ML Engineer Home Assignment

Introduction

The purpose of this assignment is to evaluate your proficiency in constructing a machine learning pipeline. The pipeline will encompass several aspects including data processing, feature calculation, model prediction and logging. You will be supplied with a set of Python scripts, datasets, and pre-trained machine learning models.

We are at your disposal for any questions you may have at Or Azar Ido Shveki Sergei Benkovich

We hope you'll enjoy it!

Assignment Details

Materials Provided

1. A data folder containing two CSV files: a train dataset and a test dataset (train.csv, test.csv)
2. Two .pkl files encapsulating pre-trained models:
    a. A model that generates two output features. It takes as input a time series of raw samples. It runs daily, per sensor
        i. Feature_creator.pkl methods:
            Transform
                a. Input: train.csv / test.csv
                    i. columns: temperature, sensor_id, 'timestamp'
                b. Output: dataframe of daily features
                    i. Columns: sensor_mac_address, timestamp, daily_mean, daily_std
    b. A clustering model based on the daily features, the model runs daily on all sensors.
        Clustering_model.pkl methods:
            1. Train
                a. Input: features dataframe per feature creation step
                b. Output:None
            2. Inference
                a. Input: features dataframe
                b. Output: predictions

Tasks

1. Construct a machine learning pipeline that:

    a. Receives a path to a file containing time series samples and models
    b. Creates the features
    c. Trains the model
    d. Predicts
    e. Log the predictions

2. Choose 3 metrics to log during both training and prediction, and log them.
3. Encapsulate the flow in a docker container.
4. Provide a README file specifying
    a. How to run the entire flow
    b. How to access outputs and logs

Bonus

1. Describe (in natural language) other possible metrics you think are worth monitoring
2. Use feature_creator_bonus.pkl for the feature creation process
    a. Support the 2 feature creation pipelines, and compare results between the models
3. Create a postgres docker container and store the predictions from the model docker in the postgres container.
   You may use this snippet (but you don't have to):
   docker run --name local-postgres -e POSTGRES_PASSWORD=pa55w0rd -e POSTGRES_USER=postgres -e POSTGRES_DB=BeeHeroTask -d -p 5432:5432 postgres

Notes

While developing your solution, please consider various facets of the system and provide explanation of your decisions and considerations.

- Specify what you would have done differently given more time.
- Specify how the system should change to accommodate scale(in number of sensors, number of prediction models).
- Specify how the system would change given long processing times of the models.

Submission Guidelines

Please submit your solution on a private github repo with permissions for the following users

ido@beehero.io, or@beehero.io, sergei@beehero.io, tair@beehero.io

The repo should include:

All relevant code, Dockerfile creation file, and a README file.
The README file should contain instructions on how to build and run your solution, along with a brief explanation of the decisions you made during the assignment and any potential improvements given more time.