Assignment instructions

Assignment description

Goal: build an end-to-end ML pipeline, from data extraction to deployment of the model endpoint. It is a group assignment, you will work with the default ESADE groups.

Tasks:

1. Select one or more datasets to use. You can find datasets here:
   - Kaggle datasets: https://www.kaggle.com/datasets
   - UC Irvine ML Repository: https://archive.ics.uci.edu/
2. Define a business question, e.g.
   - Which are the patients that get more ill?
   - Which sales will we have next quarter?
3. Build the ML pipeline. It has to:
   1. Extract the data from the source(s).
   2. Pre-process the data as needed
   3. Train a ML model that can answer the business question
   4. Fine-tune the model
   5. Deploy a prediction endpoint to receive new queries
4. Write a report describing the solution

Tools:

- You can use any tools that you know of, including those from this course or previous ones (SQL, SageMaker, containers, EC2, container registry, Jupyter notebooks, etc).
- Restriction:
  - You HAVE TO use cloud tools for the pipeline.
  - This means all or part of the pipeline should be running in AWS.
  - You can run ***part*** of the pipeline locally in your computer, but not ***all*** of it.

Important dates

- March 21st: submit the business question and dataset(s) to use.
  - I will provide feedback
  - Only one answer per group
- April 9th: submit the assignment files & peer-evaluation form

NO EXTENSIONS WILL BE GIVEN

Files to submit:

Per group:

- Technical report
- Code: you MUST submit all necessary files and code so that I can recreate your pipeline.
- The datasets are not necessary since you're using data available from the Internet.

Per student: fill in a peer-evaluation form

Grading

Assignment grade:

- 50% Report content, correctness, all sections present, precise language, brief explanations
- 25% Code can be executed
- 25% Average of peer grading

Contents of the report

- Length: Maximum 5 pages. Additional pages WILL NOT be considered.
- You don’t need to write a lot of text.
- Simple and short is preferred to complex and long.

The report MUST have the following sections.

- **Business proposal:** detail the business question you want to answer. Provide the necessary context to understand the proposition.
- **Dataset description:** briefly explain which dataset are you using, the source, and mention the most relevant variables.
- **Technical solution:** provide a diagram of the different elements you’re using, explain any relevant decisions or choices
- **Evaluation of results:** your answer to the business question
- **Reproduction instructions:** a brief summary of how to run your pipeline