O2O优惠券使用预测-项目进展报告

一、数据获取及预处理

在2016年1月1日至2016年6月30日之间真实线上线下消费行为中，数据集中有三个数据表：用户线下消费优惠券领取行为表、用户线上点击/消费和优惠券领取行为。

Table 1: 用户线下消费和优惠券领取行为

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data\_type** | **Description** | **example** |
| User\_id | String | 用户ID | 1439408 |
| Merchant\_id | string | 商户ID | 4663 |
| Coupon\_id | String | 优惠券ID：null表示无优惠券消费，此时Discount\_rate和Date\_received字段无意义 | 11002 |
| Discount\_rate | string | 优惠率：x \in [0,1]代表折扣率；x:y表示满x减y。单位是元 | 150:20:00 |
| Distance | int | user经常活动的地点离该merchant的最近门店距离是x\*500米（如果是连锁店，则取最近的一家门店），x\in[0,10]；null表示无此信息，0表示低于500米，10表示大于5公里； | 1 |
| Date\_received | String | 领取优惠券日期 | 20160528 |
| Date | String | 消费日期：如果Date=null & Coupon\_id != null，该记录表示领取优惠券但没有使用，即负样本；如果Date!=null & Coupon\_id = null，则表示普通消费日期；如果Date!=null & Coupon\_id != null，则表示用优惠券消费日期，即正样本； | 20160516 |

记录数：1754884

Table 2: 用户线上点击/消费和优惠券领取行为

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data\_type** | **Description** | **example** |
| User\_id | string | 用户ID | 13740231 |
| Merchant\_id | string | 商户ID | 18907 |
| Action | int | 0 点击， 1购买，2领取优惠券 | 2 |
| Coupon\_id | String | 优惠券ID：null表示无优惠券消费，此时Discount\_rate和Date\_received字段无意义。“fixed”表示该交易是限时低价活动。 | 100017492 |
| Discount\_rate | String | 优惠率：x \in [0,1]代表折扣率；x:y表示满x减y；“fixed”表示低价限时优惠； | 500:50:00 |
| Date\_received | String | 领取优惠券日期 | 20160513 |
| Date | String | 消费日期：如果Date=null & Coupon\_id != null，该记录表示领取优惠券但没有使用；如果Date!=null & Coupon\_id = null，则表示普通消费日期；如果Date!=null & Coupon\_id != null，则表示用优惠券消费日期； | null |

记录数：11429826

Table3: 用户O2O线下优惠券使用预测样本

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data\_type** | **Description** | **example** |
| User\_id | string | 用户ID | 4129537 |
| Merchant\_id | string | 商户ID | 450 |
| Coupon\_id | String | 优惠券ID | 9983 |
| Discount\_rate | String | 优惠率：x \in [0,1]代表折扣率；x:y表示满x减y. | 30:05:00 |
| Distance | String | user经常活动的地点离该merchant的最近门店距离是x\*500米（如果是连锁店，则取最近的一家门店），x\in[0,10]；null表示无此信息，0表示低于500米，10表示大于5公里； | 500:50:00 |
| Date\_received | String | 领取优惠券日期 | 20160712 |

记录数：113640

数据预处理：

1. 主要是类型转换，
2. 空值填充
3. 将Discount\_rate的满减形式都转换为折扣率。

二、数据分析与可视化

### 1.对用户线下消费和优惠券领取行为表进行分析：

* 有优惠券，购买商品条数： 75382
* 无优惠券，购买商品条数： 701602
* 有优惠券，不购买商品条数：977900
* 无优惠券，不购买商品条数： 0
* 在测试集中出现的用户但训练集没有出现：[2495873L, 1286474L]
* 在测试集中出现的商户但训练集没有出现：[5920L]

说明：

（1）数据中大量的购物券没有花掉(977900)，其中很多人购买商品没有优惠券(701602), 用优惠券购买商品的只有很少一部分人(75382)。

（2）dftest中有2个用户是新的，没有出现在dfoff中。dftest中有1个商铺是新的，没有出现在dfoff中。

### 2.优惠券和距离

Discount\_rate 类型:

[nan,'150:20', '20:1', '200:20', '30:5', '50:10', '10:5', '100:10',

'200:30', '20:5', '30:10', '50:5', '150:10', '100:30', '200:50',

'100:50', '300:30', '50:20', '0.9', '10:1', '30:1', '0.95',

'100:5', '5:1', '100:20', '0.8', '50:1', '200:10', '300:20',

'100:1', '150:30', '300:50', '20:10', '0.85', '0.6', '150:50',

'0.75', '0.5', '200:5', '0.7', '30:20', '300:10', '0.2', '50:30',

'200:100', '150:5']

Distance 类型:

['0', '1', nan, '2', '10', '4', '7', '9', '3', '5', '6', '8']

处理换算为打折率，处理方式：

（1）将满xx减yy类型(xx:yy)的券变成折扣率 : 1 - yy/xx，同时建立折扣券相关的特征 discount\_rate, discount\_man, discount\_jian, discount\_type

（2）将距离str转为int。

线下打折率取值：

[nan 0.86666667 0.95 0.9 0.83333333 0.8

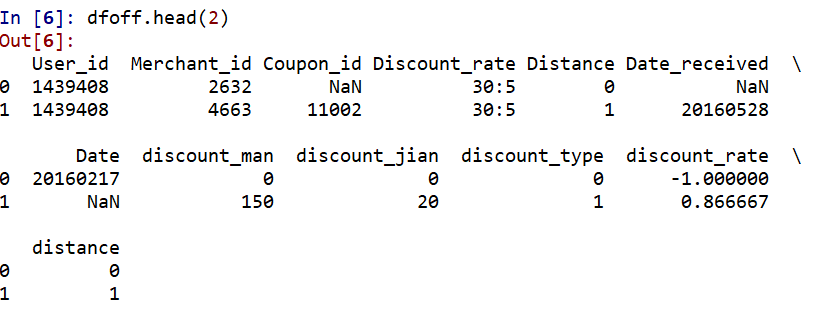
0.5 0.85 0.75 0.66666667 0.93333333 0.7

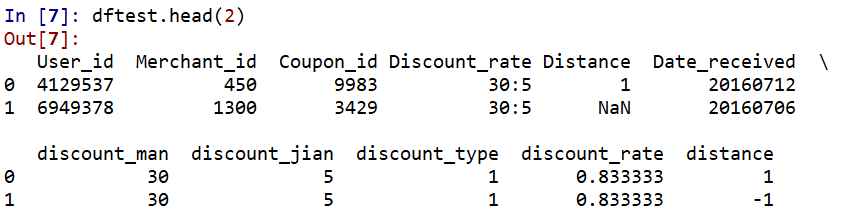
0.6 0.96666667 0.98 0.99 0.975 0.33333333

0.2 0.4 ]

线下距离取值：

[ 0 1 nan 2 10 4 7 9 3 5 6 8]

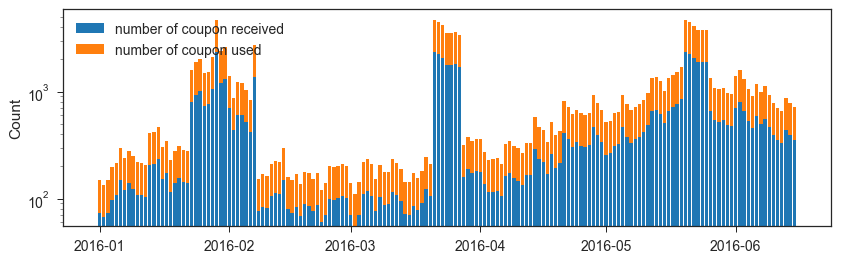




### 3.分析时间

* 优惠券收到日期: 20160101~ 20160615
* 消费日期: 20160101~ 20160630

每天的顾客收到coupon的数目，以及收到coupon后用coupon消费的数目，如图所示。



三、模型选取

* 特征值：discount, distance, weekday
* train/valid 的划分：用20160101到20160515的作为train，20160516到20160615作为valid。
* 模型：用线性模型 SGDClassifier

四、挖掘实验的结果

运行时间超过2小时，人为中断，未得到结果。

五、存在的问题

运行时间长，查找是前期处理是否存在错误，或寻找运行时间短的模型。

六、下一步工作

在数据的划分、特征提取、模型选取上进一步研究。