天猫复购预测

https://tianchi.aliyun.com/competition/entrance/231576/information

背景

商家有时会在特定日期，例如Boxing-day，黑色星期五或是双十一（11月11日）开展大型促销活动或者发放优惠券以吸引消费者，然而很多被吸引来的买家都是一次性消费者，这些促销活动可能对销售业绩的增长并没有长远帮助，因此为解决这个问题，商家需要识别出哪类消费者可以转化为重复购买者。通过对这些潜在的忠诚客户进行定位，商家可以大大降低促销成本，提高投资回报率（Return on Investment, ROI）。众所周知的是，在线投放广告时精准定位客户是件比较难的事情，尤其是针对新消费者的定位。不过，利用天猫长期积累的用户行为日志，我们或许可以解决这个问题。

我们提供了一些商家信息，以及在“双十一”期间购买了对应产品的新消费者信息。你的任务是预测给定的商家中，哪些新消费者在未来会成为忠实客户，即需要预测这些新消费者在6个月内再次购买的概率。

数据描述

数据集包含了匿名用户在 "双十一 "前6个月和"双十一 "当天的购物记录，标签为是否是重复购买者。出于隐私保护，数据采样存在部分偏差，该数据集的统计结果会与天猫的实际情况有一定的偏差，但不影响解决方案的适用性。训练集和测试集数据见文件data_format1.zip，数据详情见下表。

用户行为日志

字段名称	描述
user_id	购物者的唯一ID编码
item_id	商品的唯一编码
cat_id	商品所属品类的唯一编码
merchant_id	商家的唯一ID编码
brand_id	商品品牌的唯一编码
time_stamp	购买时间（格式：mmdd）
action_type	包含{0,1,2,3}，0表示单击，1表示添加购物车，2表示购买，3表示添加收藏夹

用户画像

字段名称	描述
user_id	购物者的唯一ID编码
age_range	用户年龄范围。<18岁为1；[18,24]为2； [25,29]为3； [30,34]为4；[35,39]为5；[40,49]为6； > = 50时为7和8; 0和NULL表示未知
gender	用户性别。0表示女性，1表示男性，2和NULL表示未知

训练数据和测试数据

字段名称	描述
user_id	购物者的唯一ID编码
merchant_id	商家的唯一ID编码
label	包含{0, 1}，1表示重复买家，0表示非重复买家。测试集这一部分需要预测，因此为空。

任务

1. 分别编写MapReduce程序和Spark程序统计双十一最热门的商品和最受年轻人(age<30)关注的商家（“添加购物车+购买+添加收藏夹”前100名）；
2. 编写Spark程序统计双十一购买了商品的男女比例，以及购买了商品的买家年龄段的比例；
3. 基于Hive或者Spark SQL查询双十一购买了商品的男女比例，以及购买了商品的买家年龄段的比例；
4. 预测给定的商家中，哪些新消费者在未来会成为忠实客户，即需要预测这些新消费者在6个月内再次购买的概率。基于Spark MLlib编写程序预测回头客，评估实验结果的准确率。

提交方式

1. git仓库地址（包括代码、结果和实验报告）
2. 报名参加竞赛，按照要求提交预测结果至网站，给出得分和排名。