Machine Learning - Assignment 2

Publication date: 16/06/2024

Due date: 06/07/2024

Data

Data source: International football results from 1872 to 2024

This dataset contains a comprehensive record of international football results spanning from 1872 to 2017. It includes match outcomes between national teams, providing a historical view of international football performance.

Your goal is to predict whether the home team won the match. And to provide additional insights into the data.

The dataset encompasses over 40,000 matches with various attributes detailing each game's context and results. The data contains multiple files that provide information regarding the game result, shootouts and goal scores:

- **results.csv**: Contains the primary data of the dataset, including details of each match such as the date, home team, away team, home score, away score, tournament, city, country, and whether the match was played on neutral ground.
- **shootouts.csv**: Provides information about matches that were decided by penalty shootouts, detailing the teams involved and the outcome of the shootout.
- **goalscores.csv**: Contains the scores of the goals in international games along with minutes of said goals.

A detailed explanation of each column can be found in the Kaggle page of the dataset.

Requirements

Section A - Data Exploration & Visualization (10 pts)

Explore the data using tables, visualizations, and other relevant methods.

- Plots should have an informative main title, axis labels and a legend (if needed).
- For each plot or table, provide a short description of **key observations**. Make sure to only include content which would be **meaningful/informative** for a team manager.
- The visualizations should be detailed and cover all relevant aspects of the data.
- The visualizations should highlight any interesting patterns or trends that can be observed in the data.

The goal of this section is to get insights on the data which may or may not be relevant for the following sections.

Section B - Data Pre-processing (30 pts)

Apply different methods of pre-processing to the data in order to prepare it for the models you wish to apply in the next sections.

- Perform feature engineering on the data, including the specific features provided below and at least six additional features of your own choosing.

  Explain why you chose these features and how they may improve model performance.

  Create the following features - add them to the current results.csv data:

  - Home team won: This is a boolean feature representing whether the home team won the match. In the case of a draw, the winner is determined by the **shootouts.csv** file.
  - Home team win rate: This feature calculates the percentage of matches the home team has won **up to that point in time**.
  - Away team win rate: This feature calculates the percentage of matches the away team has won **up to that point in time**.

    - Home team average goals: This feature calculates the average number of goals the home team has scored **up to that point in time**.
    - Away team average goals: This feature calculates the average number of goals the away team has scored **up to that point in time**.
    - At least six other features of your own choosing.
- Apply at least one type of imputation (if needed), one transformation, and one exclusion (i.e., feature selection).
  Provide an explanation to each method you apply. Your choice should reflect an understanding of the method and why it's needed.

Section C - Home team winning (25 pts)

Use at least **three** different machine learning models to predict whether the home team won the match (Home_team_won feature).

- When training a model to predict home team winning, you can use all the available data, **except** for features that reveal the winner like scored goals and any features that don't provide relevant information for the prediction task, and. (However, you can still use these irrelevant features to engineer new features that can be used in the model).
- The implementation must include parameter tuning.
- Report a suitable measure to evaluate the performance of each model.
- Present the models' results in a plot.
- Compare the results of the different models, discuss them.

Section D - Clustering (25 pts)

Apply at least **two** clustering algorithms on the prior data to cluster the different **teams**.
Make sure that before clustering the teams, you create a new set where each row represents a team and the features describe their game history (create additional features if needed). Use the win rate and average goals similarly to the features you created in section B, with the features you engineered.

- Use parameter tuning based on the algorithms you selected.
- Identify the most important features that contribute to the differences between the clusters. Discuss your findings and find a way to demonstrate **visually** what similarities the clusters may have.
- Use a method (of your own choice) to estimate the quality of the clusters you created with each clustering algorithm. Visualize the results according to the method you selected.

Presentation (10 pts)

Create a short presentation (no more than 6 slides) that includes interesting findings of your choice. 3–4 presentations will be chosen to be presented in front of the class. The goal is to learn from other students' work.

Section E - Clustering and Dimensions Reduction - Bonus (10 pts)

- Reduce dimensions of the **teams** data (from section D) using the PCA algorithm.
- Show which principal components explain the majority of the variance in data, using a plot. Identify the features that are most strongly represented in each component.
- Use the top principal components to perform clustering on the customers, using the same clustering algorithms as before.
- Visualize the clusters before and after PCA. Compare the results to the clustering performed without PCA. Are there any differences in the clusters obtained? If so, try to explain why these differences exist, and discuss how does using PCA affects the results of clustering.

Section F - Exploring Players - Bonus (10 pts)

In this section, you will explore the different players and their performance.

- Using the **goalscorer.csv** file, create a new dataset where each row represents a player and the features represent the player's goal statistics (e.g., total goals, match count, average goals…)
- Formulate a question that can be asked about this set, and suggest a machine learning algorithm that can answer it. If you decide to implement a method that was not discussed in the course so far, include references to where you studied it.
- Apply the suggested machine learning algorithm.
- Discuss the results and reflect on your question and choice of solution.

Guidelines

Please read the following section carefully before submitting the assignment.

Coding Guidelines

- Use familiar packages with explicit explanations.
- If you have installed any libraries beyond those presented in the exercises, please specify this in the report.
- The code should run without warnings or errors.
- Good documentation is **critical**.
- Indicate the exercise sections in the code as well.
- Use meaningful variable names.
- Do not use reserved words.
- Use constants where possible.

Submission Guidelines

- The assignment should be submitted in pairs (only one submission).
- You are required to submit two files including all the sections. One in **.ipynb** format and one in **.html** format. **Both files should also include the program's outputs**. In addition, you are required to upload a **PDF** file of the presentation you prepared.

- The files' names should be of the form: **ML_HW2_ID1_ID2**.
- Assignments submitted late will receive a penalty of **3 points** for each day, up to one week. Later submission will not be accepted.

Grading

You can get more than 100 points for the exercise. The exercise will be graded according to correctness, clarity, efficiency of implementation, elegance of implementation.

Self-learning

As we mentioned at the beginning of the course, self-learning is an important part of the course. Treat all sources of information carefully and critically.
Usage of LLMs is allowed, but reference it when it was used.
You can and should consult with other students in the course, but each pair must write their own work.
It is reasonable to assume that not all results and algorithms will be identical.

Questions and Reception hours

- Please post your questions on the exercise forum in Moodle, after you have read the previous posts. Professional questions sent by email will not be answered.
- If you want to schedule a reception hour with one of the instructors - please send your questions by email in advance.
- In any other case (personal questions, request for an extension with a justified reason, etc.) please email the instructor.

Good luck