|  |
| --- |
| 项目背景： |
| 在抖音商城电商平台上，连接了大量的商家店铺以及消费者， 在获得用户量级增长的同时，也面临着业务侧沉淀风控能力的需求，作为电商平台的风险包括但不限于用户、设备、地理位置等不同维度的风险特征，此外也时常受到黑灰产的攻击。我们考察的两大核心指标，用户增长和风控能力，也是我们这个项目主要评估的两方面内容。 |
|  |
| 目标：【第一部分】 |
| 对抖音商城电商平台上的用户交易业务以及行为特征进行分析，找到显著的、有区分度的特征，并写成脚本供线上策略调用，从而提升电商风控引擎对恶意退货/刷单以及黑产攻击的识别和覆盖能力，以及刻画新兴电商平台的用户增长趋势。 |
|  |
| 步骤： |
| 完成驱动表，并给黑白事件打标； |
|  |
|  |

-------------------------------------------------------------------------------

**Task 1**

**Mysql配置：**

**Username：wyz**

**Password：Qwer1234**

**Host：119.3.110.197**

**Port：3307**

**Database：data\_dy**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 完成驱动表 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| 通常我们会按照以下的步骤来做一张表的清洗，并得到驱动表，所谓的驱动表，也就是包含了黑白样本【相当于正负例】标记的记录表； | | | | | | |  |
| 在这个case里面，我们面对的是一个二分类问题，黑或白就相当于通常我们说的正负类，分类问题的本质是在于标签，所以我们通过打标构建一张驱动表 | | | | | | | |
|  |  |  |  |  |  |  |  |
| 步骤 |  |  |  |  |  |  |  |
| 1. 选出黑样本 | | 时间窗选取在20191201-20200115 | | |  |  |  |
| 2. 选出白样本 | | 时间窗选取在20191201-20200115 | | |  |  |  |
| 3. 计算黑白样本的行数，并对白样本进行采样 | | | | | | | |
| 4. 黑白事件打标并得到驱动表 (驱动表里的字段包括交易信息表的所有字段 + 黑白样本的标注字段 ) | | | | | | |  |

【mysql里面可以使用rand()函数进行采样，比如：SELECT  \*  FROM table\_name ORDER BY rand ()  LIMIT  5】

现有的表：

|  |
| --- |
| account\_all 电商平台整体的交易账户信息表 |
| account\_black 业务侧的运营同事整理好的黑名单表【包含了刷单、恶意退货等黑事件】 |

**account**\_**all 主键是 userid**

**account**\_**black 主键是event id**

1.

select \*, ‘0’ as tag from account\_black where dt between 20191201 and 20200115 (黑样本)

2.

select \*. ‘1’ as tag from account\_all left outer join account\_black on account\_all.userid = account\_black.eventid where account\_all.dt between 20191201 and 20200115 （白样本）

3

Select count(\*) from account\_black where dt between 20191201 and 20200115 (200)

select count(\*) from account\_all left outer join account\_black on account\_all.userid = account\_black.eventid where account\_all.dt between 20191201 and 20200115 (200)（计算黑白样本的行数）

select \* from account\_all left outer join account\_black on account\_all.userid = account\_black.eventid where account\_all.dt between 20191201 and 20200115 ORDER BY rand ()  （采样白样本）

**4.**

**select t1.\*,t2.\* from**

**(**

**select\*, '0' as tag from account\_black where dt between 20191201 and 20200115**

**)t1**

**cross join**

**(select \*,'1' as tag from account\_all left outer join account\_black on account\_all.userid = account\_black.eventid where account\_all.dt between 20191201 and 20200115**

**)t2 (打标，创建驱动表)**

**--整理黑白样本**

**drop table if exists event\_black; --黑样本的口径**

**create table event\_black as**

**select \***

**from account\_all -- 所有的**

**where user\_id in**

**(select user\_id**

**from account\_black**

**)**

**and dt between '20191201' and '20200115'**

**---整理白样本**

**drop table if exists event\_white;**

**create table event\_white as**

**select \***

**from (select \***

**from account\_all**

**where user\_id not in**

**( select user\_id**

**from account\_black**

**)**

**and dt between '20191201' and '20200115'**

**) a**

**where sample(27,1,eventid) = true -- sample 函数 表示数值会对每行数据进行随机哈希分配，sample(4,1,eventid) = true, 随机分配为4份，取第1份，这里的eventid是对eventid这个字段抽样；**

**;**

**select count(\*）**

**from event\_black --验证记录数**

**select count(\*）**

**from event\_white --验证记录数**

**--给事件打标，做成驱动表；**

**drop table if exists customer\_drivertable;**

**create table customer\_drivertable as**

**select \***

**,1 as target**

**from event\_black**

**union all**

**select \***

**,0 as target**

**from event\_white**

**;**