TECHJAM 2018 AUDITION — DATA SQUAD
"LOAN DEFAULT"

การเก็บดอกเบี้ยจากการให้ลูกค้ากู้เงินถือเป็นรายได้หลักอย่างหนึ่งของธนาคาร แต่อย่างไรก็ตาม การเก็บดอกเบี้ยจากการให้ลูกค้ากู้เงินนั้นต้องเสี่ยงต่อการที่ลูกค้าชำระหนี้ล่าช้า หรือไม่ยอมจ่ายหนี้ตามจำนวนเงินที่กำหนด ดังนั้นงานสำคัญอย่างหนึ่งของธนาคาร คือ การประเมินความเสี่ยงของลูกค้าที่จะมากู้เงินกับเรา หน้าที่ของคุณ คือ ทำนายว่าลูกค้าที่จะมากู้เงินกับเราแต่ละคนจะเบี้ยวการชำระหนี้หรือไม่

Loan interest is one of the major revenue of the bank. However, bank has to face risk when offering loans to the customers; for example, customers do not pay the installment payment within the due period, or they default on a loan. Therefore, the important task of the bank is risk management, to evaluate the risk of each customer who would like to apply loans. Your task is to predict whether customers who would like to apply loans will default on a loan or not.

I/O

Training Set

- File name: y_train.csv
    - Dummy customers' user id and default labels for training
    - Size: 6,361 rows

Field Name	Data Type	Description
ip_id	INT	Dummy user id
label	INT	0: Not Default 1: Default

หมายเหตุ: ข้อมูลที่ให้ไปอาจมีความไม่สมบูรณ์ ไม่สอดคล้อง ผิดพลาด และอื่นๆ ซึ่งเป็นความตั้งใจของการออกแบบโจทย์เพื่อจำลองสถานการณ์ของการทำงานจริง

TEST SET

- File name: y_test_index.csv
    - Dummy customers' user id for default prediction
    - Size: 3,938 rows

Field Name	Data Type	Description
ip_id	INT	Dummy user id

DEMOGRAPHIC DETAILS

- File Name: demo.csv
    - Dummy personal information of customers
    - Size: 10,299 rows

Field Name	Data Type	Description
ip_id	INT	Dummy user id
brth_yr	INT	Birthday in year
act_strt_dt	STRING	Account start date (the day customers become Kbank's customers)
no_of_dpnd_chl	INT	Number of dependent child
cis_income	INT	Salary
crn_bal	INT	Current balance in saving account
gnd_cd	INT	Gender 0: Female 1: Male
mar_st_cd	INT	Marital status code 1: Single 2: Married 3: Widow 4: Single

หมายเหตุ: ข้อมูลที่ให้ไปอาจมีความไม่สมบูรณ์ ไม่สอดคล้อง ผิดพลาด และอื่นๆ ซึ่งเป็นความตั้งใจของการออกแบบโจทย์เพื่อจำลองสถานการณ์ของการทำงานจริง

ctf_tp_cd	INT	Education level code 1: Lower high school 2: High school 3: Vocational diploma 4: Bachelor degree 5: Master degree 6: Doctorate
ocp_cd	INT	Occupation code 1: Government 2: State enterprise employee 3: Employee 4: Own business 5: House wife 6: Student 7: Freelance 8: Agricultural 9: Worker 10: Priest 12: Other (Work) 13: Other (Freelance)

ADDRESS DETAILS

- File Name: adr.csv
    - Dummy address information of customers
    - Size: 4,737 rows

Field Name	Data Type	Description
ip_id	INT	Dummy user id
prov	STRING	Province of residence

หมายเหตุ: ข้อมูลที่ให้ไปอาจมีความไม่สมบูรณ์ ไม่สอดคล้อง ผิดพลาด และอื่นๆ ซึ่งเป็นความตั้งใจของการออกแบบโจทย์เพื่อจำลองสถานการณ์ของการทำงานจริง

Saving account balance details

- File Name: sa.csv
    - Dummy saving account balance details
    - Size: 478,658 rows

Field Name	Data Type	Description
ip_id	INT	Dummy user id
dt	INT	Transaction date
channel	STRING	Channel of transaction
tp	STRING	Type of transaction DR: Debit CR: Credit
amt	INT	Transaction Amount

Credit card transaction details

- File Name: cc_txn.csv
    - Dummy credit card transaction details
    - Size: 627,894 rows

Field Name	Data Type	Description
ip_id	INT	Dummy user id
dt	INT	Transaction date
category	STRING	Category code
card	STRING	Card Number Hash
txn_amt	INT	Transaction amount

หมายเหตุ: ข้อมูลที่ให้ไปอาจมีความไม่สมบูรณ์ ไม่สอดคล้อง ผิดพลาด และอื่นๆ ซึ่งเป็นความตั้งใจของการออกแบบโจทย์เพื่อจำลองสถานการณ์ของการทำงานจริง

CATEGORY CODE

The category code description from credit card log is as follows:

No.	Main Category	No.	Main Category
cat1	Fashion and Apparel	cat9	Sports
cat2	Health and beauty	cat10	Children
cat3	Food and Beverage	cat11	Services
cat4	Appliance and Electronics	cat12	Education
cat5	Office supplies, books and gift shop	cat13	Pet
cat6	Automotive shops and Vehicles	cat14	Travel
cat7	Entertainment	cat15	Accommodation
cat8	Home and Garden	cat16	Others

OUTPUT FILE PREPARATION

ห้ามเปลี่ยนลำดับของ ip_id ในไฟล์ y_test_index.csv และ **ส่งเฉพาะคำตอบที่ได้จากโมเดลทำนายเป็นค่าความน่าจะเป็น (ค่าอยู่ระหว่าง 0 ถึง 1)** โดยที่ไม่ต้องใส่ column header

ตั้งชื่อไฟล์คำตอบของคุณว่า **TJ2018-AUDITION-[TEAM-ID].csv**

ตัวอย่างข้อมูลในไฟล์คำตอบ

0.2346236487
0.8375012047
0.9238235609
0.0239002407
0.5034234371

หมายเหตุ: ข้อมูลที่ให้ไปอาจมีความไม่สมบูรณ์ ไม่สอดคล้อง ผิดพลาด และอื่นๆ ซึ่งเป็นความตั้งใจของการออกแบบโจทย์เพื่อจำลองสถานการณ์ของการทำงานจริง

You must sort the predicted label as in the order of *ip_id* in y_test_index.csv file and send **only the predicted probability (values between 0 and 1)** without the column header as the format below.

File Name: **TJ2018-AUDITION-[TEAM-ID].csv**

Ex. 0.2346236487
0.8375012047
0.9238235609
0.0239002407
0.5034234371

EVALUATION

การให้คะแนนขึ้นอยู่กับ accuracy ของโมเดลของคุณ โดยดู **AUC ของ ROC curve** เป็นหลัก
The evaluation is based on **AUC of ROC curve**.

หากมีข้อสงสัยใดๆ เกี่ยวกับการทำโจทย์สามารถติดต่อสอบถามรายละเอียด ดูเพิ่มเติมได้ทางอีเมล์ contact@techjam.com หรือ inbox ของ TechJam Thailand (http://www.facebook.com/TechJamThailand)

หมายเหตุ: ข้อมูลที่ให้ไปอาจมีความไม่สมบูรณ์ ไม่สอดคล้อง ผิดพลาด และอื่นๆ ซึ่งเป็นความตั้งใจของการออกแบบโจทย์เพื่อจำลองสถานการณ์ของการทำงานจริง