O2O优惠券使用预测比赛总结

目录

[1.题目背景 2](#_Toc489801854)

[2.解决方案： 2](#_Toc489801855)

[2.1解决思路 2](#_Toc489801856)

[2.2数据集划分 3](#_Toc489801857)

[特征提取 3](#_Toc489801858)

[2.3训练+预测 3](#_Toc489801859)

[2.4规则 3](#_Toc489801860)

[3.算法介绍 4](#_Toc489801861)

[4.项目发展 4](#_Toc489801862)

[5.收获总结 4](#_Toc489801863)

## 1.题目背景

赛题提供用户在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，预测用户在2016年7月领取优惠券后15天以内的使用情况。

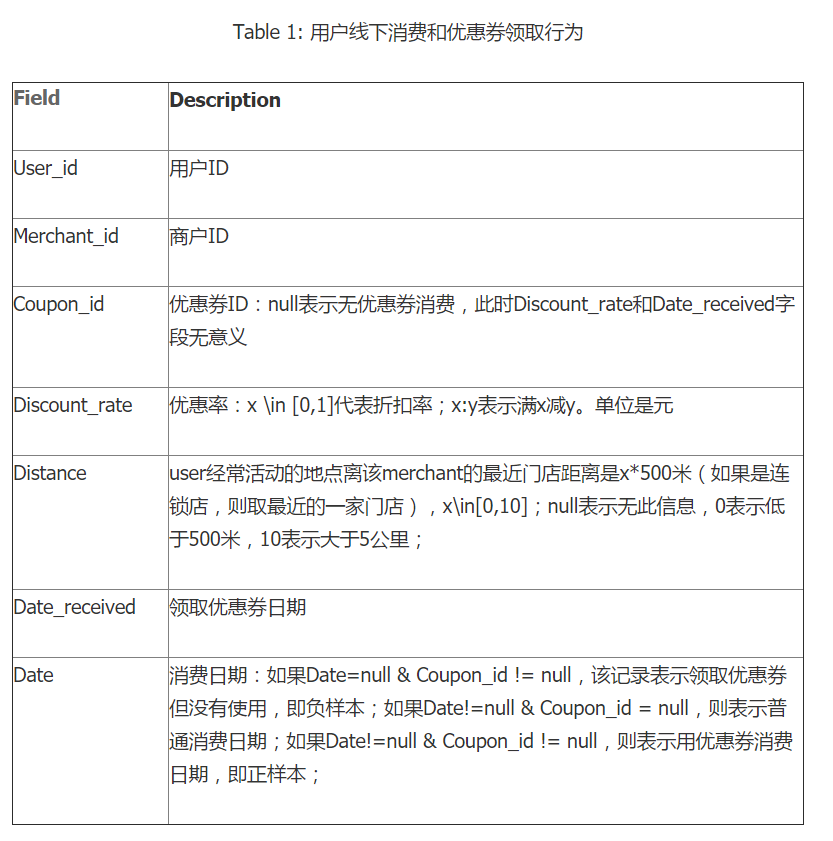
使用优惠券核销预测的**平均AUC**（ROC曲线下面积）作为评价标准。 即对每个优惠券coupon\_id单独计算核销预测的AUC值，再对所有优惠券的AUC值求平均作为最终的评价标准。

## 2.解决方案：

2.1解决思路

考虑目标是预测用户的使用情况，分为用或者不用两类情况，转化为二分类问题

数据如下



2.2数据集划分

**划窗法**的使用

|  |  |  |
| --- | --- | --- |
| 用处 | 特征提取 | 预测时间 |
| 训练(11w) | 2016/01/01-2016/04/13 | 2016/04/14-2016/05/14 |
| 训练(25w) | 2016/02/01-2016/05/14 | 2016/05/15-2016/06/15 |
| 测试(13w) | 2016/03/15-2016/06/30 | 2016/07/01-2016/07/30 |

2.3特征提取

主要包括五大特征：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，赛题包括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，不考虑线上数据。而offline数据集就提取了所有五个特征类。

2.4训练+预测

初赛使用xgboost模型，复赛在ODPS平台使用了GBDT模型

2.5规则

在赛题提供 的预测集合中，包含了一个用户在整个7月的优惠券领取情况，假设用户7月10号领取了某优惠券，在7月12号又领取了相同的优惠券，那么7月10号的优惠券的使用概率就很大，利用这个规则，效果提高了10%

## 3.算法介绍

## GBDT ??

## XGboost

## Randforest

## 4.项目发展

## 最开始的时候准确率54%-🡪 62% 🡪 73%

## 特征改变10% 体征的提取

## 使用了leakage

## 5.收获总结

**组合模型??**