Final project of Computing for Data Science

Overview and Objective

The objective of this project is to build the architecture of a library that is scalable and functional for a Data Science project. Divide yourself in groups of 2-4 students that will work as a team to design a library aiming to solve a predictive analytics task that will last for several months. There are two parts of the project: First, you will create the skeleton of the library by defining how the files and folders will be organized in a structured way. Then, you will create the first end-to-end prototype that loads the data, preprocess it, creates a bunch of features, trains a model and evaluates the predictions with some metrics. You should design the library thinking that the project will evolve, setting clear guidelines on how to scale the number of preprocessors, features, models and metrics and how will they look like.

Instructions

The project entails the following tasks:

1. Choose a dataset that will be used in the predictive analytics project. There are online resources where you can find datasets (e.g. www.kaggle.com, https://datasetsearch.research.google.com/). [Optional] You can load it to MySQL or another database management system and connect to it through Python.
2. Create a Github repository that will contain the library you are going to build for this project.
3. Create a library meant to analyse the dataset.
4. Set the folder structure of the library.
5. Create the end-to-end pipeline that will build the first model and evaluate it on a set of metrics. You can run this pipeline using Jupyter Notebooks by importing the functions and classes from the library.
    a. The pipeline should contain a preprocessing step with one or multiple preprocessors.
    b. The pipeline should contain a part to build some features to be used by the model. At least 3 features (or feature sets) should be build independently.
    c. The pipeline should contain a step to split train/test or to do cross-validation.
    d. The pipeline should train the model and generate predictions.
    e. The pipeline should contain a step that evaluates the model's performance on a set of metrics.
6. Create unit tests for the preprocessing and feature computation part.
7. Set the guidelines to scale the library to be able to add new preprocessors, features, models and metrics.
8. Prepare a 10-15 minutes presentation to show your end-to-end pipeline, how you organised yourselves as a team and the architecture of the library, explaining how a new member of the team should contribute to it.

Follow the programming principles and best practices that we saw during our classes (DRY, loops, functions, classes and object oriented programming, unit testing, loose coupling, etc.).

Evaluation

You will present the project the 21st of December 2022 to the professor and your classmates and will share the repository by turning in the assignment through Google classroom. If you create a private repository remember to invite my Github user: Icedgarr.