Github repository: https://github.com/Edvard132/TimoVillemEdvard
Team members: Timo Kaasik, Edvard Notberg, Villem Paabo

D20 An analysis of previous NBA leagues

Business understanding

Business goals

- Background

Our group doesn’t have any prior experience with this kind of databases outside of this course.

- Business goal

Our primary goal is to develop a model which could predict the odds of an NBA team winning a match, and to find out whether a team has a considerable edge when playing on its home field. Also, we plan to provide predict different game aspects, such as expected most-valuable-players, rebounds, assists, blocks, three-pointers, free-throws and steals.

- Business success criteria

All the predicted results will be compared to actual scores afterwards. The project will be judged a success if the outcome of an event can be predicted correctly at least 50% of the time.

Current situation

- Inventory of resources

Our resources are sufficient, the data of NBA matches are publicly downloadable. As a software we will be using Excel tables to store the data and Jupyter notebook to process it.

- Requirements, assumptions and constraints

The project is scheduled to be completed before the poster session. Our poster must be ready and will be presented on Thursday, December 15, 2022 at 14:00-17:00.

- Risks and contingencies

If including statistics from several years to an algorithm that predicts a winner just of a particular match-up, predictions will get less relevant as different factors such as team members, coaches etc. probably have changed over seasons.

- Terminology

Data-mining (DM) - discovering patterns from data
Machine-learning (ML) - algorithms that learn from data
Data-science (DS) - science about how to operate with data

- Costs and benefits

As we are doing this project for school, we do not have any budget and financial costs. Our only cost will be our time. Accordingly, we will not receive any money for our work, and our only benefit will be the experience and the grade at the end of the course.

Data-mining goals

- Goals for the project

1. Try building a model for predicting NBA game outcome details.
2. Visualize a chance of a team winning and game in-depth statistics indicators (such as three-pointers) at the home-court
3. Finally, create a report concluding all our work.

- Data-mining success criteria

Data should be large enough to predict an accurate outcome of the match-up for each team.
Dataset should have no or as few as possible missing values.

Data understanding

- Gathering data

The data can be easily found and downloaded as a preformatted csv file from many websites, https://www.basketball-reference.com/ being one of them. They provide different measurements from which it is up to us to decide which of them to use. The problem is doing it easily and fast for it all to be in one place and easy to access.
- Describing data

- Exploring data

As of yet we have not made many thoughtful inquiries about the data.

- Verifying data quality

The data is good, as it matches from page to page.

Project Plan

1. Collecting the data

- Estimated time: 5 hours
- Methods, tools: web-search
- Details: objective is to gather only the data that we need, as there are lots of different NBA datasets available on the internet.

2. Data preparation

- Estimated time: 5 hours
- Methods, tools: Pandas dataframe on Jupyter notebook
- Details: preparing data before operating with it. It is possible that we will drop values that we won't need .

3. Discovering patterns from data

- Estimated time: 15 hours
- Methods, tools: pandas, seaborn
- Details: discover the most frequent and relevant patterns, based on statistics and group them

4. Visualization of data

- Estimated: hours 8 hours
- Methods, tools: matplotlib, pandas, seaborn
- Details: Visualize relations of the attributes

5. Clustering data

- Estimated: hours 8 hours
- Methods, tools:
- Details: Applying clustering algorithms

6. Apply machine learning

- Estimated: hours 20 hours
- Methods, tools: Jupyter notebook sklearn.
- Details: Trying to build predictive models for classification, regression.

7. Poster

- Estimated: hours 8 hours
- Methods, tools: Microsoft Word
- Details: Create a poster to present at the poster session