Data Science Assignment

Please do not share your solution or the problem statement publicly.

This test is an evaluation of your overall reasoning and python coding skills. The solution could be rather complicated, so we suggest you focus on answering as many questions as possible to a reasonable degree in the time you have, rather than focusing too much on a single specific question. For instance, getting the perfect classifier should not be your priority, you should rather aim for a decent one you will be able to use.

Your submission should include

- A documented jupyter notebook, motivating every step you decide to take.
    - Please include any additional code you write and import in the notebook.
- [OPTIONAL] An additional docker app for inference, working locally.

Problem walkthrough

In this problem, you will build a classifier with **`loan_status`** as a binary target. In order to achieve that, you should only include loans that are either `'Fully Paid'` (your 0's) or `'Charged Off'` (your 1's).

- Download the dataset from https://www.kaggle.com/datasets/ethon0426/lending-club-20072020q1
  Note: Rename to **.csv** the **.gzip** file you will get after extracting once, and use it as a regular csv file.

1. Run a quick analysis of the data and drop the **`grade, sub_grade`**, interest related columns and any other columns you think may cause information leakage.
2. Propose at least two possible modeling approaches –including examples– based on your analysis, expand on the various pros and cons.
3. Choose one of the above approaches and implement your classifier, evaluate your model using one or more metrics you think are suited for the problem.
4. Explain (as in "explainability") your model in terms of the classifier's features.
5. Discuss additional steps you would take if you had more time or resources, and any other datasets you could use to enrich your original dataset.
6. [OPTIONAL] Discuss the scalability of your solution, in terms of the
    a. Number of loans/rows in the training data.
    b. Number of predictions a possible inference endpoint for the model will be asked to make in production.
7. [OPTIONAL] Find a way to relate your predictions to the local/global shape of the dataset, that is, data point clusters, graph node communities, etc..