**­Repeat Buyers Prediction-Challenge the Baseline**

**Problem Definition**

Merchants sometimes run big promotions (e.g., discounts or cash coupons) on particular dates (e.g., Boxing-day Sales, "Black Friday" or "Double 11 (Nov 11th)”, in order to attract a large number of new buyers. Unfortunately, many of the attracted buyers are one-time deal hunters, and these promotions may have little long lasting impact on sales. To alleviate this problem, it is important for merchants to identify who can be converted into repeated buyers. By targeting on these potential loyal customers, merchants can greatly reduce the promotion cost and enhance the return on investment (ROI). It is well known that in the field of online advertising, customer targeting is extremely challenging, especially for fresh buyers. However, with the long-term user behavior log accumulated by Tmall.com, we may be able to solve this problem.  
  
We provide a set of merchants and their corresponding new buyers acquired during the promotion on the "Double 11" day. Your task is to predict which new buyers for given merchants will become loyal customers in the future. In other words, you need to predict the probability that these new buyers would purchase items from the same merchants again within 6 months.

**Data Description**

The data set contains anonymized users' shopping logs in the past 6 months before and on the "Double 11" day,and the label information indicating whether they are repeated buyers. Due to privacy issue, data is sampled in a biased way, so the statistical result on this data set would deviate from the actual of Tmall.com. But it will not affect the applicability of the solution. The files for the training and testing data sets can be found in "data\_format2.zip".Details of the data format can be found in the table below.

|  |  |
| --- | --- |
| **Data Fields** | **Definition** |
| user\_id | A unique id for the shopper. |
| age\_range | User' s age range:   * 1 for <18; * 2 for [18,24]; * 3 for [25,29]; * 4 for [30,34]; * 5 for [35,39]; * 6 for [40,49]; * 7 and 8 for >= 50; * 0 and NULL for unknown. |
| gender | User' s gender:   * 0 for female, * 1 for male, * 2 and NULL for unknown. |
| merchant\_id | A unique id for the merchant. |
| label | Value from {0, 1, -1, NULL}.   * ' 1' denotes ' user\_id' is a repeat buyer for ' merchant\_id' , * while ' 0' is the opposite. * ' -1' represents that ' user\_id' is not a new customer of the given merchant, thus out of our prediction.   However, such records may provide additional information. ' NULL' occurs only in the testing data, indicating it is a pair to predict. |
| activity\_log | Set of interaction records between {user\_id, merchant\_id}, where each record is an action represented as '   * item\_id : category\_id : brand\_id : time\_stamp : action\_type' . * ' #' is used to separate two neighbouring elements. * Records are not sorted in any particular order. |

**Your Submission should be named as "prediction.csv" with following format.**

|  |  |
| --- | --- |
| **Data Fields** | **Definition** |
| user\_id | A unique id for the shopper. |
| merchant\_id | A unique id for the merchant. |
| prob | Predicted probability of the given user becoming a repeat buyer of the given merchant. Value should be between 0 and 1. |

**Data in another format (format1)**

We also provide the same data set in another format, which contains 4 files and may be more user-friendly for feature engineering (files can be found in "data\_format1.zip"). The details of the data formats can be found below:

* Green words’column index mean: if block is NaN, then put -1 in.
* Red words’column index mean: if block is NaN, then drop row.
* Yellow back ground mean tricky data type need to notice.

**User Behaviour Logs**

|  |  |
| --- | --- |
| **Data Fields** | **Definition** |
| user\_id | A unique id for the shopper. |
| item\_id | A unique id for the item. |
| cat\_id | A unique id for the category that the item belongs to. |
| merchant\_id | A unique id for the merchant. |
| brand\_id | A unique id for the brand of the item. |
| time\_tamp | Date the action took place (format: mmdd) |
| action\_type | It is an enumerated type {0, 1, 2, 3}, where   * 0 is for click, * 1 is for add-to-cart, * 2 is for purchase and * 3 is for add-to-favourite. |

**User Profile**

|  |  |
| --- | --- |
| **Data Fields** | **Definition** |
| user\_id | A unique id for the shopper. |
| age\_range | User' s age range:   * 1 for <18; * 2 for [18,24]; * 3 for [25,29]; * 4 for [30,34]; * 5 for [35,39]; * 6 for [40,49]; * 7 and 8 for >= 50; * 0 and NULL for unknown. |
| gender | User' s gender:   * 0 for female, * 1 for male, * 2 and NULL for unknown. |

**Training and Testing Data**

|  |  |
| --- | --- |
| **Data Fields** | **Definition** |
| user\_id | A unique id for the shopper. |
| merchant\_id | A unique id for the merchant. |
| label | It is an enumerated type {0, 1}, where   * 1 means repeat buyer, * 0 is for non-repeat buyer.   This field is empty for test data. |

**Data source:**

[**https://tianchi.aliyun.com/getStart/information.htm?spm=5176.100067.5678.2.422d1aecSDDyms&raceId=231576**](https://tianchi.aliyun.com/getStart/information.htm?spm=5176.100067.5678.2.422d1aecSDDyms&raceId=231576)

**time cost of import data into MySQL**

|  |  |  |
| --- | --- | --- |
| **table** | **seconds** |  |
| user\_info\_format1\_2 | 54.18362617492676 |  |
| train\_format1\_2 | 26.306838989257812 |  |
| user\_log\_format1\_2 | 2060.9664976596832 |  |

**Report of Basic Information**

**Data type after cleaning**

|  |  |  |  |
| --- | --- | --- | --- |
| **File Name** | train\_format1-2.csv | user\_info\_format1.csv | user\_log\_format1-2.csv |
| **Length** | 260864 | 424170 | 54925330 |
| **Cleaning time** | 0.131 | 0.227 | 33.096 |
| **To csv time** | 0.822 | 1.210 | 344.959 |
| **Column** | * 'user\_id', * 'merchant\_id', * 'label' | * 'user\_id', * 'age\_range', * 'gender' | * 'user\_id', * 'item\_id', * 'cat\_id', * 'seller\_id', * 'brand\_id', * 'time\_stamp', * 'action\_type' |
| **Cleaning Rule** | * If there has any item in row is Null, then drop whole row. * If whole column is duplicate, then drop it. * “merchant\_id” should be change to the same as 'seller\_id' from user\_log\_format1-2.csv | * If “user\_id” is null, then drop whole column. * If “user\_id” is duplicate, then drop whole column. * If “gender”is Null, then fill if with ‘2’. * If “age\_range”is Null, then fill if with ‘0’. * If “age\_range”is ‘8’, then replace it with ‘7’. * If “gender”is not 0,1,2, then replace if with ‘2’. | * If “user\_id”or “item\_id” is null, then drop whole column. * If column beside “user\_id”or “item\_id” is null, then fill it with “-1”. * If whole column is duplicate, then drop it. * Time should separate in two columns (%100) |

* Read file cost time (total): 63.93757772445679 (with pandas.read\_csv)
* Red word is suggested way to deal with data.

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Unique Numbers** | **About** |
| user\_id | 424170 |  |
| merchant\_id | 4995 |  |
| item\_id | 1090390 |  |
| cat\_id | 1658 |  |
| brand\_id | 8444 |  |

Age info

from: user\_info\_format1-2.csv

median: 3.0

mode: 0 3.0

quantile: [2.0, 0.0]

name age\_range percentage accumulate accumulate\_percentage

0.0 unknown 95131 0.224276 95131 0.224276

1.0 <18 24 0.000057 95155 0.224332

2.0 18-24 52871 0.124646 148026 0.348978

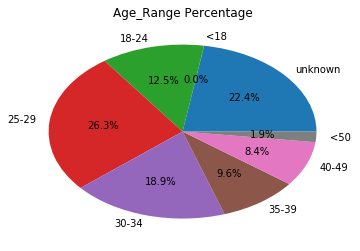
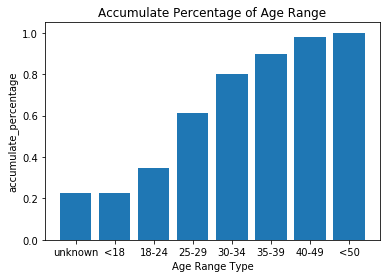
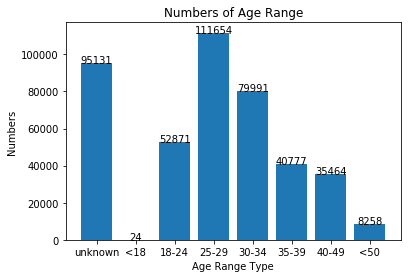
3.0 25-29 111654 0.263229 259680 0.612207

4.0 30-34 79991 0.188582 339671 0.800790

5.0 35-39 40777 0.096134 380448 0.896923

6.0 40-49 35464 0.083608 415912 0.980531

7.0 <50 8258 0.019469 424170 1.000000



## Gender info

from: train\_format1-2.csv

median: 0.0

mode: 0 0.0

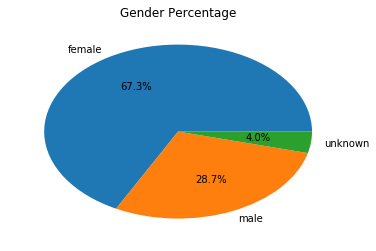
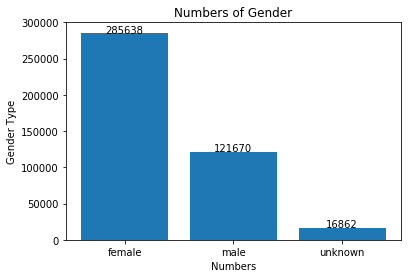
quantile: [0.0, 0.0]

name gender percentage accumulate accumulate\_percentage

0.0 female 285638 0.673405 285638 0.673405

1.0 male 121670 0.286843 407308 0.960247

2.0 unknown 16862 0.039753 424170 1.000000



## Label info[¶](http://localhost:8888/notebooks/Desktop/Tmall/DescriptiveStatistic/tmall_pilot_study_v1.3.ipynb#Label-info)

from train\_format1-2.csv

median: 0.0

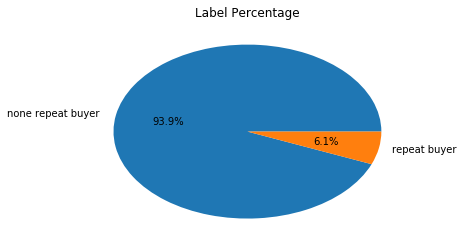
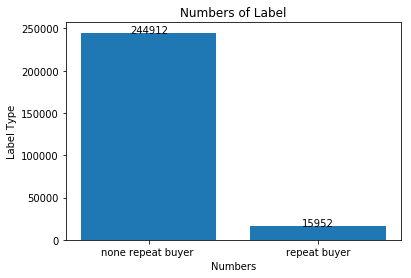
mode: 0 0.0

quantile: [0.0, 0.0]

name label percentage accumulate accumulate\_percentage

0.0 none repeat buyer 244912 0.938849 244912 0.938849

1.0 repeat buyer 15952 0.061151 260864 1.000000



## Gender and Age Range

from: user\_info\_format1-2.csv

Frequency amoung Gender and Age Range

---------------------------------------------------

gender 0.0 1.0 2.0

age\_range

0.0 66813 22655 5663

1.0 18 6 0

2.0 33692 18309 870

3.0 74609 33217 3828

4.0 52032 23708 4251

5.0 27823 11462 1492

6.0 25160 9745 559

7.0 5491 2568 199

Frequency amoung Gender and Age Range in persentage

---------------------------------------------------

gender 0.0 1.0 2.0

age\_range

0.0 15.751468 5.341019 1.335078

1.0 0.004244 0.001415 0.000000

2.0 7.943042 4.316430 0.205106

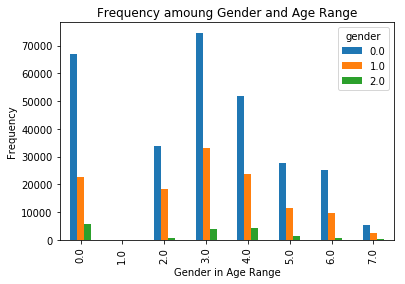
3.0 17.589410 7.831058 0.902468

4.0 12.266780 5.589268 1.002193

5.0 6.559398 2.702218 0.351746

6.0 5.931584 2.297428 0.131787

7.0 1.294528 0.605418 0.046915



## Actions to Gender

from: user\_info\_format1-2.csv & user\_log\_format1-2.csv

Frequency amoung Actions and Gender

---------------------------------------------------

gender 0.0 1.0 2.0

action\_type

0 252228 107568 14860

1 198 88 10

2 19744 8134 1131

3 13468 5880 861

Frequency amoung Actions and Gender in persentage

---------------------------------------------------

gender 0.0 1.0 2.0

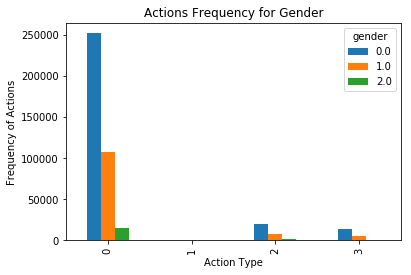
action\_type

0 59.463894 25.359644 3.503312

1 0.046679 0.020746 0.002358

2 4.654737 1.917627 0.266638

3 3.175142 1.386237 0.202985

****

## Actions to Age

from: user\_info\_format1-2.csv & user\_log\_format1-2.csv

Frequency amoung Actions and Age

---------------------------------------------------

age\_range 0.0 1.0 2.0 3.0 4.0 5.0 6.0 7.0

action\_type

0 84027 21 46785 98510 70633 36051 31344 7285

1 57 0 38 75 60 33 27 6

2 6583 3 3610 7709 5386 2711 2434 573

3 4464 0 2438 5360 3912 1982 1659 394

Frequency amoung Actions and Age in persentage

---------------------------------------------------

age\_range 0.0 1.0 2.0 3.0 4.0 5.0 \

action\_type

0 19.809746 0.004951 11.029776 23.224179 16.652050 8.499187

1 0.013438 0.000000 0.008959 0.017682 0.014145 0.007780

2 1.551972 0.000707 0.851074 1.817432 1.269774 0.639131

3 1.052408 0.000000 0.574770 1.263644 0.922272 0.467265

age\_range 6.0 7.0

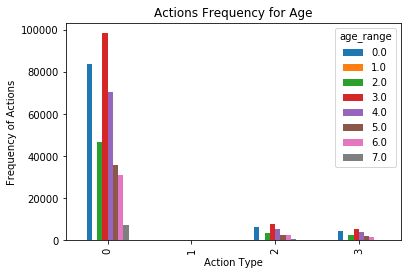
action\_type

0 7.389490 1.717472

1 0.006365 0.001415

2 0.573827 0.135087

3 0.391117 0.092887

****

## Action to Age and Gender

from: user\_info\_format1-2.csv & user\_log\_format1-2.csv

action\_type 0 1 2 3

age\_range gender

0.0 0.0 59017 44 4667 3085

1.0 20028 12 1526 1089

2.0 4982 1 390 290

1.0 0.0 16 0 2 0

1.0 5 0 1 0

2.0 0.0 29794 24 2340 1534

1.0 16232 14 1202 861

2.0 759 0 68 43

3.0 0.0 65740 44 5233 3592

1.0 29383 27 2233 1574

2.0 3387 4 243 194

4.0 0.0 45991 40 3502 2499

1.0 20899 17 1597 1195

2.0 3743 3 287 218

5.0 0.0 24583 21 1865 1354

1.0 10149 10 750 553

2.0 1319 2 96 75

6.0 0.0 22219 20 1764 1157

1.0 8628 7 633 477

2.0 497 0 37 25

7.0 0.0 4868 5 371 247

1.0 2244 1 192 131

2.0 173 0 10 16

**Some idea of new column**

**What we have:**

* The column what we can count it as a new numeric type is ‘act\_log’.
* And also, we have ‘time\_stamp’ to be another numeric data. Although it is a continue variable, it is hard to deal with different action type.
* In our class, [Cluster Analysis](https://www.yongxi-stat.com/cluster-analysis/) may be a good way to analysis for fist. To find out some relation between data.

**Hard to do:**

* Time series is too short to build a model to predict sell numbers of items (different season has different event).

**What can do:**

* Table1:
  + To grouping item then give it a new tag.
  + The idea is to separate the different item’s purchase type

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | User\_id | item\_id | Act0\_times | Act1\_times | Act2\_times | Act3\_times |
| Example | 328862 | 323294 | 4 | 1 | 34 | 2 |
| Data Type | class | class | integer | integer | integer | integer |

* Table2:
  + Find out particular user’s action frequence to items.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | User\_id | Time\_stamp | Act0 | Act1 | Act2 | Act3 |
| Example | 328862 | 829 | 1 | 0 | 0 | 0 |
| Data Type | class | date | boolean | boolean | boolean | Boolean |

* Table3:
  + Find out most popular item by time.
  + Find out which item is high sell frequence in particular term.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | Time\_stamp | Item\_id | Act0 | Act1 | Act2 | Act3 |
| Example | 829 | 323294 | 107 | 598 | 21 | 44 |
| Data Type | Date | class | integer | integer | integer | integer |

* Table4:
  + Find out most popular category.
  + Find out which cat is high sell frequence in particular term.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | Time\_stamp | cat\_id | Act0 | Act1 | Act2 | Act3 |
| Example | 829 |  |  |  |  |  |
| Data Type | Date | class | integer | integer | integer | integer |

* Table5:
  + Find out which categories are users’ prefer.
  + The action happen in same date count as once.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | User\_id | cat\_id | Act0 | Act1 | Act2 | Act3 |
| Example |  |  |  |  |  |  |
| Data Type | Class | Class | Integer | integer | integer | integer |

* Table6:
  + Find out which merchants are users’ prefer.
  + [Cluster Analysis](https://www.yongxi-stat.com/cluster-analysis/)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | User\_id | mer\_id | Act0 | Act1 | Act2 | Act3 |
| Example |  |  |  |  |  |  |
| Data Type | Class | class | integer | integer | integer | integer |

* Table7:
  + Find out which term has more purchase and its relation with age or gender.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Name | Time\_stmp | Gender | Act0 | Act1 | Act2 | Act3 |
| Example |  |  |  |  |  |  |
| Data Type | date | Class | integer | integer | integer | integer |