O2O优惠券使用预测-最终报告

崔杰

# 1 简介

随着移动设备的完善和普及，移动互联网+各行各业进入了高速发展阶段，这其中以O2O（Online to Offline）消费最为吸引眼球。据不完全统计，O2O行业估值上亿的创业公司至少有10家，也不乏百亿巨头的身影。O2O行业天然关联数亿消费者，各类APP每天记录了超过百亿条用户行为和位置记录，因而成为大数据科研和商业化运营的最佳结合点之一。以优惠券盘活老用户或吸引新客户进店消费是O2O的一种重要营销方式。然而随机投放的优惠券对多数用户造成无意义的干扰。对商家而言，滥发的优惠券可能降低品牌声誉，同时难以估算营销成本。个性化投放是提高优惠券核销率的重要技术，它可以让具有一定偏好的消费者得到真正的实惠，同时赋予商家更强的营销能力。

我们分别从数据处理，数据分析，结果评价和结果展示四个部分入手，解决O2O优惠券使用的预测问题。

# 2 问题陈述

我们需要解决的问题是，如何从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，对2016年7月领取优惠券后15天以内的使用情况进行预测。

比赛中提供了三个数据表：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及用户O2O线下优惠券使用预测样本。

Table 1: 用户线下消费和优惠券领取行为

|  |  |  |  |
| --- | --- | --- | --- |
| **Field** | **Data\_type** | **Description** | **example** |
| User\_id | int | 用户ID | 1439408 |
| Merchant\_id | int | 商户ID | 4663 |
| Coupon\_id | int | 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 | int | 用户ID | 13740231 |
| Merchant\_id | int | 商户ID | 18907 |
| Action | int | 0 点击， 1购买，2领取优惠券 | 2 |
| Coupon\_id | int | 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 | Int | 用户ID | 4129537 |
| Merchant\_id | int | 商户ID | 450 |
| Coupon\_id | int | 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 | 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 | 20160712 |

记录数：113640

本课题基于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，预测用户在2016年7月领取优惠券后15天以内的使用情况，如Table4，其中user\_id,coupon\_id和date\_received均来自Table 3,而Probability为预测值。使用优惠券核销预测的平均AUC（ROC曲线下面积）作为评价标准。 即对每个优惠券coupon\_id单独计算核销预测的AUC值，再对所有优惠券的AUC值求平均作为最终的评价标准。

Table 4：预测结果

|  |  |
| --- | --- |
| **Field** | **Description** |
| User\_id | 用户ID |
| Coupon\_id | 优惠券ID |
| Date\_received | 领取优惠券日期 |

# 3 技术方案

数据集提供数据的区间是2016-01-01~2016-06-30，预测七月份用户领券使用情况，即用或者不用，转化为二分类问题，然后通过分类算法预测结果。首先就是特征工程，其中涉及对数据集合的划分，包括提取特征的区间和训练数据区间。之后是从特征区间中提取特征，包括用户特征、商户特征、优惠券特征、用户商户组合特征、用户优惠券组合特征，最后选择训练模型学习并预测。

## 3.1 数据预处理

在提供的数据集中，存在很多空值，但有指定意义，不需要填充。

由于折扣率Discount\_rate通过两种方式表示，优惠率或满x减y，需要预处理，都转变为优惠率。

（1）处理方式：将类型str变成 float。

（2）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。

## 3.2 数据分析

### 1.读取数据以及简单分析

对用户线下消费和优惠券领取行为表进行分析，结果如下：

* 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）测试集中有2个用户是新的，没有出现在训练集中。测试集中有1个商铺是新的，没有出现在训练集中。

2.时间分析

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1所示。

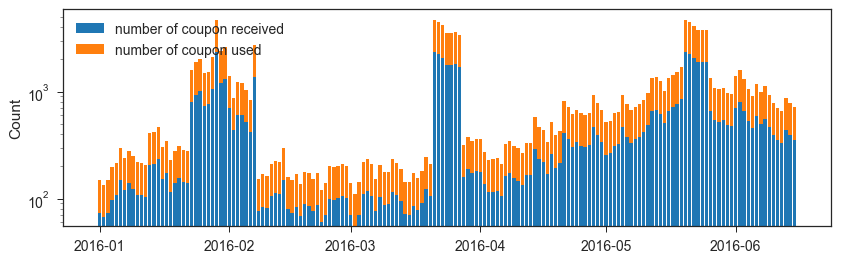


图1

### 2.特征工程

本次课题提取了三大特征类：用户特征、商户特征、用户商户组合特征，数据集包括online和offline的数据，由于里面只有部分用户重合，商户优惠券等并未有重合，因此只在online提取了用户特征。而offline数据集就提取了所有三个特征类，以下是各部分特征：

1. 用户特征

* 用户收到的优惠券数量
* 用户购买离线的次数（带或不带优惠券）
* 用户离线购买次数（带优惠券）
* 领取优惠券的商户数
* 用户经常活动的地点离商家门店距离最小值
* 用户经常活动的地点离商家门店距离最大值
* 用户经常活动的地点离商家门店距离平均值
* 用户经常活动的地点离商家门店距离中位值

1. 商户特征

* 商家的优惠券数量
* 商家销售数量（带或不带优惠券）
* 使用优惠券的商家销售数量
* 商家门店与用户经常活动的地点距离的最小值
* 商家门店与用户经常活动的地点距离的最大值
* 商家门店与用户经常活动的地点距离的平均值
* 商家门店与用户经常活动的地点距离的中位值

1. 用户和商户组合特征

* 用户在商店消费或领优惠券合计次数
* 用户在商店消费次数
* 用户在商店领优惠券次数
* 用户在商店使用优惠券次数
* 用户在商户消费率
* 用户在商户优惠券使用率
* 用户在商户消费中用优惠券率

### 3.数据标注

在训练集中，进行类标注，分为三类：

（1）没领优惠券消费，计701602元组。

（2）领取优惠券后15天以内消费，计64395元组。

（3）其他情况，计988887元组。

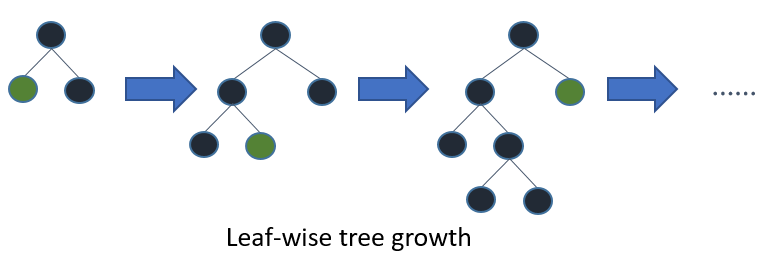
# 4 实现与结果分析

## 4.1 LightGBM简介

数据挖掘目的就是要根据大量的数据构建一个可以准确预测未知情况的模型，当前训练模型的算法有很多，很多算法都可以训练出一个可以准确预测数据的模型。在本次的项目中利用GBM light进行训练。

LightGBM 是一个梯度 boosting 框架，使用基于学习算法的决策树，是分布式的，高效的，与常用的机器学习算法速度快，具有以下优势：

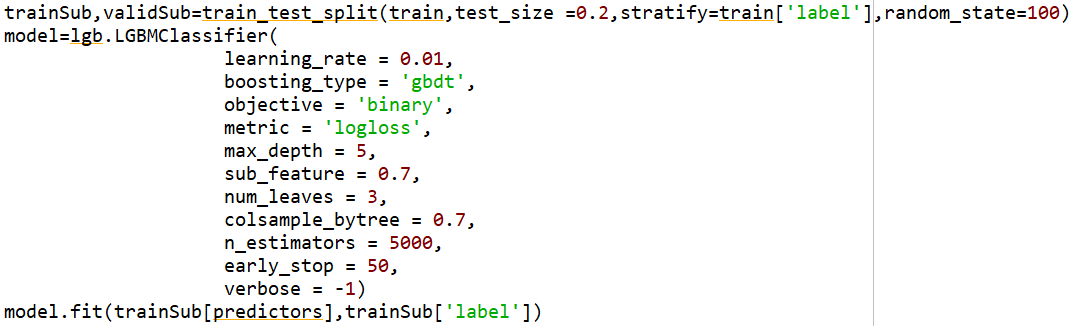
* 速度和内存使用的优化。LightGBM 利用基于 histogram 的算法，通过将连续特征（属性）值分段为 discrete bins 来加快训练的速度并减少内存的使用。
* 稀疏优化。
* 准确率的优化。
* Leaf-wise (Best-first) 的决策树生长策略。它将选取具有最大 delta loss 的叶节点来生长。 当生长相同的 #leaf，leaf-wise 算法可以比 level-wise 算法减少更多的损失。



* 类别特征值的最优分割。
* 网络通信的优化。
* 并行学习的优化。包括特征并行、数据并行、投票并行。
* 应用和度量。回归，目标函数为 L2 loss；二分类，目标函数为 logloss（对数损失）；多分类；lambdarank目标函数为基于NDCG的lambdarank。
* 可处理大规模数据

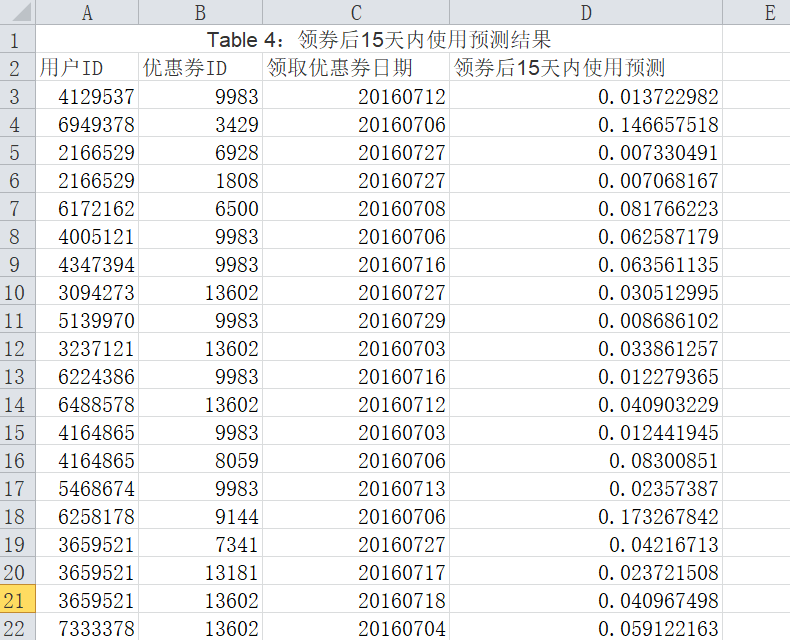
## 4.2 模型建立

在LightGBM中,验证数据与训练格式一致，在20160101到20160515之间的数据提取特征，20160516-20160615的数据作为训练集，参数设置如图：



## 4.3 结果分析

经训练，对每个优惠券coupon\_id单独计算核销预测的AUC值，再对所有优惠券的AUC值求平均作为最终的评价标准，结果为0.6350835527024095，对用户O2O线下优惠券使用预测样本进行预测，结果如图所示。



此次计算，优惠券核销预测的平均AUC（ROC曲线下面积）为0.6350835527024095，说明准确率较低，还需在数据划分、特征值选取、模型方面做进一步研究。

## 4.4 总结

此课题是天池新人实战赛一个经典赛题，提供帮助教程，通过学习实践，对学习数据挖掘很有帮助，此次只练习了一种模型，还需进一步学习实践，提高预测准确率。