TERM PROJECT

KAI 1Q

20200045 Geonwo Kim, 20210240 Sumin Park , 20210297 Seongjae Seo

Contents

01 DataSet/Task Description

DATASET

- Information of 80,000 users using Hana 1Q pay for 238 days

Input

1) Categorical: Region / Age / Regiion

2) Cardinal : # of logins / logins with money transfer / Duration of staying with the app

Label

{1 = small business owner , 0 = general(non-business) }

TASK

Task 1 - Predict the probability of 'small business' for a given 20,000 another users

Task 2 - Select user to send pop-up add

Task 3 - Select users to invite to the survey promotion

02 Data Preprocessing

Target Encoder

Why is it need?

Region and age data

- We need to change Categorical to Cardinal
- These are high cardinality features with low redundancy.

How to Solve?

The target encoder method was adopted

Region and age data ⇨ # ∈[0,1]

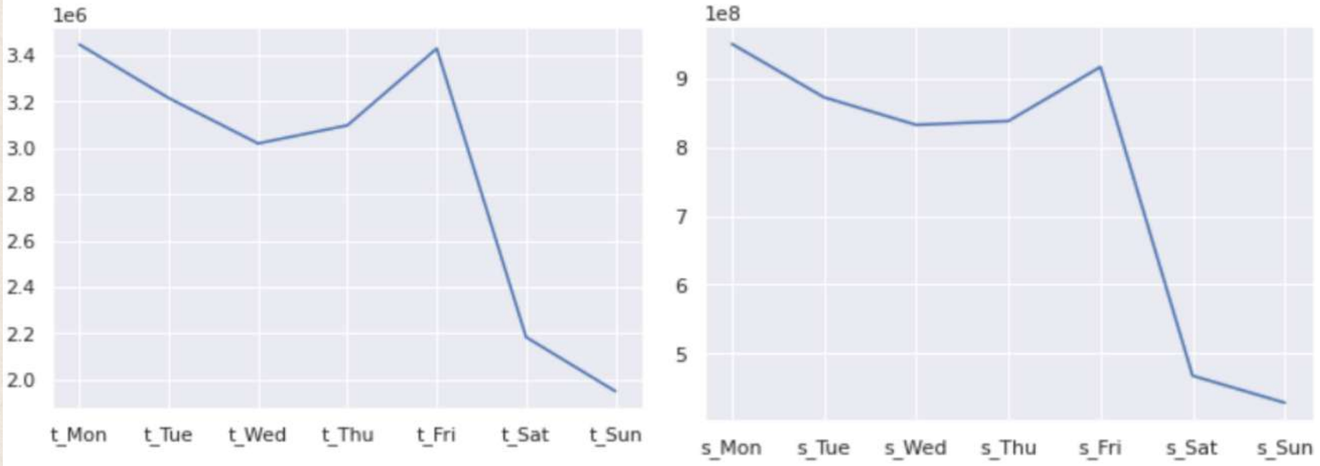
Fill "NaN" to "0"

There is an **empty cell** in Excel.

This means zero and mu st be filled with zero.

02 Data Preprocessing

Reduce Number of Data



The tendency of business owner and general people by day of the week is similar.

Decide to reduce the number of data by combining data per week.

1/1 (Sat)	1/2 (Sun)	1/3 (Mon)	1/4 (Tue)	1/5 (Wed)	1/6 (Thu)	1/7 (Fri)

Average of 7 days

Total

238 / 7 = 34 datas

about c,s,t,s_c,t_c

03 Task I - Small Business Owner Prediction

Goal : To predict the probability of small business owner

1. Split the Data

Before we first train the model, we randomly selected 20 percent of the total data given and designated it as validation data.

2. Training Models

After training using several different models, the model with the highest AUC score was selected.

Candidate

Random Forest
CatBoost
LGBM
Adaboost
Bagging Classifier

LGBM Classifier

03 Task Ⅰ - Small Business Owner Prediction

Goal : To predict the probability of small business owner

3. Hyperparameter Tuning

Before learning the model, you should identify the best hyperparameter.

HYPERPARAMETER

boosting_type='dart'	num_iterations = 1000
learing_rate = 0.01	metric = 'binary_logloss'
objective ='binary'	verbose=-1
	class_weight={0:1, 1:15})

other values are default values

03 Task Ⅰ - Small Business Owner Prediction

Goal : To predict the probability of small business owner

Result

After training the test data with a model with a given hyperparameter,
the result value of the validation data is expected, and the following accuracy is shown.

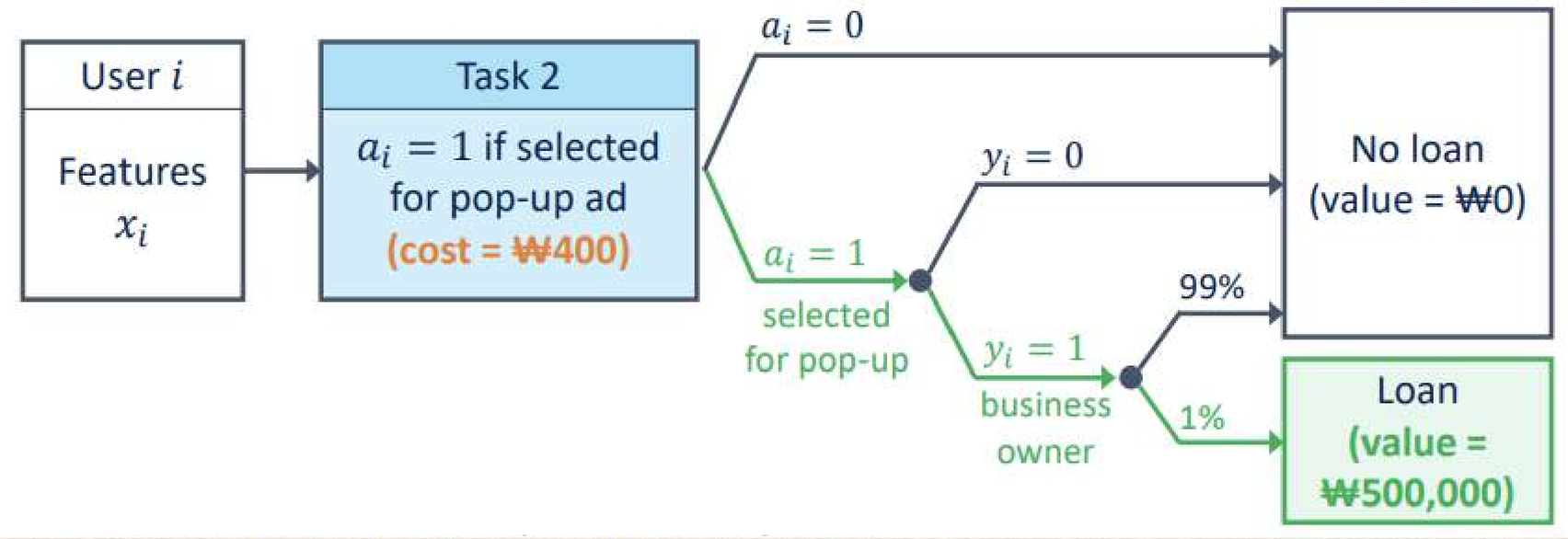
	AUC Score	F1 Score
Training	0.911	0.75
Validation	0.879	0.74

04 Task II - Pop-up Ad Planning

Goal : Decide whether to send pop-up ad

1. Process Description

Only probability of small business owner is a variable that changes net profit.



After calculating the threshold for the probability of small business owner, the a dvertisement should be sent to people with p values higher than the threshold.

04 Task II - Pop-up Ad Planning

Goal : Decide whether to send pop-up ad

2. Caculate Threshold

By Net Profit

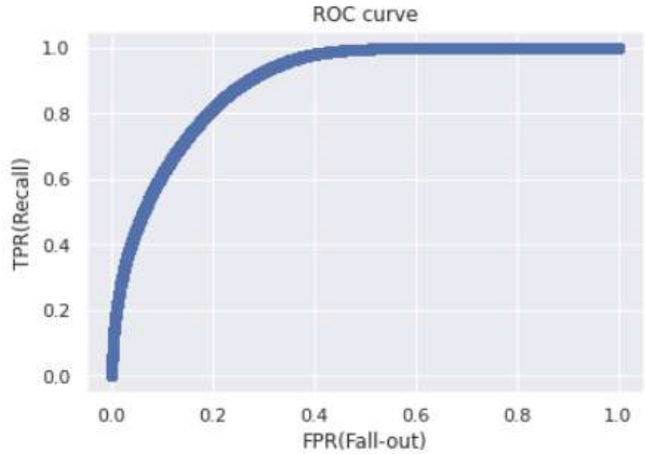
When a pop-up add is sent for each individual, the net profile is as follows

{p*0.01*500,000-400}

p : probability of small business owner

∴ **Threshold = 0.08(8%)**

By AUC-ROC CURVE



Threshold(tpr - fpr highest) = 0.05 (5%)

More than 8% by net profit, it is already beneficial, so the threshold is chosen as 0.08.

Total number of people selected = 11,1801

05 Task III - Survey Ad Planning

Goal: Decide whether to invite him/her to the survey promotion

(Calculate net profit)

$$\{p'*0.18*p*0.2*500,000 - p'*0.18*5,000\}$$

p = probability of small business owner, p' = probability of login during 5 days

Threshold of p = 0.05(5%)

# of people with a net profit of 0 or higher
117,912

Restrction #
50,000

Need to erase more people based on the number of recent connections

05 Task III - Survey Ad Planning

Goal: Decide whether to invite him/her to the survey promotion

Assumption	**The more people you invite, the higher your net profile is.** Because the threshold is low and the expected effect is higher than the risk. ⟹ the number of people to choose should be the upper limit of 50,000. **To reduce the number of people, the threshold should be at least 0.5.**

(Analyze the number of logins during August)

The accounts of people who have rarely accessed in August are judged as sleeper accounts.

People classified as sleeper accounts are excluded from the invitation of the survey promotion.

05 Task III - Suvey Add Planning

Goal: Decide whether to invite him/her to the survey promotion

1. Sleeper account classification criteria

Let. Sleeper Account = people with less than or equal to N access during the month of August
Rule: If 50,000 people are selected for those who exceed N times, the threshold must be at lea st 0.5.

➪ **N ≥ 2 : Only people with more than 3 connections are handled**

2. Select the people

- Except for those with less than 3 days of access during August
- Among them, 50,000 people are selected in the order of high pr obability of small business owners

Threshold 0.509

06 Analysis & Discussion

TASK 1

It is regrettable that the AUC score is relatively low at 0.879.

Improvement

Train with more models and then do the voting

Need additional preprocessing process

ex) Dividing it into weekends and weekdays, Calculating correlation and deleting very low values

TASK 2&3

Task2 and Task3 had many difficulties in selecting threshold.

Improvement

Since there is also a correlation between variables, the influence of other features should be considered.

Add process of testing whether Threshold is suitable is necessary

TEAMMATE

20200045 Geonwo Kim	35%
20210240 Sumin Park	30%
20210297 Seongjae Seo	35%