Term Project Poster

KrispyKreme

Jaeyun Song 20190327
Jaewoo Kim 20180140
Seonkyu Lee 20220487

Contents

Task1 : Small Business Owner Prediction

At first, we designed simple but representative features to our dataset for the baseline.

- number of days with login
- number of days with transaction
- number of days with non-zero duration
- maximum number of logins
- maximum number of transaction
- maximum duration

Model	Training AUC-ROC Score	Validation AUC-ROC Score
XGBoost	0.874	0.873
LightGBM	0.883	0.877
CatBoost	0.899	0.878
AdaBoost	0.869	0.869
GradientBoost	0.873	0.872
Stacking Classifier	0.889	0.878

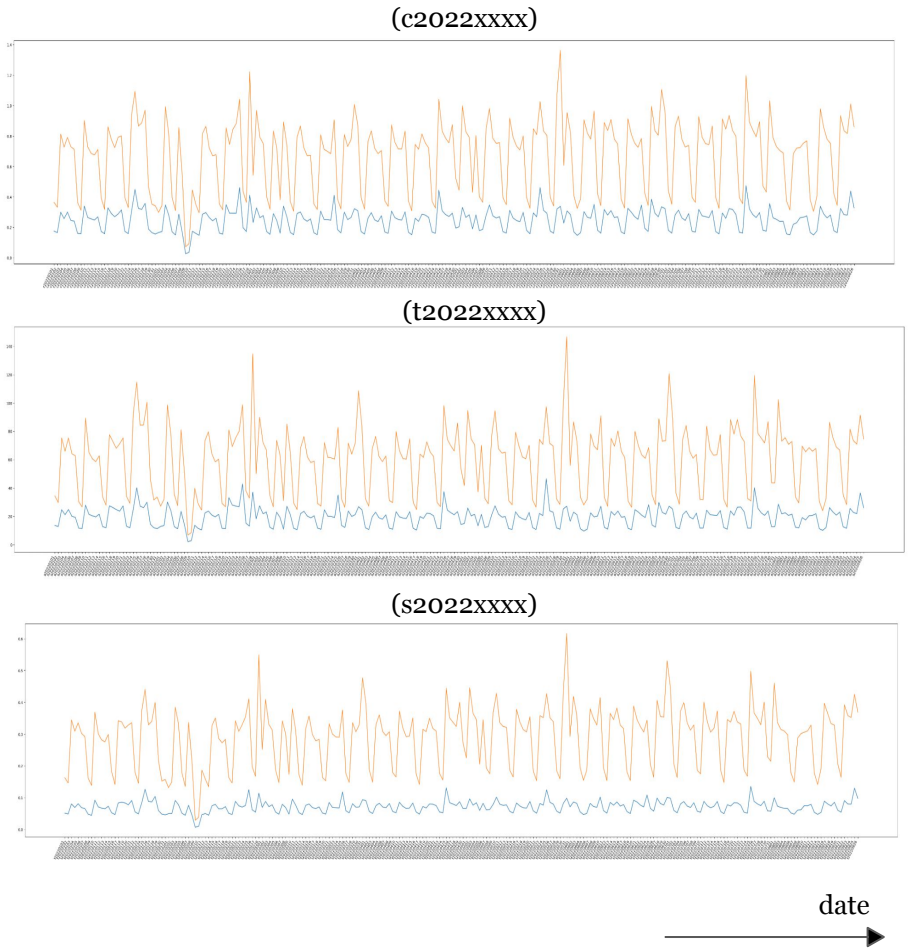
* Stacking classifier :
  - base estimators : CatBoost, LightGBM
  - final estimator : - LogisticRegression

- We found that designing good feature affects a lot on performance of prediction.
- With this top 3 baseline model, we tried to find helpful features for our prediction performance

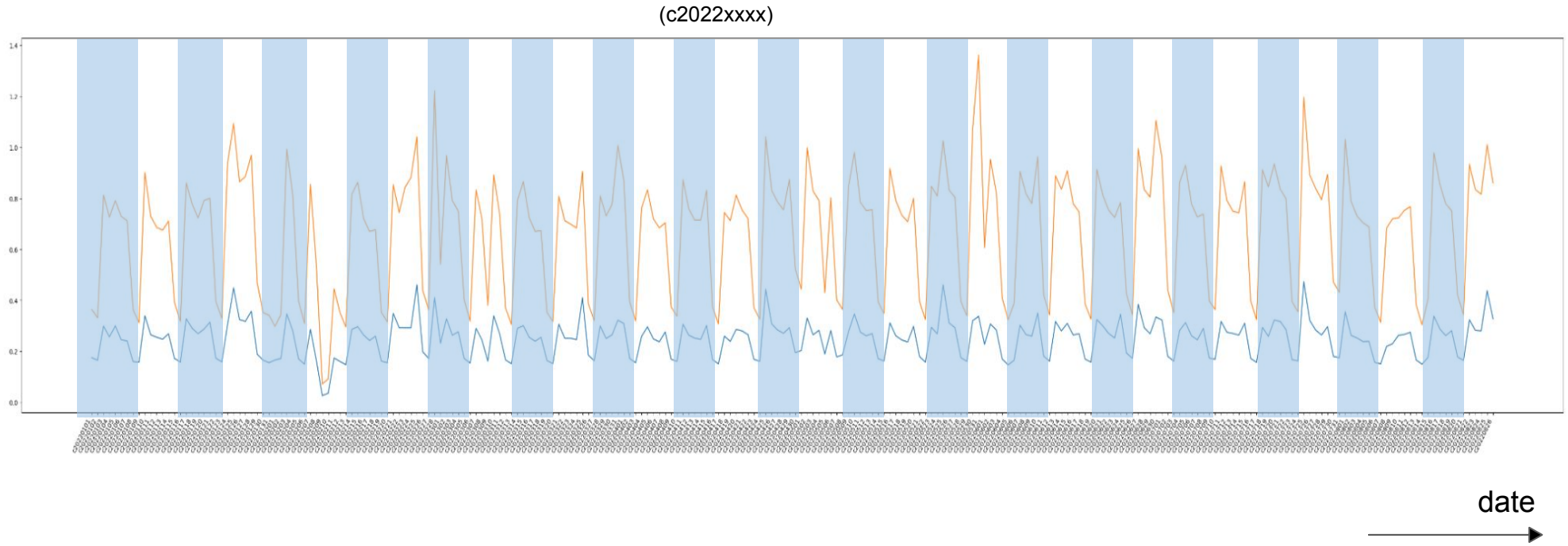
Task1 : EDA / Feature Extraction

We analyzed time series features by visualization over time period.

- We visualized c, s, t values over time period classifying users by business.
- Orange line represent the average value of each date for small business owners. Blue line represent the average value of each date for other users.
- All of three features look having similar periodicity and tendency in terms of high or low peaks, and thus we decided to focus especially on which time feature is notable for characterizing business owners.



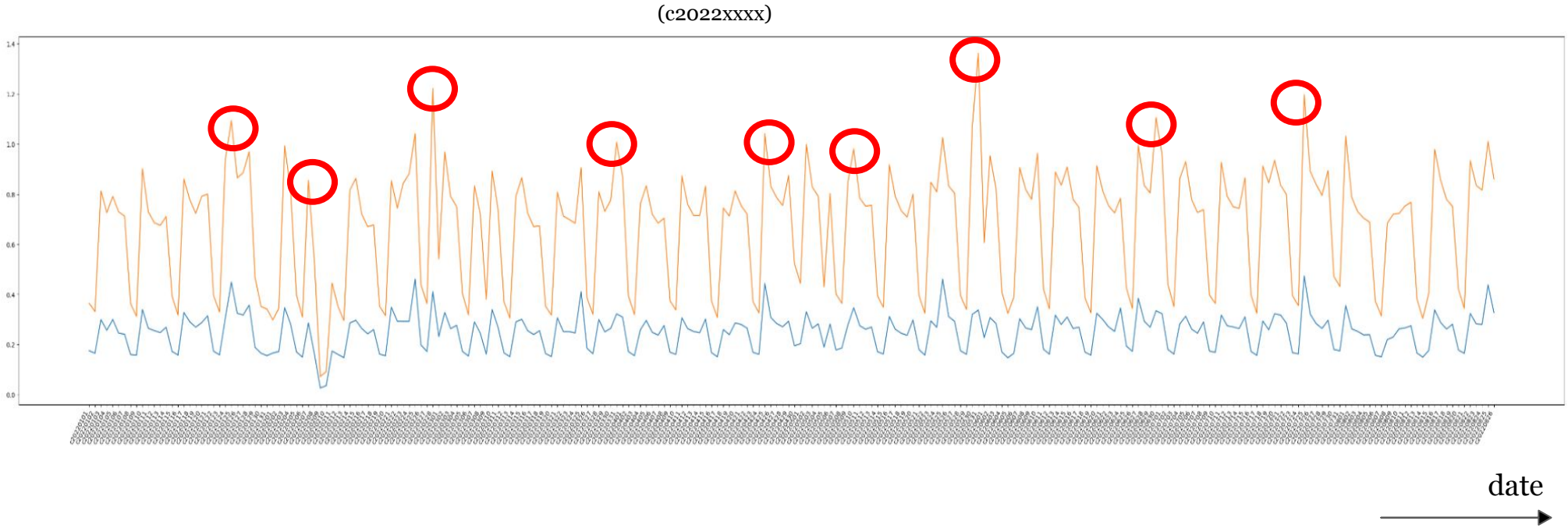
Task1 : Feature extraction



We observed there is an obvious periodicity for weeks :

- It shows low peak at each weekend.
- It shows high peak at Monday and Friday.
  (a day after weekends and before weekends)
- With more examination, we found that same tendency is shown
  not only to the weekend, but also non-weekend holiday
- We added features using the periodicity and tendency that occurs for each holiday.

Task1 : Feature extraction



We also examined the high peaks:

- There are huge high peaks at each first and last businessday of each month.
- Each 10th and 25th for each month have relatively higher value.
- With more examination, we found that same tendency is shown not only to the weekend, but also non-weekend holiday
- We conclude that these dates are related to business or finance (for example, every 25th is payday for many people)

Task1 : Model Selection

We also experimented with features for #login/#transaction average of each date (1th, 2nd, 3rd...) and standard deviation for #login.

- number of days with login
- number of days with transaction
- number of days with non-empty duration
- maximum number of logins
- maximum number of transaction
- maximum duration

} baseline features

+

Average value of #login/#transaction over each date.

Standard deviation of #login over each date.

Average value of #login/#transaction/duration over businessdays

Average value of #login/#transaction/duration over holidays

Average value of #login/#transaction/durationn over the businessdays 1 day before holiday

Average value of #login/#transaction/duration over the businessdays 1 day before holidays

Average value of #login/#transaction/durationn over the first businessdays of the month

Average value of #login/#transaction/duration over the last businessdays of the month

Average value of #login/#transaction/durationn over the first businessdays of the month

Average value of #login/#transaction/duration over the last businessdays of the month

Average value of #login/#transaction/durationn over the first businessdays of the month

Average value of #login/#transaction/duration over the last businessdays of the month

Model	Validation AUC-ROC Score
LightGBM	0.8954
CatBoost	0.8979
Stacking Classifier	0.8987

This is our final model for task1!

Task2 : Pop-up Ad Planning

- We decided to pop-up advertisement for user who is predicted to own small business

- Expected reward when business prediction is correct : 500000
- Pop-up cost : 400 (whenever pop-up)

$\gamma_{ij} = profit\ if\ decide\ i\ as\ j$

Expected profit for decision 0 : $\hat{p}_0\gamma_{00} + \hat{p}_1\gamma_{10}$

Expected profit for decision 1 : $\hat{p}_0\gamma_{01} + \hat{p}_1\gamma_{11}$

0 : non-business owner
1 : small business owner

$$\begin{pmatrix} \gamma_{00} & \gamma_{10} \\ \gamma_{01} & \gamma_{11} \end{pmatrix} =$$

	0	1
0	0	-400
1	0	4600

Profit Matrix

For optimal expected profit decision,

$$\hat{p}_0\gamma_{01} + \hat{p}_1\gamma_{11} \geq \hat{p}_0\gamma_{00} + \hat{p}_1\gamma_{10}$$

$$(1 - \hat{p}_1)\gamma_{01} + \hat{p}_1\gamma_{11} \geq 0$$

$$\hat{p}_1 \geq \frac{\gamma_{01}}{\gamma_{01} + \gamma_{11}} = 0.08$$

Task2 : Model Selection

We find the optimal threshold by experiment as well as our theoretical analysis with profit matrix.

Model	Threshold	E[Profit]
LightGBM	**0.81**	**174.583**
CatBoost	**0.08**	**175.955**
Stacking Classifier	**0.85**	**175.859**

CatBoost model showed the best performance with threshold 0.08.

- The experimental decision threshold value coincides with our analysis.
- **We selected CatBoost for Task 2.**

Task3 : Expected Profit

$A = number\ of\ people\ invited\ on\ survey\ and\ logged\ in\ during\ Aug\ 27-31$
$P_s = Probability\ of\ participating\ on\ survey = 0.18$
$B = number\ of\ people\ who\ participated\ on\ survey = A * P_s$

$C = number\ of\ business\ owners\ who\ participated\ on\ survey$
$P_i = Probability\ of\ using\ loan\ service = 0.2$
$D = expected\ number\ of\ business\ owners\ to\ use\ loan\ service = C * P_i$

$$Expected\ Profit = -5000B + 500{,}000D$$

Task3 : Strategy

Strategy 1

1. Make a login_model to predict whether to log in during July 27-31th.
   To make a login_model, train data from Jan 1st to July 26th.
2. Using business probability of task 1, calculate login & business probability.
   login & business probability = login probability * business probability.
3. Select top 50,000 users whose login & business probability is high.

Strategy 2

1. Using business probability of task 1, calculate login & business probability.
   Select top 50,000 users whose login & business probability is high.

Performance of login_model

```
CatBoost
  Training   AUROC score = 0.8954
  Training   Precision = 0.807
  Training   True Positive Rate = 0.741
  Training   f1 = 0.772
  Validation AUROC score = 0.8827
  Validation Precision = 0.784
  Validation True Positive Rate = 0.722
  Validation f1 = 0.752
```

Choose Strategy 2

	Data for Strategy 1
X_train	January 1 ~ July 26
Y_train	July 27 ~ July 31

Strategy	Expected Profit(₩)
1	112,632,750
2	114,090,750

Discussion

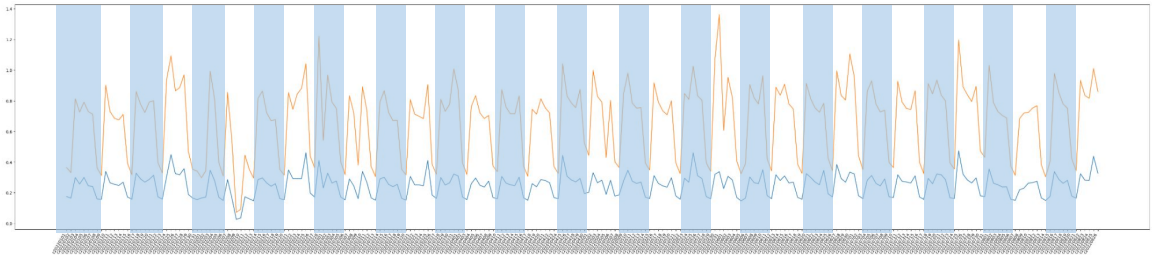
What we judged,

Improving the **performance of task 1 model is the most important** since it affects both task 2 and 3's profit.

To improve AUC-ROC score of task 1, **'Feature Engineering' is critical.**

- Using EDA, we tried to add new features which is related to business probability. Visualising data was very helpful for analysing features.

(c2022xxxx)

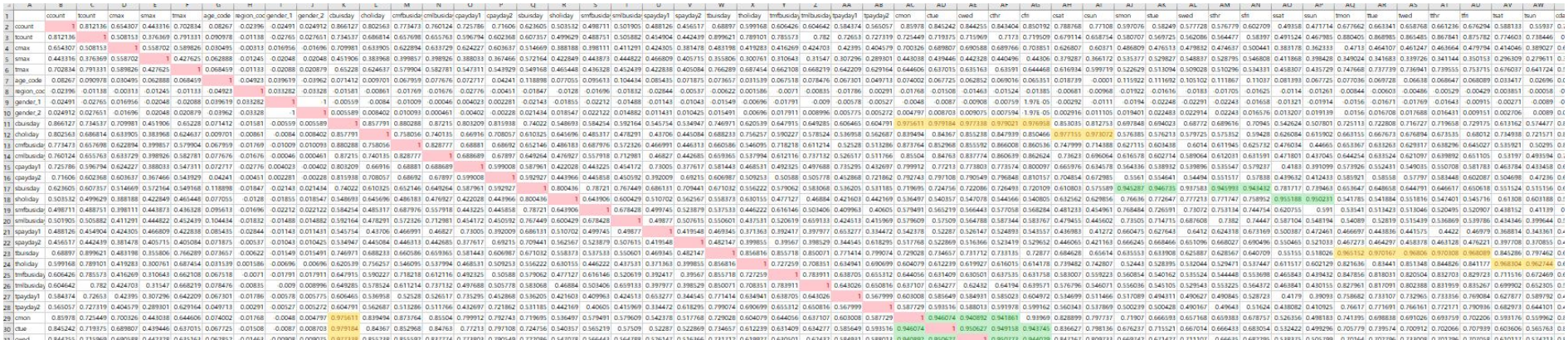


- Using PCA, we tried to reduce the dimension of features. However, applying PCA didn't affect AUC-ROC critically so we didn't applied it for the final model.

	eigen	기여율	누적
0	1.221058e+08	9.957277e-01	0.995728
1	4.945100e+05	4.032547e-03	0.999760
2	8.375739e+03	6.830106e-05	0.999829
3	6.181434e+03	5.040732e-05	0.999879
4	4.004972e+03	3.265907e-05	0.999912
5	2.819360e+03	2.299084e-05	0.999935

Discussion

Using corr() function, we analysed the correlation between features.
We eliminated features which have high correlation coefficient, but it didn't affected on AUC-ROC critically so we didn't applied it for the final model.



	Range of Correlation Coefficient
Red	0.98 ~ 1
Yellow	0.96 ~ 0.98
Green	0.94 ~ 0.96

Contribution

Jaeyun Song 20190327 (35%)
- Feature Engineering, Derivation of Profit, Experiment

Jaewoo Kim 20180140 (35%)
- EDA, Feature Engineering, Experiment

Seonkyu Lee 20220487 (30%)
- PCA, Feature Engineering, Experiment