Machine Learning Engineer

Technical test

Instructions:

- Read carefully each of the proposed exercises. Do not hesitate to contact us in case you need additional explanation reading any of the exercises.
- For the exercises, please include the code used for your solutions. Write the code in such a way that it can be easily read and executed by the examinator, and follow good coding standard rules. We prefer you to use Python (or R), but it is not mandatory.
- You can take as much time as you like in order to complete this test. However, we would appreciate it if you can submit your answers within one week.

Tasks:

The business model of one of our products, a dating website, is a subscription based payment model. Users join the website and pay a small amount for the first day. After one day, their subscription is converted to a full amount (higher price) subscription, which is charged monthly.

Explanation of the data set:
The files contain anonymized data about the behaviour and spending patterns of our customers. We use a similar data set to model the amount a user will spend during his/her membership on our websites. This metric is called customer lifetime value (LTV).

The training set data are from users that have joined our website from Dec/2018 through June/2019. The validation set contains users that joined in July/2019, and the test set Aug/2019.

The data set contains the following columns:

- **user_id:** a numerical identifier of the user.
- **join_date:** the moment when the user joined our site.
- **hidden:** a user segment estimated when joining our site according to available data.
- **STV (short-term-value)**: the amount the user has spent on our site in the first 4 days of his activity
- **target (lifetime value):** the outcome variable. It is the amount the user has spent after a little bit more than 2 months. This variable is only contained in the training and validation set.
- **credit_card_level:** one of two types of credit cards
- **is_lp:** characterizes the traffic source
- **aff_type:** another traffic source characterization
- **is_cancelled:** did the user cancel his subscription?
- **country_segment:** from which country is the user?

First exercise:

Please develop a machine learning model with the training data set train.csv to predict the target variable.
Note that it is very ambitious to model single user behaviour, that's why we are interested in the average error of the model when aggregating many users. So please check the performance of your model by evaluating the mean absolute error of the target for each country segment when aggregating the data by

- join day
- join month

You can use the valid.csv file for the evaluation.

A simple solution here is sufficient, the focus of this test is on the second exercise.

Second exercise:

This part is about packaging your model into an API, and writing a script that communicates with the API.

1. API requirements:
    a. Connects to a database (ideally a relational database).

    b. Receives data in JSON format, and stores it in the database. Please store all data in the database.
    c. Preprocesses data to have the correct input format for the model. Make your implementation robust against missing data in some of the columns. Take also into account that in a real scenario there might be new country segments appearing and you don't have a target column.
    d. Predicts the target variable with the model from exercise 1. The predictions should be stored in the database in a separate table than the features.
    e. Returns predictions.
2. Command line script requirements:
    a. Reads validation/test data and sends it to the API.
    b. Gets the predictions either from the API or from the database, and evaluates the error of the mean by using the same aggregations as in exercise 1.
    c. Note that the provided test data set does not have a target column, but we will add it when we run your code with the test data set.
3. Bonus point (choose one):
    a. Package everything into Docker.
    b. Deploy the API to an online server.

The idea of the command line script is that you test the API and the model using the validation set, but when we evaluate your work we will use the test set. To do so we add the target column to the test set during evaluation.

We would appreciate it if you write the API and the script in Python (or in R), and we highly value that you use best coding practices with a clean architecture and concepts like ORM.

Please include instructions on how to execute your code into your submission. The version that you submit should read the test data set and evaluate the performance of the algorithm on this unknown dataset.

Looking forward to your solution!