Deep Data First Round

Problem Statement

In the age of disruption, we cannot run away from basic needs. Money is one of them. There are so many ways to make money. Most people earn money on a regular basis via payroll.

Banks are the places for peace of mind to save our money. They are safe and take care of our money. Most of the time, banks know spending activities of their customers via all kinds of services. However, banks only know "income" from registered payroll or declared statements.

With customers' spending behavior and other financial activities, is it possible to determine their *income*? You are data scientists who are eager enough to help Kasikorn Bank estimating customer income.

Given Features

1. ID
2. Age
3. Gender
4. Credit Card number (each person may have more than 1 card)
5. Credit card spending
6. Aggregated expense via K+
7. Number of K+ activities

There are 5 datasets (tables). The data is mocked to represent financial activities during Jan 2018 - Jun 2018.

1. demographics.csv

Field Name	Data Type	Description
id	INTEGER	Dummy customer ID
cc_no	INTEGER	Dummy credit card number
gender	INTEGER	Gender 1: Male 2: Female
ocp_cd	INTEGER	Encoded group of occupations. We don't give details on this. Example of groups of occupations: "students"
age	INTEGER	Range of ages 0: [0-15], 1: [16-25] 2: [26-35], 3:[36-45] 4: [46-55], 5: [56+]

Example:

id	cc_no	gender	ocp_cd	age
1	98397	2	9	5

2. cc.csv (credit card spending)

Field Name	Data Type	Description
cc_no	INTEGER	Dummy credit card number
pos_dt	STRING	Date of the transaction Format: yyyy-mm-dd
cc_txn_amt	FLOAT	Spending amount

Example:

cc_no	pos_dt	cc_txn_amt
37069	2018-05-10	5000
37069	2018-06-04	12000
37069	2018-04-03	5000
37069	2018-04-22	1600
37069	2018-01-21	5000

3. kplus.csv (K+usage)

Field Name	Data Type	Description
id	INTEGER	Dummy customer number
sunday	STRING	Aggregated data is represented by the end of each week. *For example, 2018-05-27 represents the aggregate data of 2018-05-21 to 2018-05-27
kp_txn_count	INTEGER	Frequency of K+ usage per week
kp_txn_amt	FLOAT	Monetary amount of K+ usage

Example:

id	sunday	kp_txn_count	kp_txn_amt
14802	2018-01-14	2	2400
14802	2018-04-01	9	33900

4. train.csv (Training set with labels)

Field Name	Data Type	Description
id	INTEGER	Dummy customer number
income	FLOAT	Income labels for training

Example:

id	income
1	20000
2	106000

5. test.csv (list of customers to predict

Field Name	Data Type	Description
id	INTEGER	Dummy customer number

Example:

id
50001
50002

The score from the submission engine online might not reflect the real score. It is just a guideline score. Not every row is evaluated.

Metric of evaluation

Modified SMAPE (symmetric mean absolute percentage error)

$$Score = 100 - \frac{100}{N}\sum_{i=1}^{N}\frac{|F_i - A_i|^2}{(\min(2|A_i|, |F_i|) + |A_i|)^2}$$

where A_i = Actual value (answer key), F_i = Forecast value
*Note, this score can be negative. If your answer is negative, the system will show your score as 0.
** If you get score = -1, this means there is something wrong about your submitted file(s).

Output Format

The output must follow the format:

1. Two columns: ID (INTEGER) and Income (FLOAT).
2. Must include the header (column names).
3. Include every ID and only **ONCE.**

If the format is violated, the predicted income of that ID will be assigned to 0 automatically.

Example of output

id	income
55001	21321.32
55002	32293.01
55003	29329.93
55004	12000.00

Final Submission

For the *final answer*, please submit .zip file. This zip file **MUST** contain

1. Model file(s)
   - Use your team number as filename:
     M_[team_name].xxx
2. Output file
   - Use your team number as filename:
     O_[team_number].csv
3. Other dependency file(s)

Remark: Some of the data provided maybe incomplete, inconsistent, missing, noisy, erroneous, etc. as can occur in the real-world setting. It is the participants' task to recognize such cases as the challenge intentionally posed by the problem designer.

If there's any concern or question, please contact Facebook page TechJam Thailand (www.facebook.com/TechJamThailand)