Deep Data First Round

โจทย์รอบที่ 1

ในยุคปัจจุบันที่โลกของเรามีความเจริญก้าวหน้าในด้านต่างๆ อย่างรวดเร็ว “เงิน” ยังคงเป็นสิ่งสำคัญในการดำเนินชีวิตของคนทั่วไป และคนเราต้องทำงานเพื่อหาเงินเพื่อใช้ดำเนินชีวิต โดยคนบางกลุ่มจะได้รับเงินจากองค์กรที่ตนทำงานเป็นรายเดือน หรือที่เรียกว่า “เงินเดือน” นั่นเอง

ธนาคารเป็นที่คอยเก็บรักษาเงินและดูแลกิจกรรมทางการเงินต่างๆ ของเรามักจะรู้ถึงข้อมูลเงินเดือนของประชาชนทั่วไปได้ผ่านบริการของธนาคาร แต่ถึงกระนั้น ธนาคารจะรู้ข้อมูลเงินเดือนเฉพาะข้อมูลของคนที่ผูกบัญชีเงินเดือนกับธนาคาร หรือมีเอกสารมายืนยันเท่านั้น

พวกคุณ ในฐานะนักวิทยาศาสตร์ข้อมูล จะสามารถนำข้อมูลที่กำหนดให้เกี่ยวกับพฤติกรรมการใช้เงินของผู้ใช้งานในธนาคาร เพื่อทำนาย “เงินเดือน” ของผู้ใช้งานกลุ่มนั้นได้หรือไม่

ข้อมูลที่ให้ (โดยสังเขป)

1. ID
2. Age
3. Gender
4. Credit Card number (แต่ละคนสามารถมีบัตรได้มากกว่า 1 ใบ)
5. Credit card spending
6. สรุปรายจ่ายผ่าน K+
7. จำนวนการใช้ K+

ในข้อมูลที่เราให้ไปจะมี 5 ไฟล์ ซึ่งข้อมูลนี้เป็นการจำลองจากการทำธุรกรรมทางการเงินของลูกค้าธนาคารช่วง ม.ค. 2561 ถึง มิ.ย. 2561

1. demographics.csv

Field Name	Data Type	Description
id	INTEGER	ID ลูกค้า
cc_no	INTEGER	เบอร์บัตรเครดิตจำลอง
gender	INTEGER	เพศ 1: ผู้ชาย 2: ผู้หญิง
ocp_cd	INTEGER	รหัสอาชีพ ทางเราจะไม่ให้ความหมายของแต่ละรหัสอาชีพ ตัวอย่างของอาชีพ “students”
age	INTEGER	ช่วงอายุ 0: [0-15], 1: [16-25] 2: [26-35], 3: [36-45] 4: [46-55], 5: [56+]

ตัวอย่าง:

id	cc_no	gender	ocp_cd	age
1	98397	2	9	5

2. cc.csv (credit card

Field Name	Data Type	Description
cc_no	INTEGER	เบอร์บัตรเครดิต จำลอง
pos_dt	STRING	วันของการใช้จ่าย Format: yyyy-mm-dd
cc_txn_amt	FLOAT	จำนวนเงิน

ตัวอย่าง:

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 (การใช้งาน K+)

Field Name	Data Type	Description
id	INTEGER	ID ลูกค้า
sunday	STRING	วันสรุปยอดของแต่ละสัปดาห์ (โดยใช้วันอาทิตย์เป็นตัวแทน) *ตัวอย่าง 2018-05-27 หมายถึง วันสรุปยอดตั้งแต่ 2018-05-21 ถึง 2018-05-27
kp_txn_count	INTEGER	จำนวนครั้งในการใช้จ่ายผ่าน K+ ของสัปดาห์นั้นๆ
kp_txn_amt	FLOAT	จำนวนเงินที่ใช้จ่ายผ่าน K+ ของสัปดาห์นั้นๆ

ตัวอย่าง:

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	ID ลูกค้า
income	FLOAT	รายได้ (คำตอบสำหรับการ train โมเดล)

ตัวอย่าง:

id	income
1	20000
2	106000

5. test.csv (ลิสลูกค้าสำหรับส่งคำตอบ)

Field Name	Data Type	Description
id	INTEGER	ID ลูกค้า

ตัวอย่าง:

id
50001
50002

ตัวตรวจคะแนนบนเวปจะบอกตัวอย่างคะแนนจากการสุ่มตรวจบางข้อมูลเท่านั้น ซึ่งไม่ได้แสดงถึงคะแนนจริงที่จะใช้ในการวัดผลไปรับรอบสุดท้าย

วิธีการวัดผล

Modified SMAPE (symmetric mean absolute percentage error)

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

โดยกำหนดให้ A_i = ค่าจริง (คำตอบ), F_i= ค่าที่ส่งมาให้ระบบตรวจ

**หมายเหตุ

- คะแนนนี้อาจติดลบได้ ถ้าเกิดคะแนนติดลบ ระบบจะแสดงคะแนนให้เป็น 0
- ถ้าได้คะแนน -1 หมายถึงเกิด error จากตัวไฟล์ เช่น ส่งไฟล์ผิดฟอร์แมต

Output Format

ส่งคำตอบมาตาม format นี้

1. ไฟล์คำตอบมี 2 columns: ID (INTEGER) และ income (FLOAT).
2. ต้องใส่ Header (column names) มาด้วย
3. ต้องใส่ครบทุก ID ที่ให้ไป และไม่มี ID ซ้ำ

*ถ้าส่งคำตอบที่ไม่ตรงตามเงื่อนไข ระบบจะถือว่าคำตอบของบรรทัดนั้น predicted income เป็น 0

ตัวอย่าง ตารางคำตอบ

id	income
55001	21321.32
55002	32293.01
55003	29329.93
55004	12000.00

การส่งคำตอบสุดท้าย

สำหรับคำตอบสุดท้าย ให้ส่งมาในฟอร์แมต .zip ซึ่งใน zip file นี้ต้องประกอบไปด้วย

1. Model
   - ใช้ M_ ตามด้วย เลขทีมของผู้เข้าแข่งขันเป็นชื่อไฟล์: M_[team_name].xxx
2. Output
   - ใช้ O_ ตามด้วย เลขทีมของผู้เข้าแข่งขันเป็นชื่อไฟล์: O_[team_number].csv
3. ไฟล์อื่นๆที่เกี่ยวข้องที่ใช้ในการ run model

หมายเหตุ: ข้อมูลที่ให้ไปอาจมีความไม่สมบูรณ์ไม่สอดคล้อง ผิดพลาด และอื่นๆ ซึ่งเป็นความตั้งใจของการออกแบบโจทย์เพื่อจำลองสถานการณ์ของการทำงานจริง

หากมีข้อสงสัยใดๆ เกี่ยวกับการทำโจทย์สามารถติดต่อสอบถามรายละเอียดเพิ่มเติมได้ทาง inbox ของ TechJam Thailand (www.facebook.com/TechJamThailand)