Data Science Assignment

Wongnai would like to thank you for your application to the Data Scientist position. In the next step, we'll ask you to complete 2 assignments within 7 days.

Assignment 1: SQL

Since SQL is a language for retrieving structured data from Wongnai's relational database, it is required that Wongnai's data scientists have good command of SQL.

Tools

1. Bigquery Sandbox mode: https://console.cloud.google.com/bigquery?p=bigquery-public-data&page=project
2. We will focus on **bigquery-public-data:san_francisco_bikeshare** dataset.

Deliverables

1. Submit an image file of the diagram for the first question.
2. Submit only ONE query for each question in ONE SQL file and name file as "sql_assignment_<number of question>.sql" for the rest of questions.

Questions (bigquery-public-data:san_francisco_bikeshare)

1. Draw an Entity-Relationship diagram, aka. ER diagram, showing the relationship between all tables using crow's foot notation. Note that there is no score for the aesthetic of the diagram.

2. Find the top 5 start stations with the highest total number of trips.

3. How many trips are there that have the region Oakland as their start station?

4. From bike-share trips that started from year 2015 onwards, find

   - Minimum, maximum, mean of duration (in seconds)
   - Which year has the highest total number of trips?

5. In 2015, find top 5 stations that got their first trip earlier than other stations. Considering only the start station, please provide trip_id and station_id only, and consider year 2015 only.

6. In 2015, given the situation where your business did not go well and you need to reduce operation cost to a certain level. Your business unit decided to shutdown some stations which perform poorly. Find a minimum station you can keep in order to maintain a number of trips per year greater than or equal 320000, consider year 2015 only.

Assignment 2: Improve promotion section in Wongnai

This assignment is about a promotion section in Wongnai which allows users to upload restaurant promotions consisting of either **image and text or both**. Some promotion may have either text or image, or some could have both combination

You can see promotions in our system here: https://www.wongnai.com/promotions

Assignments

There are 2 sub-assignments you need to complete. Assume that you're working as a Data Scientist in Wongnai and the business team requests these tasks. You need to complete the tasks while taking into account both limited time-frame (7 days) and getting the best result.

1. Spam Promotion Detection

There are too many spam promotions which we believe are not useful for users and we would like to remove them.

Therefore, our team defines spam promotions as:

- Alcohol
- People images
- Storefront
- No promotion detail
- Loyalty program
- Screenshot

If the promotion matches any of these cases, it will be considered as a spam promotion. Using these criteria, we have already labelled a number of promotions as spam/non-spam in the training set.

An example of promotion images and their descriptions can be downloaded here https://drive.google.com/file/d/13zvKkwNVjpn9Jfy6gLNrDcepEnD1NfSM/view?usp=sharing

Your tasks:

1. Given image and promotion description, create a predictive model to predict whether this promotion is spam or not.
2. Use the model from 1. to predict promotions in the testing set (testing_set.csv + testing_images). Please include the predictions in your submission..

Note: Some promotions do not contain description data. That is, the promotion is simply an image, no caption or description at all.

2. Recommendation systems in promotions

One way to increase the number of user engagement is to recommend items that users might like. There are many options for creating a recommendation system. Assume that our promotion feature in Assignment 2 part 1 above was just launched about 1 month ago and we have limited data in our hands.

Your task:

1. Given both the image and text of the promotions, use them to find similar promotions.

Note: Please focus on explaining how you would approach this problem. The goal is for us to understand your approach and reasoning. Building a simple prototype to help illustrate your approach is encouraged, but not required.

Dataset

The dataset can be downloaded via this link
https://drive.google.com/open?id=1QLbzskHxeGglvUrguZQlq1DRZfonhrt0

The dataset is used for the interview process only. Send/forward the datasets to anyone else or use the data for other purposes are prohibited.

Tools

Please only use Python or R in Jupyter Notebook/Lab. You can use any libraries in Python and you should specify all required libraries inside requirements.txt so we can install and check your results.

Criteria

Here are the capabilities we look for in a candidate. We believe they are required to solve the business problem we are facing right now.

- Problem definition
- Exploratory data analysis (EDA)
- Data preparation/wrangling
- Feature engineering
- Model selection
- Model evaluation
- Writing quality code

Please demonstrate these capabilities in your Jupyter notebook/lab as you work through the problems.

Timeline

Please complete all assignments within 7 days after the time the email has been sent.

Submission

Once you have finished the assignments, please zip all required files and submit using a link inside an email.

Please make sure that once we extract the zip and open Jupyter Notebook, we can re-run all commands and get the same result that you sent.

Do not forget to include prediction results of testing dataset in promotion spam detection task inside a zip file.

Please organize your submission as following

|- name-surname-assignment

    |- sql

        |- sql_assignment_er.jpg

        |- sql_assignment_2.sql

        |- sql_assignment_3.sql

        |- sql_assignment_4_1.sql

        |- sql_assignment_4_2.sql

        |- sql_assignment_5.sql

        |- sql_assignment_6.sql

    |- spam-detection

        |- spam.ipynb

        |- prediction.csv