Customers Behavior Analysis of Taobao & TMall

Group Member: Weiteng Li, Shikui Wang, Yuqi Liu

Agenda

I. Introduction and Overview

II. Interesting Findings Through Analysis

III. Returned Customer Analysis

Introduction & Overview

Dataset: traffic data of Taobao & TMall from 05/2018 to 11/2018, found it from Database Lab of Xiamen Universty

Variables:
user_id | item_id | cat_id | merchant_id | brand_id
month | day
action | age_range | gender | province

Tools: Spark

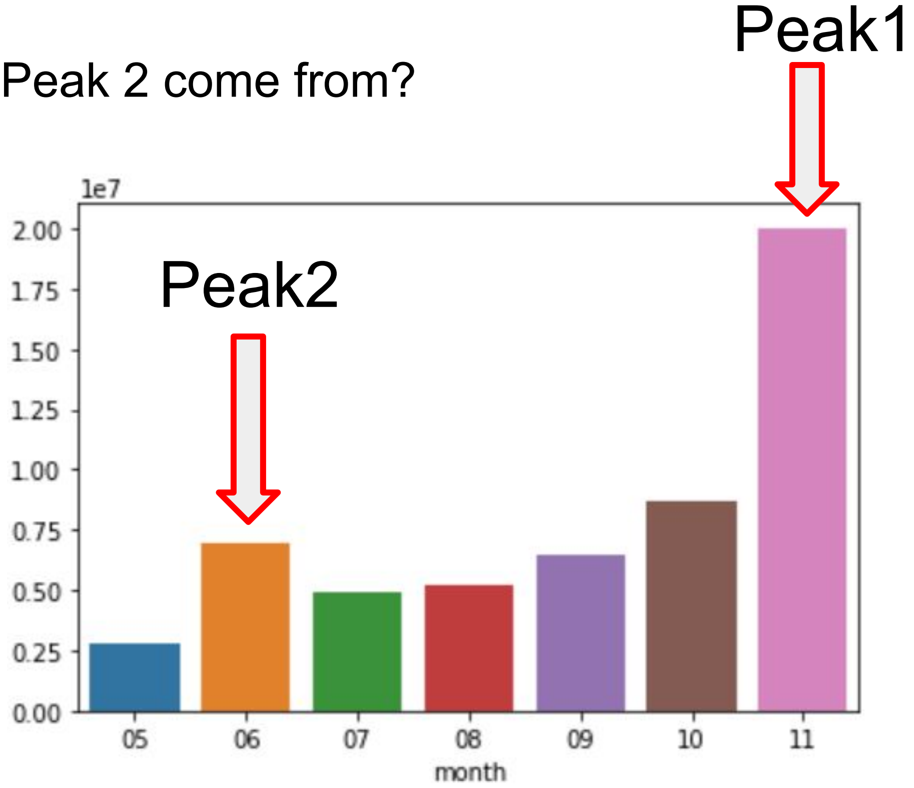
Objective: better understand the e-commerce market &
customer behavior

Finding About Shopping Festival

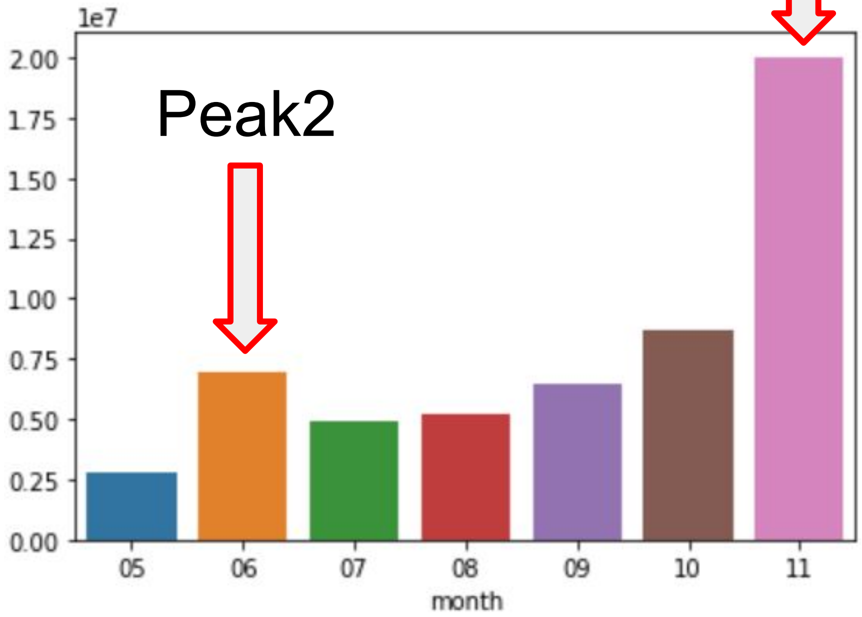
Pattern Through Time

GLOBAL SHOPPING FESTIVAL 2018

Where does the Peak 2 come from?



GLOBAL SHOPPING FESTIVAL 2018

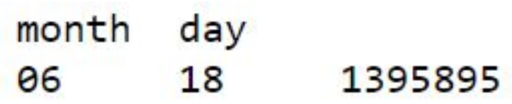


Peak1

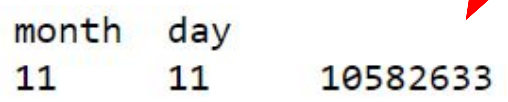
Peak2

618 Festival

10 times larger!



```
month   day
06      18        1395895
```

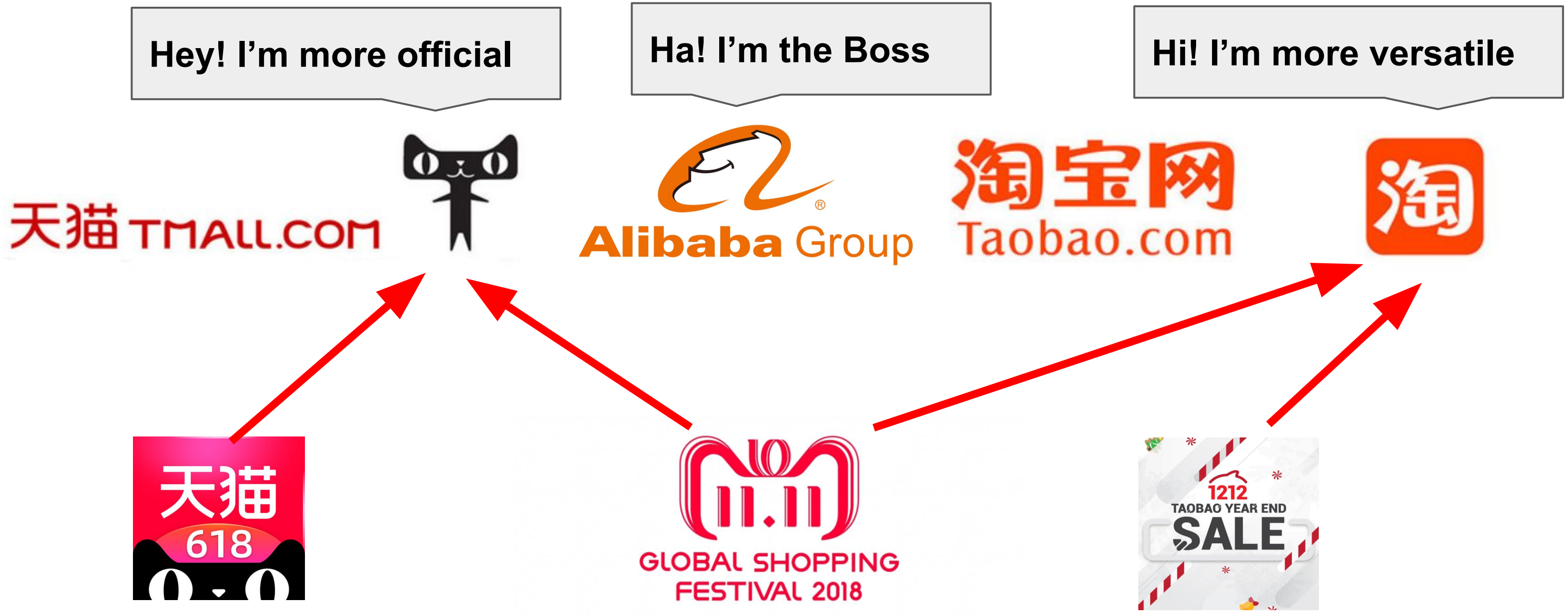


```
month   day
11      11        10582633
```

618 Festival

Why 618 Festival is far less influential than double 11 Festival?

More About Shopping Festival



Hey! I’m more official

Ha! I’m the Boss

Hi! I’m more versatile

6.18
TMALL Only
Special Category Focus

11.11

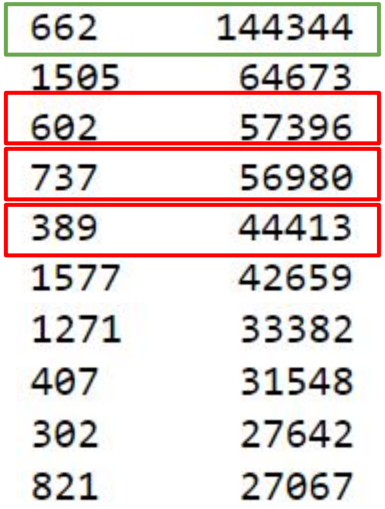
12.12

Maybe we can find some hint from the data to verify the difference.

Follow Up Validation

6.18

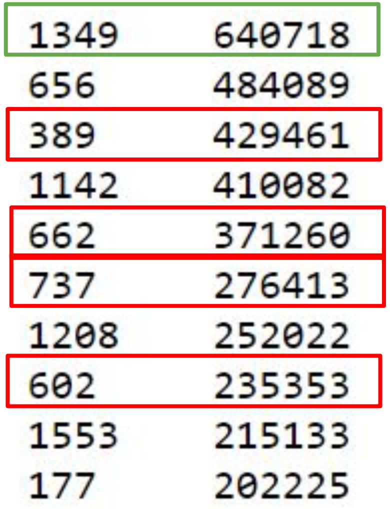
Blowing Up !



Cat.ID	Count
662	144344
1505	64673
602	57396
737	56980
389	44413
1577	42659
1271	33382
407	31548
302	27642
821	27067

Overlap

11.11



Cat.ID	Count
1349	640718
656	484089
389	429461
1142	410082
662	371260
737	276413
1208	252022
602	235353
1553	215133
177	202225

Returned Customer Analysis

Data Manipulation & Variables

How to define repeat customer?

1. User_ID + Merchant_ID
2. Different purchased date

If one user purchased in three different shops --> Three unique observations
If one user purchased in the same shop for more than one time --> Returned customer

Variables Used:

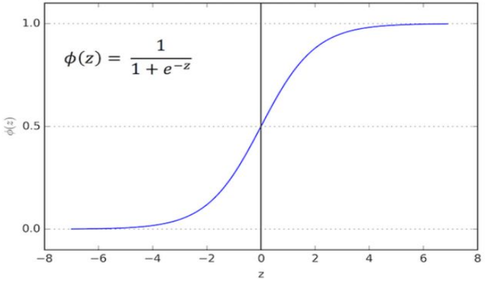
Age_Group:

1 [0 ,18]
2 [18,24],
3 [25,29],
4 [30,34],
5 [35,39],
6 [40,49]

Gender: Male / Female

Model & Result Analysis

· **Logistics Regression:**



· **Result Formular:**
y = sigmoid (-3.19 + 0.302×Age_Range + 0.044×Is_female)

· **Accuracy** = 96.27% (based on 15% of test set)
· **AUC** = 0.5413 (slightly better than random guess)
· **Why → Super imbalanced data**
(3.71% Positive observations)

What can we do next?

1. Adding more Features: Item catrgory, Merchant, Brand, City etc. (ha
encoding becuaseof high cardinality)

2. Features engineering - dates variables, extract extra features (is holid

3. Try more models (SVM, RF, GBT) / Ensemble

4. Data augmentation - upsampling / downsampling

5. Clustering customers and build model to identify future customer seg