## 电商用户复购预测

## 1.需求分析

大家对电商都应该非常热悉了，在“双十一”“双十二”或者黑色星期五的疯狂购买中肯定少不了它。在2019年，其“双十一”销量额更是突破了2684亿元

我们先来看看电商用户重复购买的研究背景：商家一般会在“双十一”、“双十二”、黑色星期五进行大规模的促销，比如采用折扣券或者现金券的形式让利给消费者，同时这也是吸引新用户的重要形式。然而，被低价、折扣、各种让利吸引的用户往往在这次消费之后就再也没有购买，其主要是为了“薄羊毛”，成为一次性用户，针对这些用户的促销并没有给店铺带来未来销量的增加，只是增加了相应的营销成本。因此，店铺有迫切的需求，想知道哪些用户可能会成为重复购买其店铺商品的忠诚用户，以便对这些有潜力的忠诚用户进行精准营销，这样就可以大幅降低其促销成本，提高投资报率(ROI)。在这样的背景下，我们研究项目的预测目的就是用户购买商品的复购率。

在了解了需求之后，我们又该怎么寻找这些对店铺忠诚的用户呢？什么样的用户被认为是忠诚用户呢？你买东西的时候如果经常去某家店，店长一般会这样和你说：“都是老顾客了，给你最低价”。也就是说，经常到某家店铺购买的用户就是该店铺的忠实用户。这也便引出了店铺重复购买的概念，用更准确的数学形式对其定义就是：一个用户在一家店铺购买次数大于1。某个用户在一家店铺买过商品之后，觉得其很靠谱，会再次在这家店铺购买相同或者其他商品的行为叫作重复购买行为。

现在我们就可以使用店铺复购率来衡量店铺的用户忠诚程度。复购率有两种计算方法：

复购率=重复购买用户数量/用户样本数量

复购率=重复购买行为次数（或者交易次数）/用户样本数量

下面举例来帮助大家理解复购率的概念：假设有100个用户，其中有20个有复购行为，其中10个复购1次（总购买次数=2），10个复购2次（总购买次数3）。用两种计算方法分别计算复购率：

计算方法1：复购率=重复购买用户数量/用户样本数量，如表2-1-1所示。

表2-1-1

|  |  |  |
| --- | --- | --- |
| 用户样本数量 | 重复购买人数 | 复购率=（10\*1+10\*1）/100 |
| 100 | 20 | 20% |

计算方法2：复购率=重复购买行为次数（或者交易次数）/用户样本数量，如表2-1-2所示。

表2-1-2

|  |  |  |
| --- | --- | --- |
| 用户样本数量 | 重复购买人数 | 复购率=（10\*1+10\*2）/100 |
| 100 | 20 | 30% |

那么，哪些因素会影响用尸的复购率呢？可以通过两个方面来考虑：一个是用户维度,另一个是商品维度。不同用户的性别、身高、体重、月收入、喜好、网购习惯等都可能会影响用户对店铺的复购行为。比如，如果某个女性客户有明显的品牌偏好，那么在其喜爱的品牌下购买的次数会更多，金额也会更高；同样而言，男性用户更容易买数码产品、鞋子，而女性用户对服饰、包包购买的频率会更高，这就是不同性别之间的差异。与此同时，像手机这种类目的复购率往往会偏低，母婴类的复购率却很高。这些都是业务类别、客户自身和商品属性带来的影响。因此，针对不同的店铺准确识别其未来会复购的用户并不简单，而且每个类和客户都有其自身的内在特点。下面让我们使用现有的数据和算法来尝试解决这个问题。

## 2.数据集介绍

数据描述如下：数据集主要包括在“双十一”之前和之后的6个月内，匿名用户的购买为日志数据、用户的画像数据，以及相关的指示标签，标明其是否为重复购买用户。

注明，由于隐私问题，教据以偏移量进行采样，因此该数据集的统计结果和电商的实际数据会存在一定的偏差，但不会影响解决方案的适用性。

下面继续了解数据的详细信息，主要有用户行为日志表、用户特征表以及训练数据和试数据表三张表。

用户行为日志表的字段含义，如表2-1-3所示：

|  |  |
| --- | --- |
| Data Fields | Definition |
| user\_id | 顾客（用户）ID |
| item\_id | 商品ID |
| cat\_id | 商品类目ID |
| seller\_id | 商家（店铺）ID |
| brand\_id | 品牌ID |
| time\_tamp | 行为发生时间 |
| action\_type | 取值范围：{0，1，2，3)，0表示点击，1表示加入购物车，2表示购买，3表示收藏 |

可以看出，表中包括用户在天猫的各种用户行为的记录数据。

用户特征表的字段含义，如表2-1-4所示：

|  |  |
| --- | --- |
| Data Fields | Definition |
| user\_id | 顾客（用户）ID |
| age\_range | 顾客（用户）年龄范围：1表示<18；2表示[18，24]；3表示[25，29]；4表[30，34]；5表示[35，39]；6表示[40，49]；  7 and 8表示>=50；0 and Null表示未知 |
| gender | 顾客性别：0表示女性；1表示男性；  2 and Null 表示未知 |

可以看出，表中呈现的是用户的画像信息。

训练数据和测试数据表的字段含义，如表2-1-5所示：

表2-1-5

|  |  |
| --- | --- |
| Data Fields | Definition |
| user\_id | 顾客（用户）ID |
| merchsnt\_id | 商家（店铺）ID |
| label | 取值范围：{0，1}，1表示重复购买，0表示非重复购买，测试集数据的这个字段为Null |

合并后的训练数据集：

|  |  |
| --- | --- |
| 字段 | 备注 |
| user\_id | 顾客（用户）ID |
| merchsnt\_id | 商家（店铺）ID |
| age\_range | 年龄范围 |
| gender | 性别 |
| user\_cnt | 用户购买总次数 |
| seller\_nunique | 不同店铺个数 |
| cat\_nunique | 购物车不同品类个数 |
| brand\_nunique | 不同品牌个数 |
| item\_nunique | 不同商品个数 |
| label | 标签 |

关键代码：

* 导包

import pandas as pd

import numpy as np

import warnings

warnings.filterwarnings("ignore")

* 加载文件

train\_data=pd.read\_csv("./data\_format1/train\_all01.csv",nrows=1500)

test\_data=pd.read\_csv("./data\_format1/test\_all.csv",nrows=500)

* 查看数据信息

train\_data.head()

test\_data.head()

* 特征值、目标值

features\_columns=[col for col in train\_data.columns if col not in ['user\_id','label']]

train=train\_data[features\_columns].values

test=test\_data[features\_columns].values

target=train\_data['label'].values

* 线性回归模型

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error,r2\_score

import numpy as np

X\_train,X\_test,y\_train,y\_test = train\_test\_split(train,target,test\_size=0.2,random\_state=0)

lin\_reg=LinearRegression()

lin\_reg.fit(X\_train,y\_train)

predictions=lin\_reg.predict(X\_test)

print(predictions)

results\_df = pd.DataFrame(data=X\_test)

results\_df['Predicted'] = predictions

results\_df.head(5)

mse=mean\_squared\_error(y\_test,predictions)

r2=r2\_score(y\_test,predictions)