**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)