Assignment 1

CSCI 5901 Applied ML - Summer 2022

Start Date: May 31, 2022,
Due Date: June 8, 2022 11:30PM Halifax Time

Background Information

We have discussed the model development is not the only part of providing an ML service in production. In this assignment, we assess your knowledge on developing an ML model and design a simple pipeline from data ingestion to model deployment. The assignment has two parts. 1) model development 2) Pipeline development.

Your Task

Step 0: Explain the problem you are going to solve using the selected data. [5 points]

Step 1.1: Find a public dataset in a domain you like. This dataset must be have the following features. 1- there is a possibility of finding versions of the data. 2- possibility of change in the data. 3- possibility of receiving future updated on the data. 4- have at least two protected features.

Step 1.2: Define some ML metrics to evaluate your model. [5 points]

Step 1.3: Define some business metrics to evaluate your model. [5 points]

Step 1.4: Define some software metrics to evaluate your model. [5 points]

Step 2.1: Describe the dataset objective and the features made you decide on selecting it. [5 points]

Step 2.1: Describe the quality of dataset using a radar chart with enough explanation. [5 points]

Step 2.2: upload your data in a public repo on `https://git-lfs.github.com/`. [5 points]

Step 3: What are the features in your dataset? what is the target variable?. [10 points]

Step 4: Among the features what are the features with more predictive value? [5 points]

Step 5: Identify all protected features? (for example in some domains we can say Gender is a predictive feature.) [5 points]

Step 6: Build a model to predict your target value. [5 points]

Step 7: Explain what model you utilized and the reason of choosing it. [5 points]

Step 8: using the identified ML metrics evaluate your model. [5 points]

Step 9: Perform error analysis on your dataset and try to improve the performance of your solution by investigating the samples in your dataset. [5 points]

Step 10: evaluate the fairness of your model. Use subsets of data to assess the fairness in regards to the protected features [5 points]

Step 11: build a pipeline to ingest your data, train the model and deploy it in a specific folder. [5 points]

Step 12: suggest a way to calculate the software and business metrics for your pipeline. [5 points]

How to Submit

After you finished your coding and make sure it is working well, you run your code step by step in a Python notebook and print the notebook in a PDF file. We call code_A1.PDF

Now you submit the (1) code_A1.PDF, (2) your notebook and (3) link to the data you used to BrightSpace.