TDS Project: part 1- Dataset Selection & EDA, Basic Model Analysis

---

In the 1st part of the project, you'll find a dataset of your choice followed by analyzing it's nature, characteristics and properties by creating meaningful visualizations. Then, you'll define a **regression** problem relevant to said dataset, and build a simple ML pipeline to serve as a baseline. We recommend using XGBoost regressor or an equivalent state-of-the-art off-the-shelf model, with a **default configuration**.

Lastly, you'll conduct an error analysis on the model's performance.

You can use the example you've seen in class as a reference.

---

1. Dataset Selection

Choose a real-life publicly-available tabular dataset. Possible sources are the Kaggle website or Google Datasets Search. Do not use "easy" text-book ones such as the Flowers dataset.

- The dataset should contain at least 1K rows.
- The dataset should contain at least 10 different attributes.
- The dataset should mostly contain numeric and categorical attributes, but can also contain additional, more complex types of data such as text, time series or images. These additional data types can be used to improve the DS pipeline in the next stage of the project.
- Different teams will **not** be allowed to use the same dataset.
- You are required to provide a reference to the dataset you've selected.

---

2. Data Analysis

Include a section of 5 visualizations that are relevant for understanding the nature of the data, mainly focused on the target attribute and it's relationship with other attributes. The visualization's quality is highly dependant on it's readability- meaningful x and y axis labels, ticks, and titles. Furthermore, each visualization should come with a markup that explains why you chose it and why it's significant for the task at hand. **Important! Choose only the most insightful ones, and don't repeat yourself!**

---

3. Basic Model Pipeline

This section contains the execution and results of the basic pipeline (We recommend using XGBoost regressor or an equivalent model, but will also accept simpler models). **Important-** as in the next stage of the project you will improve your DS pipeline based on an analysis of the model's error – **the predictive performance should not be significantly high.** If this is the case, consider altering the dataset or target column, as otherwise it will be more difficult to improve an already good model.

---

4. Error Analysis

Conduct an error analysis that examines the model performance and find its weaknesses. Present and explain what you've found. Some pointers you can use for example:

- On which items the model performs the worst? The best? Why do you think that is?
- Is the model mostly overestimating? underestimating? Why?
- Are some features sabotaging the model? How?
- Are you able to find commonalities between the erroneous samples?

You should include code snippets, markups and visualizations that support your statements and they should be readable.

Submission Guidelines

You will work in groups of 2 people. Each group should create a public github repo, and it should consist of (by the submission date) a single notebook (ipynb file) which contains all relevant code, text, visualizations and references, as well as the dataset (in a “data” folder).

The notebook should be reproducible, I.e. I should be able to run it on my machine- If you use non trivial packages (not introduced in class), add a requirements.txt file

refer to the google sheets file in the course site- fill it ASAP, the datasets are accepted in a first come-first served manner.

By 10.12.2024, you should upload a Jupyter notebook containing:

You must add a reference to every source you used (blogposts, notebooks, as well as usage of LLMs and any type of assistance).

Good Luck!

For special requests/approvals, email me- itay.elyashiv@gmail.com