Version 0.0.2

Customer Lifetime Value Project

Thank you for considering the Data Science Python developer position at Lesara. This take-home test is designed to gather more information about your experience and get a sense of how you approach problems we care about. Feel free to send questions directly to juliet.matsai@lesara.de if anything is unclear. Please also acknowledge the receipt of the email. The test should take you no more than half a day since we want to be respectful of your time. Our intention of the test is to also use this problem for further discussions on the on-site interview.

Problem description

Customer lifetime value (CLV) -- discounted value of future revenue generated by a customer, is particularly important in online fashion retail, since CLV helps us make important business decisions about sales, marketing, product development, and customer support. Now, you are collaborating with Lesara data scientists to work on a project, which predicts the CLV of each customer based on his/her historical orders.

About dataset

Features of the dataset consist mostly of customer order (e.g. customer ID, order ID, order item ID, number of items, revenue, order creation time). Some of the characteristic are anonymized.
Data fields:
client_id -- ID of a customer
order_id -- Order ID
order_item_id -- Item ID in the current order: one order could have multiple item IDs
num_items -- Number of items with the same item ID
created_at_date -- Order creation date

Our dataset is contained in a csv file: orders.csv
You can download the dataset from here

Task description

Our data scientists found that the most important features to predict a customer CLV are:

1. Max number of items in one order: if a customer has more than one order, take the one with more items.
2. Max revenue in one order: if a customer has more than one order, take the one with largest revenue.

3. Total revenue of a customer: including all orders
4. Total number of orders
5. Days since last order: number of days from the last order until 2017-10-17
6. The longest interval between two consecutive orders (in unit of days). If a customer has only one order, use the formula below

$$Avg(longest\ interval) + days\ since\ last\ order$$

A model has already been developed by our data scientists, and dumped as dill file. You can load the model instance using `dill.load(file)`. The model contains a method `predict`, which load the aforementioned data in numpy.array format, and return the predicted CLV. For example, an input dataset could be
numpy.array([[3, 92.6, 109.3, 2, 12, 26],[2, 10.4, 43.5, 3, 26, 5]])
And the `predict` method returns a numpy array ([244.9, 89.9]) as the predicted CLV.

Your tasks:

1. Simple ETL: transform the raw data to the desired input format of the prediction model.
2. Run the model on the dataset and output the predictions in CSV format. At least two columns [customer_id, predicted_clv] should be included in the output.
3. Design a RESTful application with Flask (for storing data you can use some data storages e.g. mysql, redis, sqlite etc). Provide one GET endpoint, which returns the predicted CLV for a given customer_id.

Requirements:

1. Required libraries should be included in the requirements.txt file.
2. The application must be runnable if installed on an average computer running with Python 3.6 or higher version.
3. Follow OOP and PEP8 principles.
4. Handle errors and exceptions.
5. Unittest should be included.
6. Please upload everything to github and provide us a link and instructions how to run app
7. Bonus point if you will do it with docker (app inside a docker container, instruction how to run and test it)

Questions to think about:

1) How would you deploy the app?
2) How to schedule the ETL job?