Soccer score prediction

The goal of this project is to apply all material learned in INF161 and successfully complete a data science project. Please read the description carefully!

This project is a compulsory part of the course. This project contributes 45% to the final grade. The grade will be based on good choice of methods, correctness of answers, clarity of code and thoroughness and clarity of reporting.

Requirements

You will build a machine learning model to predict soccer scores. The system takes as its input the teams to play each other and will output the predicted score. The data used to fit the model comes from the previous years of play in the same soccer league. Note that data from newly promoted teams may be missing. The work will consist of four parts:

- Data preparation (40 pts):
  - Input: raw data
  - Output: clean data
  - Features: This systems takes the provided data and generates a dataframe that can be used in the machine learning model. Data description, visualisation, and feature engineering are important parts you should report on to explain the choices made within this system.
- Modelling and prediction (40 pts):
  - Input: prepared data
  - Output: machine learning model, expected generalisation RMSE
  - Features: This system takes the prepared dataframe and builds a machine learning model for predicting scores. Model selection, feature selection and handling missing data are important parts of this system. You should evaluate at least 3 fundamentally different modelling approaches before selecting the final model. We evaluate the performance of the system by comparing the predicted scores with the known scores on a validation/test data set. Specifically, the system should be evaluated with the root mean squared error (RMSE) of predictions, i.e.

    $$\sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}},$$

    where N is the number of predictions, $\hat{y}_i$ is the i-th prediction and y_i is the corresponding true score. The system should report the expected generalisation RMSE.
- Prediction (10 pts):
  - Input: machine learning model and 2020 data
  - Output: prediction of scores
  - Features: Given new data, this system should return predictions of the scores. You may (optionally) test your prediction RMSE on kaggle (`https://www.kaggle.com/c/inf161-innforing-i-dat`
- Website (10 pts):

  - Features: The website should allow users to enter two teams and return a predicted score. Note that this is a HTML document that exists on your personal computer that you open with your browser and not a website hosted on the internet.

Deadlines

The project consists of three parts with three distinct deadlines. In the first part you will prepare the data for analysis. In the second part of the project you will design a machine learning model and predict scores. The last part of the project consists of creating a simple website that runs your score prediction system.

- Deadlines:
  - Part 1: Sunday, 19.09, 23.59
  - Part 1 - peer review: Sunday, 26.09, 23.59
  - Parts 1&2: Sunday, 10.10, 23.59
  - Part 1&2 - peer review: Sunday, 17.10, 23.59
  - Complete project: Sunday, 31.10, 23.59
- Deliver at MittUIB.no/assignments

Deliverables

All dealdines are mandatory. The first two parts will not be graded, but you will give and receive peer feedback that will improve your final project.

For the final submission, please provide the following:

- A jupyter notebook `preparation.ipynb` for data preperation The notebook acts as both report and submitted code. It should contain all the code to reproduce your work and a report of all your methodological choices and results. Please "restart and run all" before submission, so that you submit a clean version.

- A jupyter notebook `model.ipynb` for modelling, generating a machine learning model that is saved to disk. The notebook acts as both report and submitted code. It should contain all the code to reproduce your work and a report of all your methodological choices and results. Please "restart and run all" before submission, so that you submit a clean version.

- A file `predictions.csv` with predictions for each user and each game ID in 2020.

- A zip file that contains the file `app.py` for running a local website and all neccesary files such that `app.py` runs with the command `python3 app.py`. The website should then run at `localhost:8080/`.

Note that each notebook and the app need to run independently.

In addition to packages from the standard library, you may use the following python packages: `xlrd`, `numpy`, `pandas`, `scipy`, `sklearn`, `matplotlib`, `seaborn`, `requests`, `plotly`, `flask`, `django`, `waitress`. If you use any other packages we will not be able to run your app and you will fail the project.

Code should be documented and tricks (e.g. to avoid division by zero, to make sure it takes finite time to run, etc.) should be reported. The rational behind all steps in the code should be clear from the report.

NOTE: This project is a learning experience. If we see that you have copied your answers from online resources, you will get 0 points.

Model selection is an important part of the task and will be graded accordingly. Before applying machine learning algorithms, you should always consider (and report) what results you expect. When you have successfully applied machine learning algorithms, you should always comment on how well the results match your expectations.

Leaderboard

There will be a leaderboard at kaggle (`https://www.kaggle.com/c/inf161-innforing-i-data-science-2021/leaderboard`). Note that your grade will not depend on your leaderboard position.