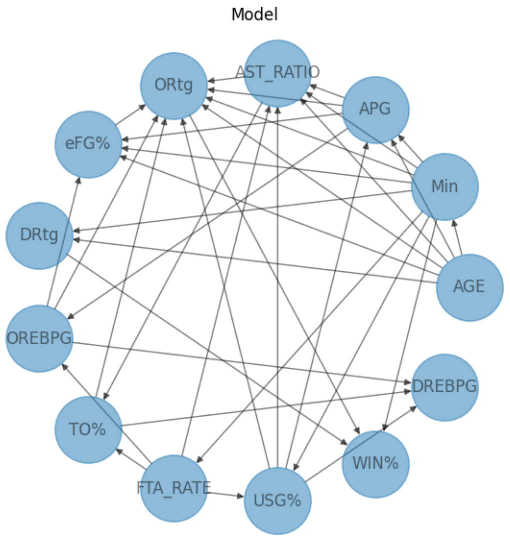

Figure 1: Bayesian Network

Analysis

Experimental setup

During the development of the project there were some questions that came up and that we wanted to answer through some queries: firstly, we wanted to understand if there is a specific age range where a player reaches the peak performance, and, to answer this question, we queried the offensive and defensive rating given the age, because those two statistics are the ones that describe more comprehensively the game of a player. After that we wanted to see if a high offensive and defensive rating translates to more wins, and, to check that, we queried the WIN% (winning%) given those ratings. Another interesting subject is the age, and, mainly, if the usage of a player drops after reaching a certain age, so we queried the USG% (usage%) given the age. The last question that came up was: given a player that makes a lot of assits, does he tend to loose the ball more? To answer we queried the TO% given the assist ratio.

Results

From the queries described in section before, we have found that capturing the performance of a player starting from this simple network is difficult: we have seen how WIN% changes given ORtg and DRtg (Offensive rating and Defensive rating), but we have also seen how ORtg and DRtg do not change with the age of the player. One possible explanation is that, since our datasets are formed only by players that played at least 61 games, all those players are so good that the variable age does not influence the other variables. This is particularly evident when we consider USG% given AGE: the two variables seem independent even if they are not (older players tend to play less as they get tired before and are more injury prone).

In the last query we try to understand the behaviour of TO% given AST_RATIO (Assist Ratio). From our results we have seen that if AST_RATIO $> 35\%$, the probability of TO% increases. This is an expected result as passing the ball to another player is one of the easiest way of getting the ball stolen. In fact in our datasets only few player have a AST_RATIO $> 35\%$.

Conclusion

After all the work done, we can confirm that the support of sports experts is essential for the application of machine learning techniques in this field. The problem of having a tiny dataset is not simple to solve since a bigger dataset would for sure result in a more detailed case study, but it also would not capture the essence of the modern game since it tends to evolve really fast. Also, from our queries, even spotting dependencies between variables that seem correlated is hard. Anyway, this makes sense, since it is a complex game and there are many other variables that should be taken into account to have a more comprehensive view on the subject.

Links to external resources

- This is the Github repository where all the work can be found: `https://github.com/CacioCavalloIsNotReal/FAIKR_project_NBA`
- These are all the datasets that were used:
  - `https://www.nba.com/stats/players/traditional?Period=4&Season=2021-22`
  - `https://www.nbastuffer.com/2021-2022-nba-player-stats/`
  - `https://www.kaggle.com/datasets/amirhosseinmirzaie/nba-players-stats2023-season/data`
  - `https://www.nbastuffer.com/2022-2023-nba-player-stats/`

References

D'Urso, P.; De Giovanni, L.; and Vitale, V. 2023. A bayesian network to analyse basketball players' performances: a multivariate copula-based approach. *Annals of Operations Research* 325(1):419–440.