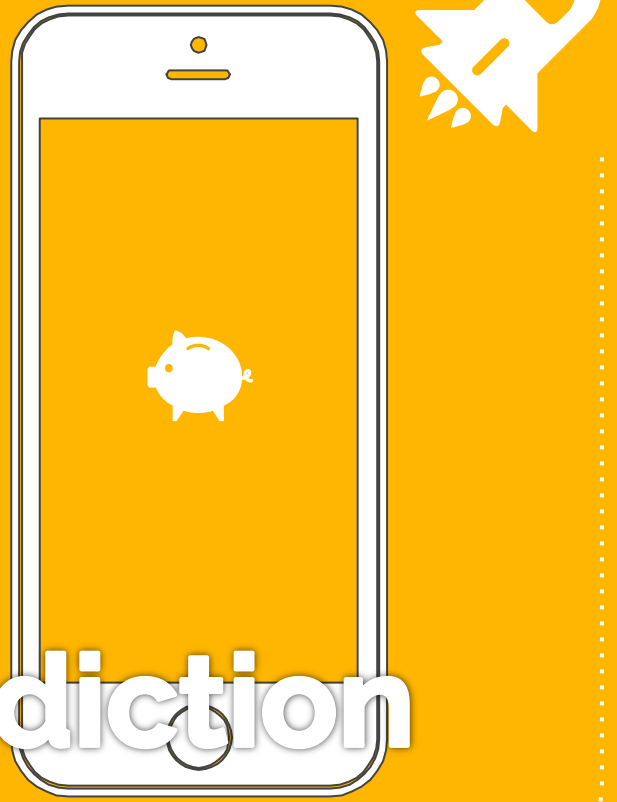
Customer Repurchase Prediction



Guo Jing

Task description

MOVTIVATION
The objective of the vouchers is to motivate these buyers to start to buy things from the platform again and hopefully they will continue the purchases subsequently.

It is important to automatically identify the group of 'inactive' buyers, who are likely to (1) use the "reactivation" vouchers sent to them (2) continue the purchase behavior subsequently.

OBJECTIVE
The main objective of the task is to predict which buyers will use the vouchers and continue to purchase after receiving coupons.

The prediction is made based on a user's purchase behavior generated from the shopping history.

Model construction

1

Data

Data source
Data probing

2

Feature engineering

Three super groups (13 subgroups): user, voucher, date

3

Model construction

Feature extraction
Normalization and missing values
Parameter tuning
Model selection

4

Result & Discussion

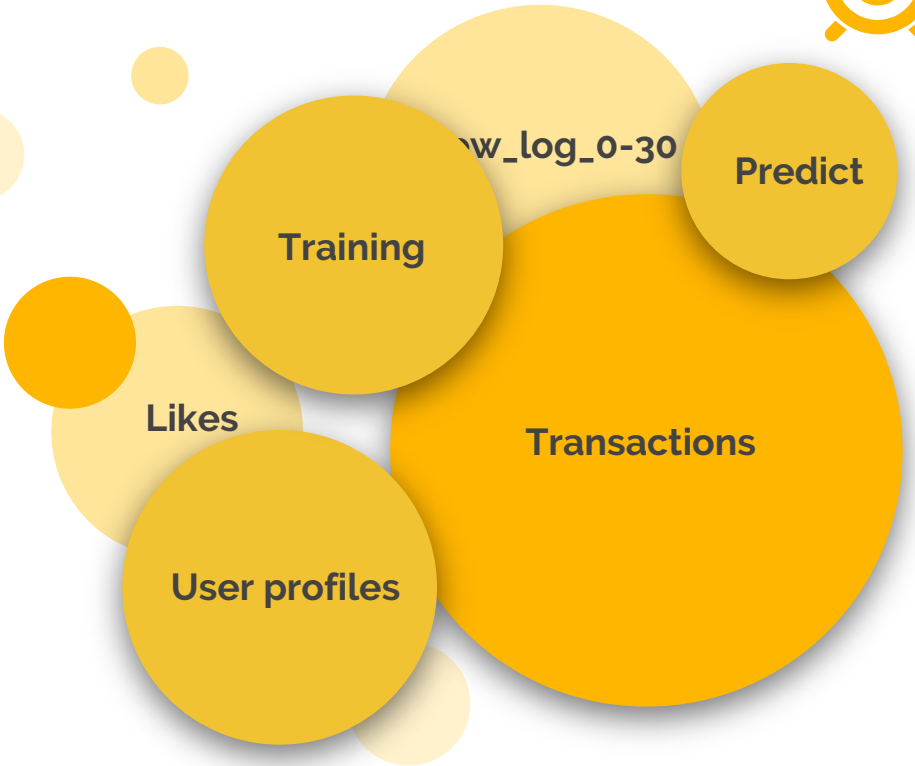
Data balance & cleaning

Data source

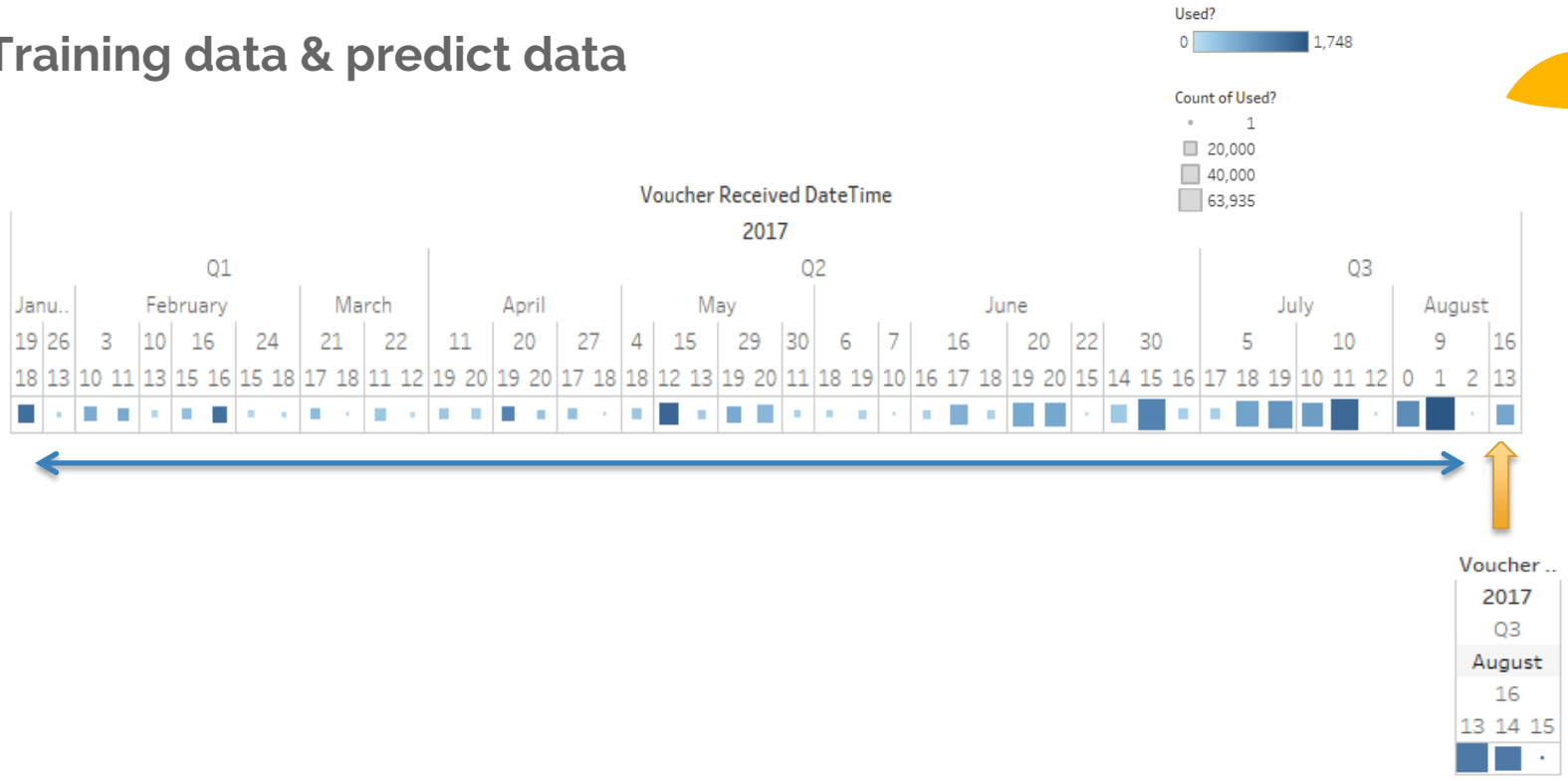
shopping history for a large set of buyers who were targeted for the reactivation campaign.

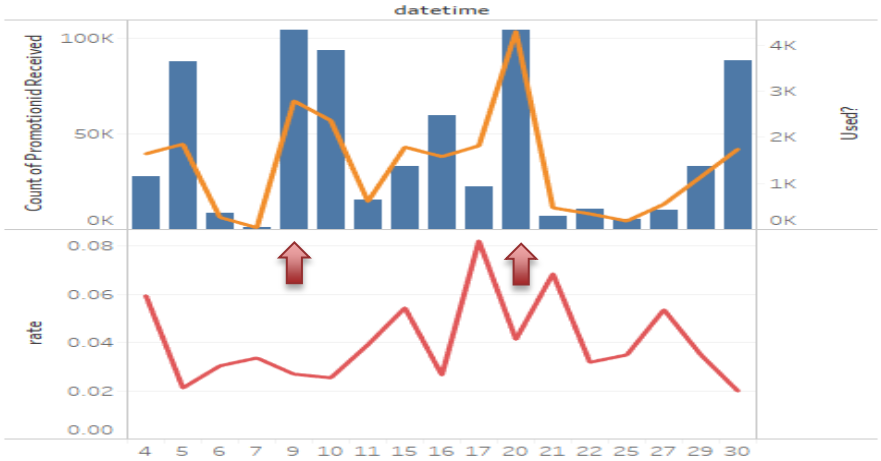
Source	Attributes	Time span
Training (710078,9)	**userid**, **promotionid_recieved** (voucher_code), **voucher_received_time**,used?,repurchased_x (15d,30d,60d,90d)	2017.1-8 (sparse)
Predict (78903,4)	**userid**, **promotionid_recieved** (voucher_code), **voucher_received_time**	2017.8.16
Transactions	Userid,shopid,total_price,order_time,promotionid_used	2015 Q2- 2017 Q3
User profiles	Userid,registration_time,is_seller,gender,...	
Likes	Userid,voucher_received_date,status,ctime	
view_log_0-30	Userid,promotionid_received,date,event_name, count	2017.4-2017.8
Voucher_distribution_active_date	Userid,promotionid_received,time,date,active sessions counts	2017.2(16)-2017.8(16)
Voucher_mechanics	Promotionid,discount,max_value	

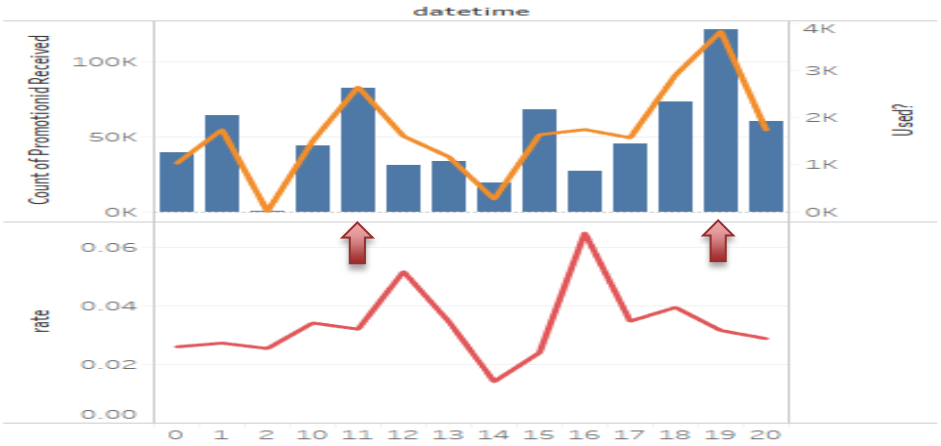
Data probing

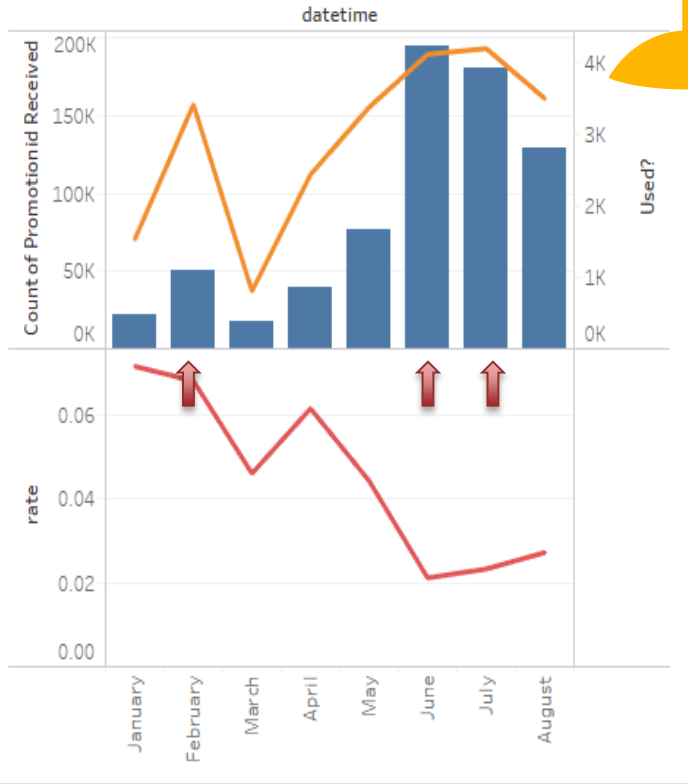


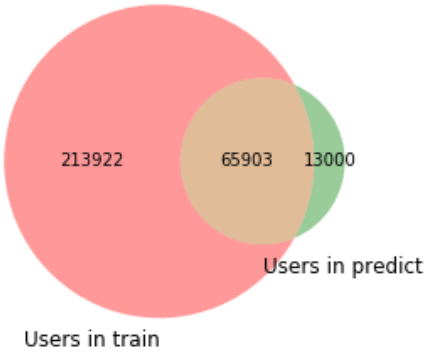
Training data & predict data

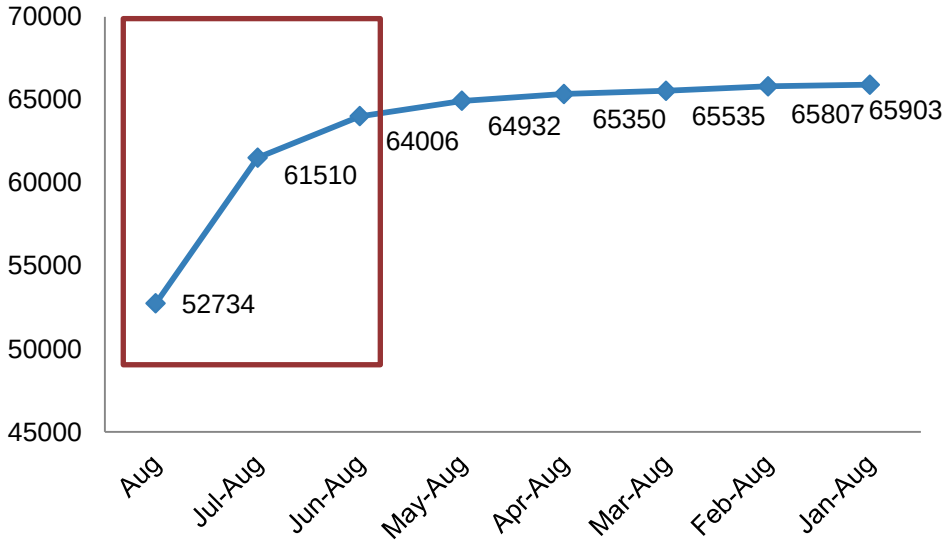


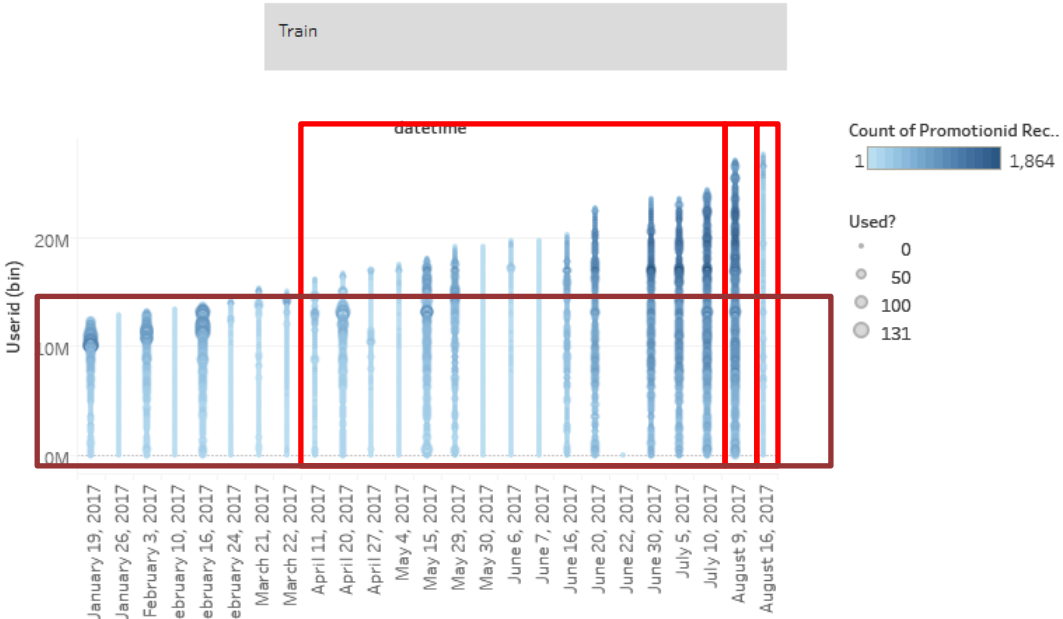

datetime
Count of Promotionid Received
100K
50K
0K
4K
3K
2K
1K
0K
Used?
rate
0.08
0.06
0.04
0.02
0.00
4 5 6 7 9 10 11 15 16 17 20 21 22 25 27 29 30

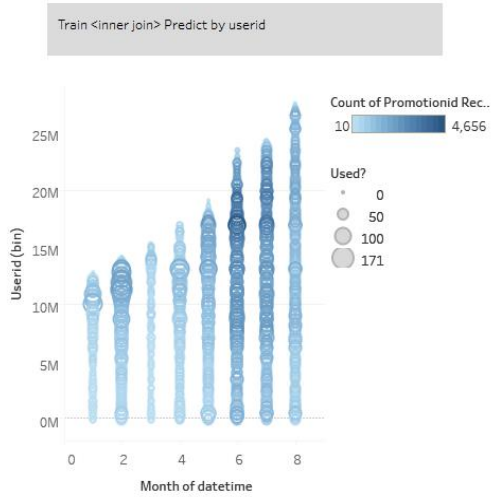

datetime
Count of Promotionid Received
100K
50K
0K
4K
3K
2K
1K
0K
Used?
rate
0.06
0.04
0.02
0.00
0 1 2 10 11 12 13 14 15 16 17 18 19 20


datetime
Count of Promotionid Received
200K
150K
100K
50K
0K
4K
3K
2K
1K
0K
Used?
rate
0.06
0.04
0.02
0.00
January
February
March
April
May
June
July
August


213922
65903
13000
Users in predict
Users in train


70000
65000
60000
55000
50000
45000
52734
61510
64006
64932
65350
65535
65807
65903
Aug
Jul-Aug
Jun-Aug
May-Aug
Apr-Aug
Mar-Aug
Feb-Aug
Jan-Aug

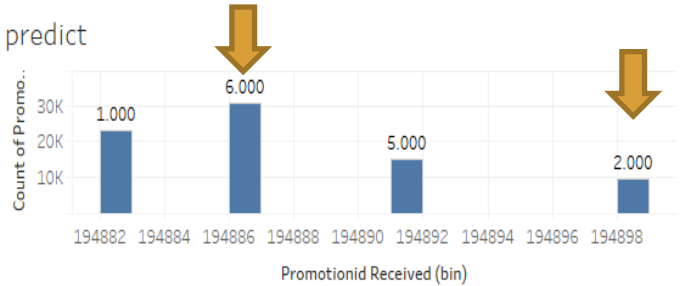

Train
datetime
20M
10M
0M
Userid (bin)
January 19, 2017
January 26, 2017
February 3, 2017
February 10, 2017
February 16, 2017
February 24, 2017
March 21, 2017
March 22, 2017
April 11, 2017
April 20, 2017
April 27, 2017
May 4, 2017
May 15, 2017
May 29, 2017
May 30, 2017
June 6, 2017
June 7, 2017
June 16, 2017
June 20, 2017
June 22, 2017
June 30, 2017
July 5, 2017
July 10, 2017
August 9, 2017
August 16, 2017
Count of Promotionid Rec..
1
1,864
Used?
0
50
100
131

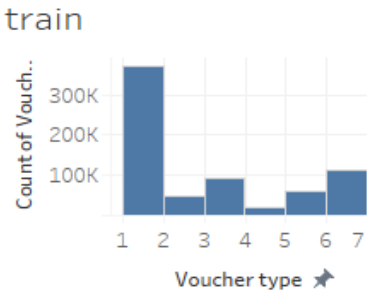

Train <inner join> Predict by userid
25M
20M
15M
10M
5M
0M
Userid (bin)
0
2
4
6
8
Month of datetime
Count of Promotionid Rec..
10
4,656
Used?
0
50
100
171

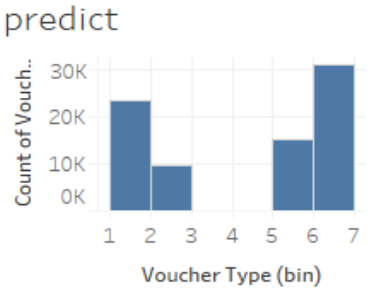
Voucher mechanics

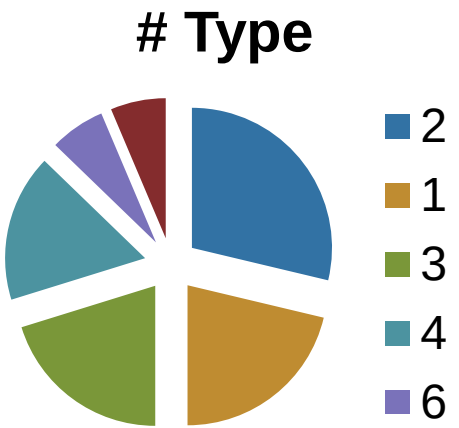
Discount	Max_value	Type
20	1000000	1
50	2000000	2
50	1500000	3
20	1500000	4
50	1000000	5
30	1000000	6

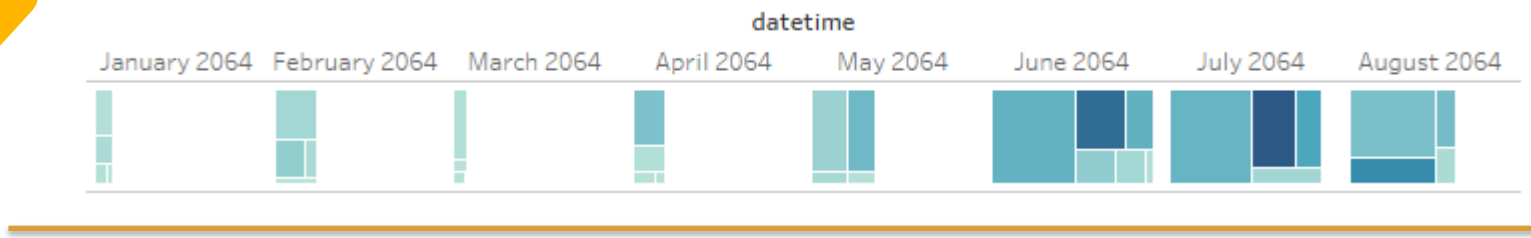
94 promotion IDs (92 in train, 4 in predict)



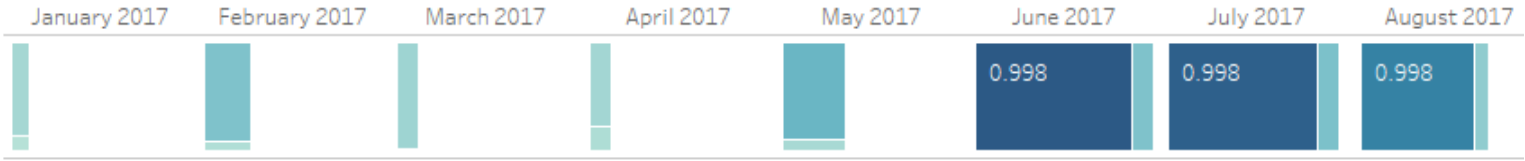


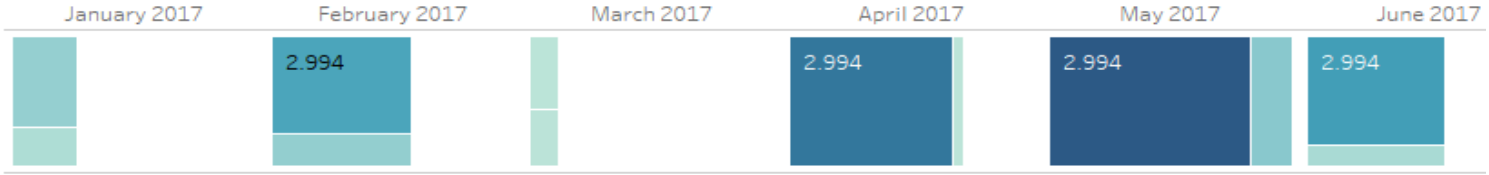






Type 1,2

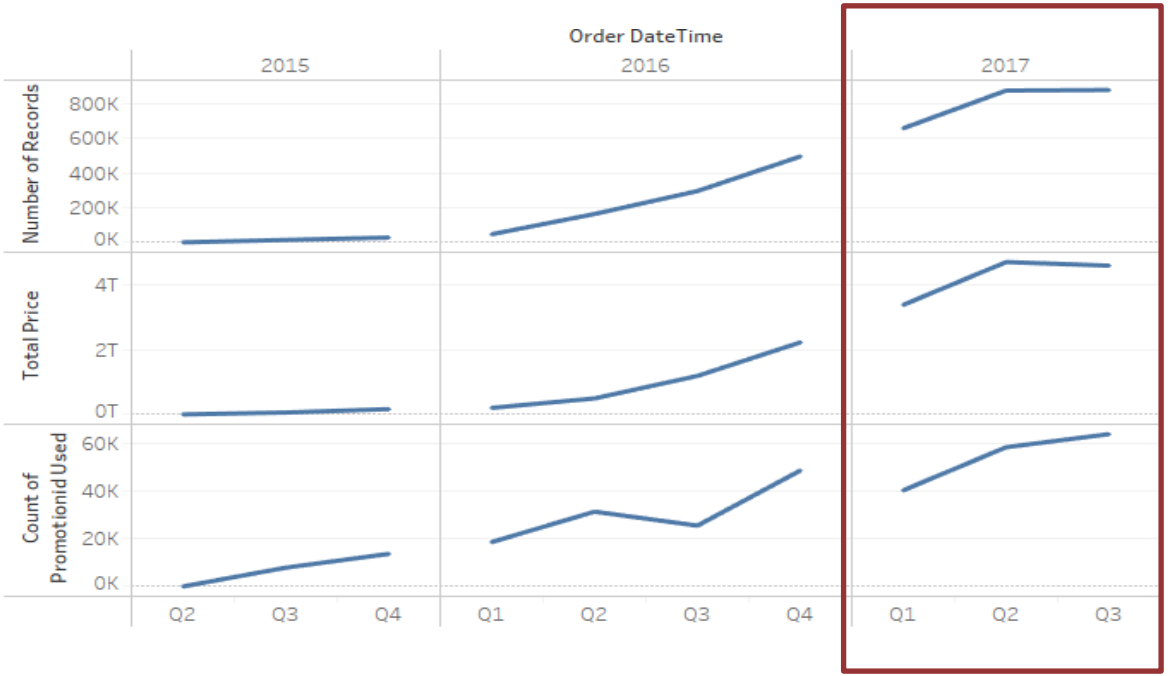




Type 3,4

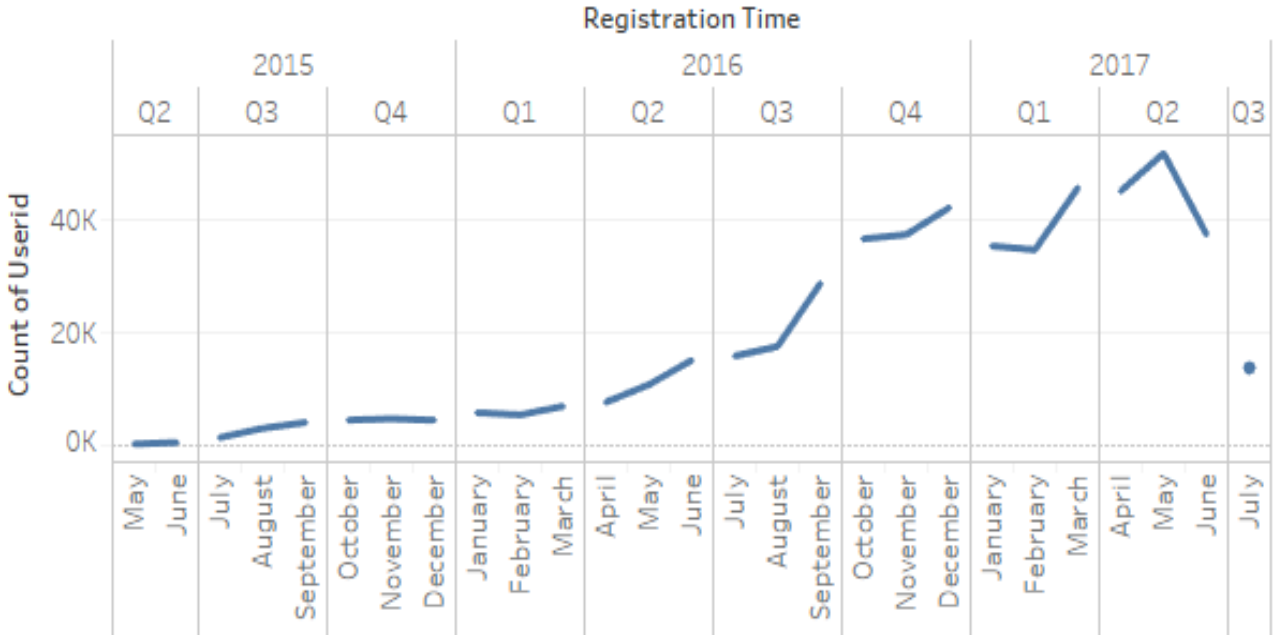
Type 5,6

Transaction history



279803 (out of 279825 users in the **training data**) can be found in the transactions table.
78895 (out of 78903 users in the **predict data**) can be found in the transactions table.

User profiles



View logs

	Voucher Received Date							
	2017							
	Q1			Q2			Q3	
Event Name	January	February	March	April	May	June	July	August
Null	Abc	Abc	Abc	Abc	Abc	Abc	Abc	Abc
addItemToCart				Abc	Abc	Abc	Abc	Abc
trackGenericClick				Abc	Abc	Abc	Abc	Abc
trackGenericScroll				Abc	Abc	Abc	Abc	Abc
trackGenericSearchPageV..				Abc	Abc	Abc	Abc	Abc
trackSearchFilterApplied				Abc	Abc	Abc	Abc	Abc

Active sessions

Voucher Received Date								
2017								
Q1			Q2			Q3		
January	February	March	April	May	June	July	August	
Abc	Abc	Abc	Abc	Abc	Abc	Abc	Abc	

Model construction

1

Data

Data source
Data probing

2

Feature engineering

Three super groups (13 subgroups): user, voucher, date

3

Model construction

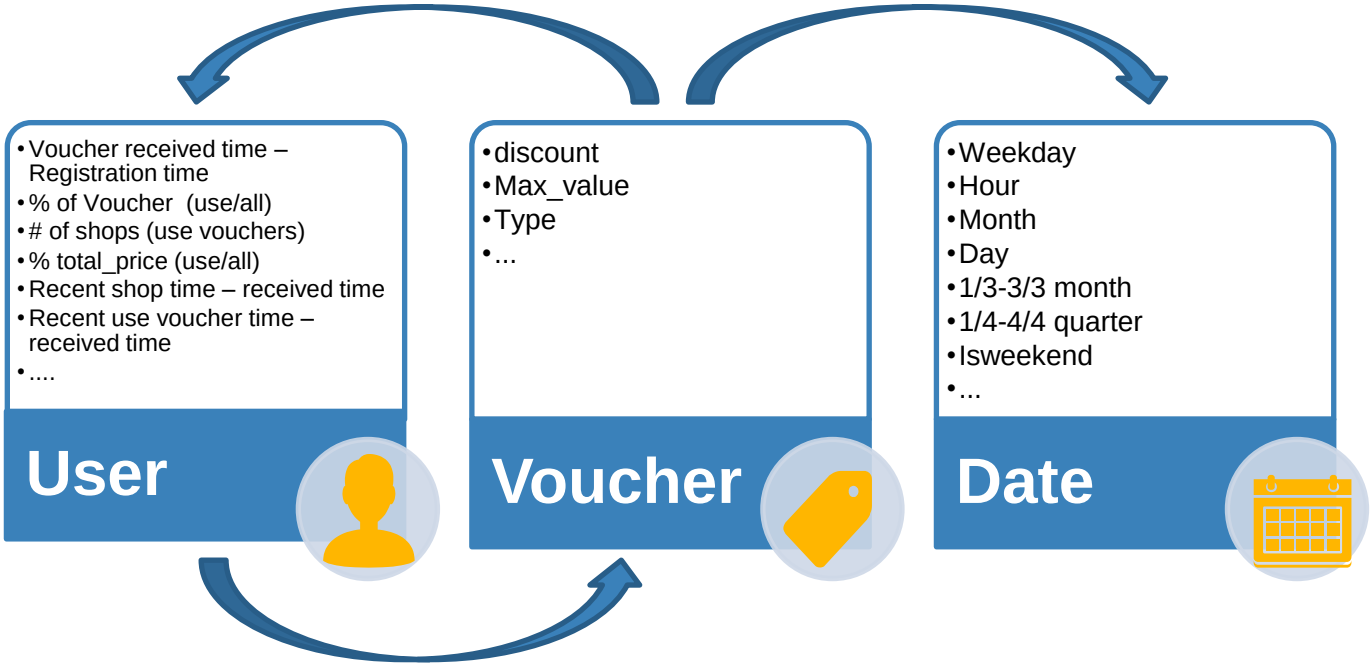
Feature extraction
Normalization and missing values
Parameter tuning
Model selection

4

Result & Discussion

Data balance & cleaning

Three super groups (13 subgroups): user, voucher, date



Source
Training (710078,9)
Predict (78903,4)
Transactions
User profiles
Likes
view_log_0-30
Voucher_distribution_active_date
Voucher_mechanics

Model construction

1

Data

Data source
Data probing

2

Feature generation

Three super groups (13 subgroups): user, voucher, date

3

Model construction

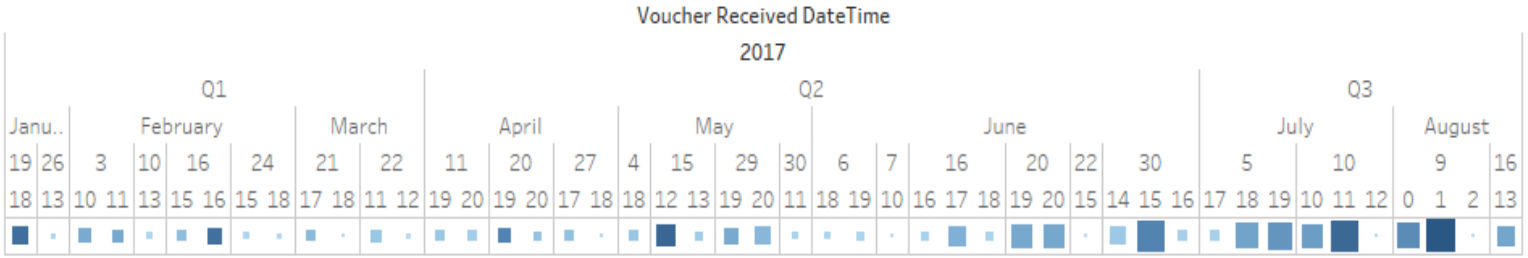
Feature extraction
Normalization and missing values
Parameter tuning
Model selection

4

Result & Discussion

Data balance & cleaning

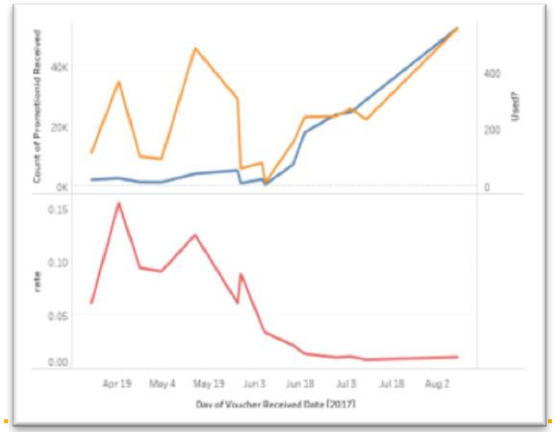
Data splitting



Training

Set	Time span	N/P
Training set 1	4.11-8.9	34:1
Training set 2	4.11-8.9	13:1
Validation set	8.16 (training data)	33:1
Predict set	8.16 (predict data)	?

Z-score ~~(MinMaxScale)~~ is used to normalize continuous values.



Validation

Predict

Model construction

1

Data

Data source
Data probing

2

Feature generation

Three super groups (13 subgroups): user, voucher, date

3

Model construction

Feature extraction
Normalization and missing values
Parameter tuning
Model selection

4

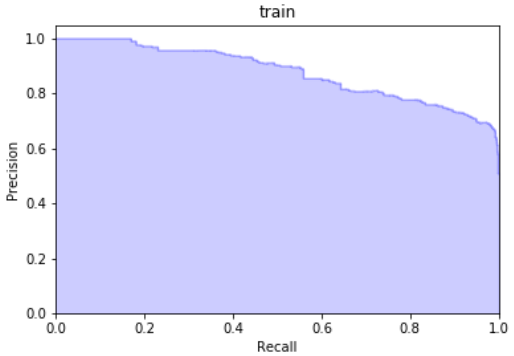
Result & Discussion

Data balance & cleaning

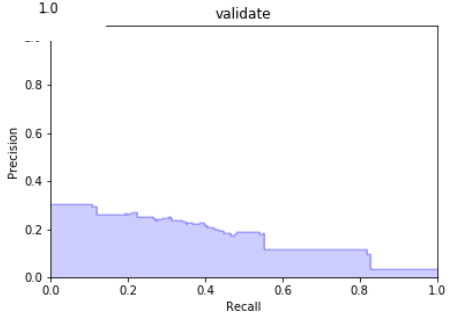
Model selection: Xgboost

```
params={
        'eta':0.3, #0.05-0.3 0.1
        'max_depth':8, #3-10 xxx
        'subsample':0.7, #xxx
        'gamma':50, #xxx
        'colsample_bytree':0.3, #xxx
        'lambda':10,
        'alpha':1,
        'silent':1,
        'verbose_eval':True,
        'max_delta_step': 1,
        'scale_pos_weight': 1,
        'objective': 'binary:logistic',
        'eval_metric': ['map'], #auc
        'seed':12
        }

      'min_child_weight': 10,
```



	0	1	
0	30	3	33
1	0.4	0.6	1



Discussion

Likes table

Data balance & cleaning

- SMOTEENN, CNN

Model selection

- RF, GBDT, LR
- Ensemble methods

Thanks!

GitHub: https://github.com/NetLand-NTU/Shopee