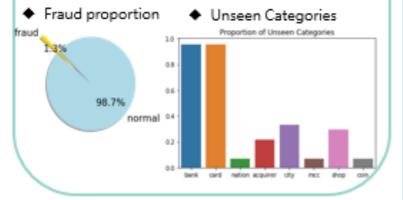# EDA

◆ Data description
- Dummy: Online, Install, Fallback, 3ds, excess
- High cardinality categorical:

Bank, Card, Nation, City, Acquirer, Shop, Mcc , Coin
- Low cardinality categorical:

Trade_type, Trade_cat, Status, Pay_type
- Continuous: Money, Term

◆ Fraud proportion ◆ Unseen Categories

Proportion of Unseen Categories

fraud 1.3%

98.7% normal

Feature Engineering

◆ Preliminary

Feature	Public	Private
High cardinality columns count / day	0.605375	0.603649
Money statistical feature	0.608321	0.608321
Bank and other columns aggregation	0.615974	0.611734
Columns and Columns aggregation	0.626992	0.630726
Money cumulated sum on high cardinality columns	0.640071	0.638095

◆ With adding boost rounds and removing useless features, we improved the model to 0.656894 / 0.655731

◆ Advanced

1. Build three NN models predicting trade_type, money and online columns respectively and embed high cardinality columns to dimension 5.

➡ Improved to 0.664304 / 0.671522

2. Graph embed on high cardinality columns. Construct daily graphs by bank relations and practice random walks. Feed them into the Word2Vec model.

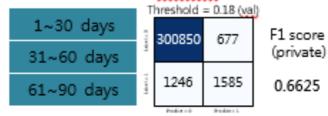➡ Improved to 0.670664 / 0.674588

Model

◆ Cross Validation Scheme

◆ Threshold

Choose the threshold with high average score and low variance

Threshold	Fold 1	Fold 2	Fold 3	Fold4	Fold5
0.001	0.27	0.29	0.34	0.37	0.36
...	...	...	...	...	...
0.999	0.03	0.12	0.18	0.22	0.15

◆ Model

1. 30-days based model (Lightgbm)

Threshold = 0.18 (val)

| 1~30 days |
| 31~60 days |
| 61~90 days |

| 300850 | 677 |
| 1246 | 1585 |

F1 score (private)

0.6625

2. 90-days based model (Lightgbm)

Threshold = 0.26 (val)

| 1~90 days |

| 300741 | 786 |
| 1177 | 1654 |

F1 score (private)

0.6745
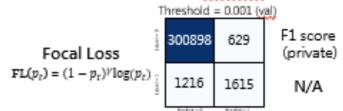
3. 30-days based Focal model (Lightgbm)

Threshold = 0.001 (val)

Focal Loss

$$FL(p_t) = (1 - p_t)^\gamma \log(p_t)$$

| 300898 | 629 |
| 1216 | 1615 |

F1 score (private)

N/A

4. 90-days based model (Neural Network)

Threshold = 0.325 (val)

| 299887 | 1640 |
| 1896 | 935 |

F1 score (private)

0.4464

Ensemble