

Artificial Intelligence

Project – Part 1

April 20th, 2023

Instructions:

1. Work in groups of two members.
2. You have to use python programming language.
3. Plagiarism of any kind (copying from others and copying from the internet, etc.,) is not allowed and can result in zero marks in whole assignment category.
4. Any part of the assignment marked 'AI Generated' by ZEROGPT, DETECTGPT can result zero marks.
5. Your code must be properly commented.
6. Any assignment marked late by google will be considered late.
7. No marks will be assigned if any of the following deliverables are missing.
 - a. The source code of the program.
 - b. A pdf or word report containing a brief explanation of the steps involved in the
 - c. program (each question) and the results obtained.
8. Put both source code and report in one folder, ZIP it and submit it. Your folder must be named as `ROLLNO01_NAME_ROLLNO02_NAME.ZIP`
9. Only one student is required to make submission on GCR. Multiple submissions may lead to plagiarism.

1. Background

As less than 10% of worlds citizens own automobiles, the frequency at which citizens commute on taxis, buses, trains, and planes is very high. Uber (or to some extent Careem), the dominant ride-hailing company, processes over 11 million trips, plans over 9 billion routes and collects over 50TB of data per day. To meet needs of riders, Uber must continually innovate to improve cloud computing and big data technologies and algorithms in order to process this massive amount of data and uphold service reliability. Supply-demand forecasting is critical to enabling Uber to maximize utilization of drivers and ensure that riders can always get a car whenever and wherever they may need a ride. Supply-demand forecasting helps to predict the volume of drivers and riders at a certain time period in a specific geographic area. For instance, demand tends to surge in residential areas in the mornings and in business regions in the evenings. Supply-demand forecasting allows Uber to predict demand surges and guide drivers to those areas. The end result is higher earnings for drivers and no surge pricing for riders!

2. Definition and Evaluation Criteria

1. Definition

A passenger calls a ride(request)by entering the place of origin and destination and clicking "Request Pickup" on the Uber app. A driver answers the request (answer) by

taking the order. Uber divides a city into n non-overlapping square regions $D = d_1, d_2, d_3, d_4, \dots, d_n$ and divides one day uniformly into 144 time slots $t_1, t_2, t_3, \dots, t_{144}$ each 10 minutes long. In region d_i , and time slot t_j , the number of passengers' requests is denoted as r_{ij} , and drivers' answers as a_{ij} . In region d_i and time slot t_j the demand is denoted as $\text{demand}_{ij} = r_{ij}$ and the supply as $\text{supply}_{ij} = a_{ij}$, and the demand supply gap is: gap_{ij} : $\text{gap}_{ij} = r_{ij} - a_{ij}$. Given the data of every region d_i and time slot t_j , you need to predict gap_{ij} , $\forall d_i \in D$.

2. Evaluation Metrics

Given i regions and j time slots, for region d_i in time slot t_j , suppose that the real supply-demand gap is gap_{ij} , and predicted supply-demand gap is s_{ij} , then:

$$\text{MeanAbsoluteError} = \frac{1}{n} \sum_{d_i} \left(\frac{1}{q} \sum_{t_i} |\text{gap}_{ij} - s_{ij}| \right)$$

The lowest mean absolute error will be the best.

The detailed description of each field is as follows:

| Data Name | Data Type | Example |
|------------------|-----------|---|
| Region ID | String | 1,2,3,4 (the same as region mapping ID) |
| Time slot | String | 2016-01-23-1 (The first time slot on Jan. 23rd, 2016) |
| Prediction value | Double | 6.0 |

3. Data Format

The training set contains three consecutive weeks of data for City M in 2016, and you need to forecast the supply-demand gap for a certain period in the fourth and fifth weeks of City M. The test set contains the data of half an hour before the predicted time slot. The specific time slots where you need to predict the supply-demand gap are shown in the 1 explanation document in the test set.

The Order Information Table, Weather Information Table and POI Information Table are available in the database. All sensitive data has been anonymized.

1. Order Information Table

| Field | Type | Meaning | Example |
|-------------------|--------|-------------|----------------------------------|
| order_id | string | order ID | 70fc7c2bd2caf386bb50f8fd5dfef0cf |
| driver_id | string | driver ID | 56018323b921dd2c5444f98fb45509de |
| passenger_id | string | user ID | 238de35f44bbe8a67bdea86a5b0f4719 |
| start_region_hash | string | departure | d4ec2125aff74eded207d2d915ef682f |
| dest_region_hash | string | destination | 929ec6c160e6f52c20a4217c7978f681 |
| Price | double | Price | 37.5 |

| | | | |
|------|--------|------------------------|---------------------|
| Time | string | Timestamp of the order | 2016-01-15 00:35:11 |
|------|--------|------------------------|---------------------|

The Order Information Table shows the basic information of an order, including the passenger and the driver (if driver id =NULL, it means the order was not answered by any driver), place of origin, destination, price and time. The fields order id, driver id, passenger id, start hash, and dest_hash are made not sensitive.

2. Region Information Table

The Region Information Table shows the information about the regions to be evaluated in the contest. You need to do the prediction given the regions from the Region Definition Table. In the submission of the results, you need to map the region hash value to region mapped ID.

| Field | Type | Meaning | Example |
|-------------|--------|-------------|----------------------------------|
| region hash | string | Region hash | 90c5a34f06ac86aee0fd70e2adce7d8a |
| region id | string | Region ID | 1 |

3. POI Information Table

The POI Information Table shows the attributes of a region, such as the number of different facilities. For example, 2#1:22 means in this region, there are 22 facilities of the facility class 2#1. 2#1 means the first level class is 2 and the second level is 1, such as entertainment#theater, shopping#home appliance, sports#others. Each class and its number is separated by

| Field | Type | Meaning | Example |
|-------------|--------|--------------------------|----------------------------------|
| region hash | string | Region hash | 74c1c25f4b283fa74a5514307b0d0278 |
| poi class | string | POI class and its number | 1#1:41 2#1:22 2#2:32 |

4. Weather Information Table

The Weather Information Table shows the weather info every 10 minutes each city. The weather field gives the weather conditions such as sunny, rainy, and snowy etc; all sensitive information has been removed. The unit of temperature is Celsius degree, and PM2.5 is the level of air pollutions.

| Field | Type | Meaning | Example |
|-------------|--------|-------------|---------------------|
| Time | string | Timestamp | 2016-01-15 00:35:11 |
| Weather | int | Weather | 7 |
| temperature | double | Temperature | -9 |
| PM2.5 | double | pm25 | 66 |