基于用户行为的黑产识别

By Susanna Lee

一、 背景介绍

随着互联网+这一概念的不断发展，电商、出行、外卖等行业近些年也持续发展壮大，越来越多的商家进入这一市场。为了在激烈的竞争中拉取新用户，培养用户的消费习惯，各种类型的营销活动和补贴活动也是层出不穷。在为正常用户带来福利的同时，也催生了一批专注于营销活动的“羊毛党”。目前，羊毛党的行为越发专业化，团伙化和地域化，同套利黑产团伙的斗争，是一场永无止境的攻防战。机器学习模型是风控系统中实时识别和对抗黑产攻击的有效手段。面对黑产攻击手段快速多变，黑样本数据标签缺失等问题，目前除了 LR,RF 等耳熟能详的机器学习模型，基于 RNN 的深度学习模型，无监督学习模型等技术也被应用到同黑产的对抗中。

本文通过训练学习用户在消费过程中的关联操作、交易详单信息，来识别交易风险。

1.1 数据来源

https://www.dcjingsai.com/common/cmpt/2018%E5%B9%B4%E7%94%9C%E6%A9%99%E9%87%91%E8%9E%8D%E6%9D%AF%E5%A4%A7%E6%95%B0%E6%8D%AE%E5%BB%BA%E6%A8%A1%E5%A4%A7%E8%B5%9B_%E7%AB%9E%E8%B5%9B%E4%BF%A1%E6%81%AF.html

1.2 数据集基本情况简介

- **operation**：用户操作详情表单，训练集共 1460843 条数据，测试集共 1140578 条数据。

操作详单数据字典

字段名	中文解释	字段说明
UID	用户编号	
day	操作日期	连续的日期标识， E.g., 1为第一天，2为第二天，以此类推
mode	操作类型	操作类型（例如：修改密码、查询余额…）
success	操作状态	
time	操作时间点	
os	操作系统	
version	客户端版本号	
device1	操作设备参数1	设备名称加密，原字段如“*Jack’s iphone*”
device2	操作设备参数2	设备型号
device_code1	操作设备唯一标识1	设备号唯一标识加密，可用于安卓类设备的唯一标识
device_code2	操作设备唯一标识2	设备号唯一标识加密，可用于安卓类设备的唯一标识
mac1	MAC地址	操作设备MAC地址编码加密， 原字段如“38:XX:XX:XX:XX:92”
ip1	IP地址	操作设备IP地址编码加密
ip2	IP地址	操作电脑IP地址编码加密
device_code3	操作设备唯一标识3	设备号唯一标识加密，可用于苹果类设备的唯一标识
mac2	MAC地址	WIFI MAC地址编码加密， 原字段如“02:XX:XX:XX:XX:03
wifi	WIFI名称	WIFI名称，原字段如“A的wifi”
geo_code	地理位置	经纬度GeoHash编码
ip1_sub	IP地址	前三位操作设备IP地址编码加密（ip1前三位IP地址） 比如，原字段为12，34，56，7和12，34，56，8的ip地址前三位都为12，34，56，故脱敏后的值是一样的
ip2_sub	IP地址	前三位操作电脑IP地址编码加密（ip2前三位IP地址）

- **transaction：**交易详情表单，训练集共共 264654 条数据，测试集共 128382 条数据

交易详单数据字典

字段名	中文解释	字段说明
UID	用户编号	
channel	平台	平台类型
day	交易日期	连续的日期标识， 1为第一天，2为第二天，以此类推
time	交易时间点	
trans_amt	脱敏后交易金额	保留大小关系
amt_src1	资金类型	交易资金源类型，例如“余额”、“银行卡”
merchant	商户标识	商户编码加密
code1	商户标识	商户子门店编码加密
code2	商户终端设备标识	商户交易终端设备编码加密
trans_type1	交易类型1	交易类型，例如“消费”，“退款”
acc_id1	账户相关	用户交易账户号编码加密
device_code1	操作设备唯一标识1	设备号唯一标识加密，可用于安卓类设备的唯一标识
device_code2	操作设备唯一标识2	设备号唯一标识加密，可用于安卓类设备的唯一标识 （唯一标识码并不会只是一种 但都能达到效果）
device_code3	操作设备唯一标识3	设备号唯一标识加密，可用于苹果类设备的唯一标识
device1	操作设备参数1	设备名称加密，原字段如“*Jack’s iphone*”
device2	操作设备参数2	设备型号
mac1	MAC地址	操作设备MAC地址编码加密， 原字段如“38:XX:XX:XX:XX:92”
ip1	IP地址	操作设备IP地址编码加密
bal	脱敏后账户余额	保留大小关系
amt_src2	资金类型	交易资金源类型，与1类型相似，2对银行卡做了细分
acc_id2	账户相关	转账操作的转出账户号编码加密
acc_id3	账户相关	转账操作的转入账户号编码加密
geocode	地理位置	经纬度GeoHash编码
trans_type2	交易类型2	交易类型，例如“线上”、“线下” trans_type2与trans_type1的维度和侧重不同
market_code	营销活动号编码	营销活动号编码加密
market_type	营销活动标识	营销活动类型
ip1_sub	IP地址	前三位操作设备IP地址编码加密（ip1前三位IP地址）

1.3 总体思路

本项目主要构建了两类模型，一类是基本特征模型，主要提取用户单体信息作为特征，围绕用户、商户和设备等特征从流行度统计和频次统计两个方面来衍生新特征；另一类为关系网络特征模型，主要运用 graph embedding 构建一度关联、二度关联的异构网络，通过 Node2Vec 方法提取用户与商户、设备及账户关系信息作为模型特征。最后采用 LightGBM 对两类模型进行训练，加权融合模型结果。

二、 数据清洗

2.1operation&transaction 数据集：

1. 将特征 day 转换为完整日期，第一天为'2018-08-31'，以此往后推。
2. 根据 device_code1、device_code2、device_code3 构造新的特征 device_type 用于判断用户设备是 ios 系统还是 Android 系统，变量‘1’表示 ios，‘2’表示 Android，‘3’为未知。
3. 将 device_code1、device_code2、device_code3 整合为一个变量 device_code。
4. 从 device2 特征中提取出手机品牌作为新的一列特征 device_brand。
5. 对 geo_code 特征进行地址解析，从而得到经纬度(longitude\latitude)、国家(nation)、省份(province)、城市(city)等特征。
6. 从 time 特征中提取出 hour。
7. 根据 nation 判断是否在境内操作从而衍生特征 is_china。
8. 判断设备信息缺失程度 device_miss_cnt 和环境缺失程度 device_miss_cnt。

2.2operation 数据集：

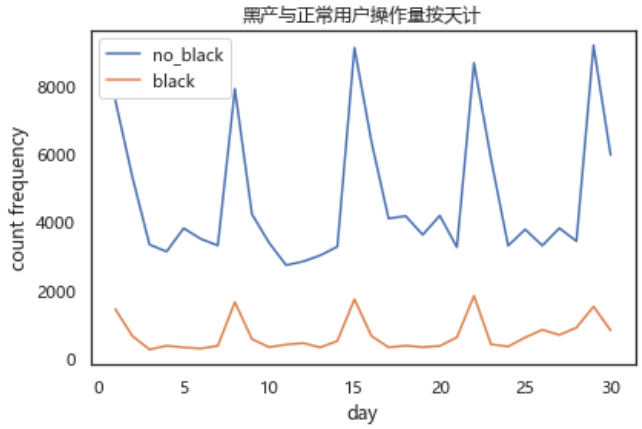
1. 判断是否连接 wifi 操作 is_wifi_env。
2. 根据 ip1、ip2 判断用户是电脑操作还是手机操作，‘1’为电脑，‘2’为手机，‘3’为两者都用。
3. 将 ip1、ip2 合成一个特征 ip，ip1_sub、ip2_sub 合并为 ip_sub。

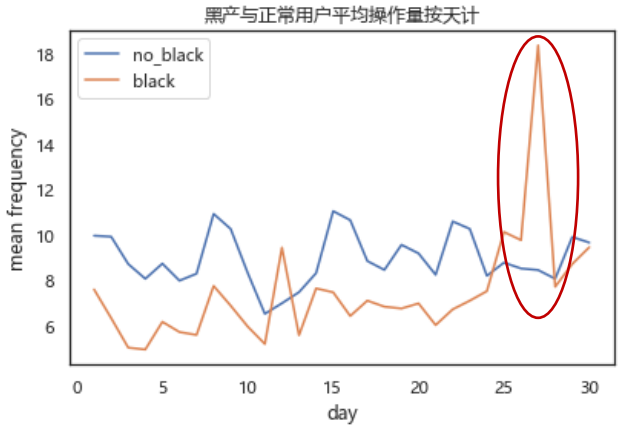
三、 探索性数据分析 EDA

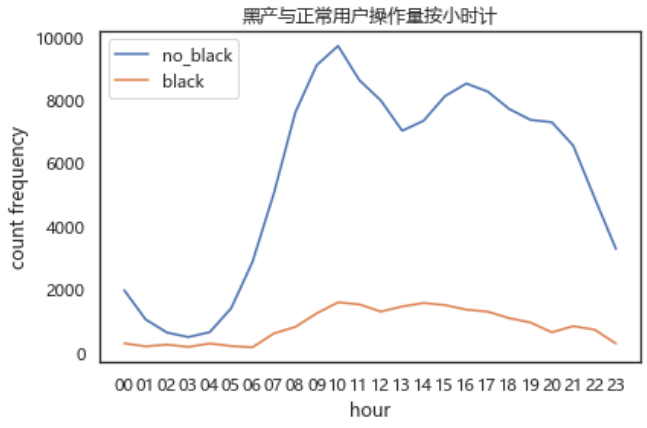
训练集中同时有操作和交易的用户有 29091 个，只有操作记录的用户有 637 个，只有交易记录的用户有 1451 个。

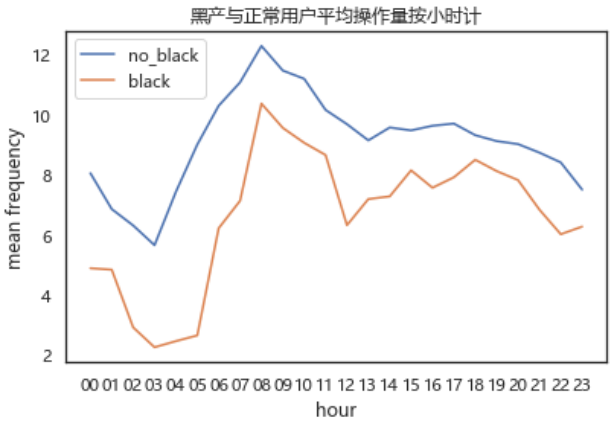
3.1operation 数据集

1. 通过统计用户平均操作次数来观察黑产用户与正常用户之间是否存在明显区别，其中黑产用户平均操作次数为 36 次，而正常用户平均操作次数 51 次，可以看出黑产用户平均操作次数更少，可以认为他们想通过较低的成本获得利益。
2. 分别以天和小时为统计单位，比较黑产与正常用户在一定时间内的总操作量和平均操作量。按天来统计的操作量呈现周期性的趋势，而从按天计的平均操作量来看，绝大时间正常用户的平均操作量高于羊毛党，而是某些时间羊毛党的操作量则显著高于正常用户，后续特征构造会考虑时间对黑产用户的影响。

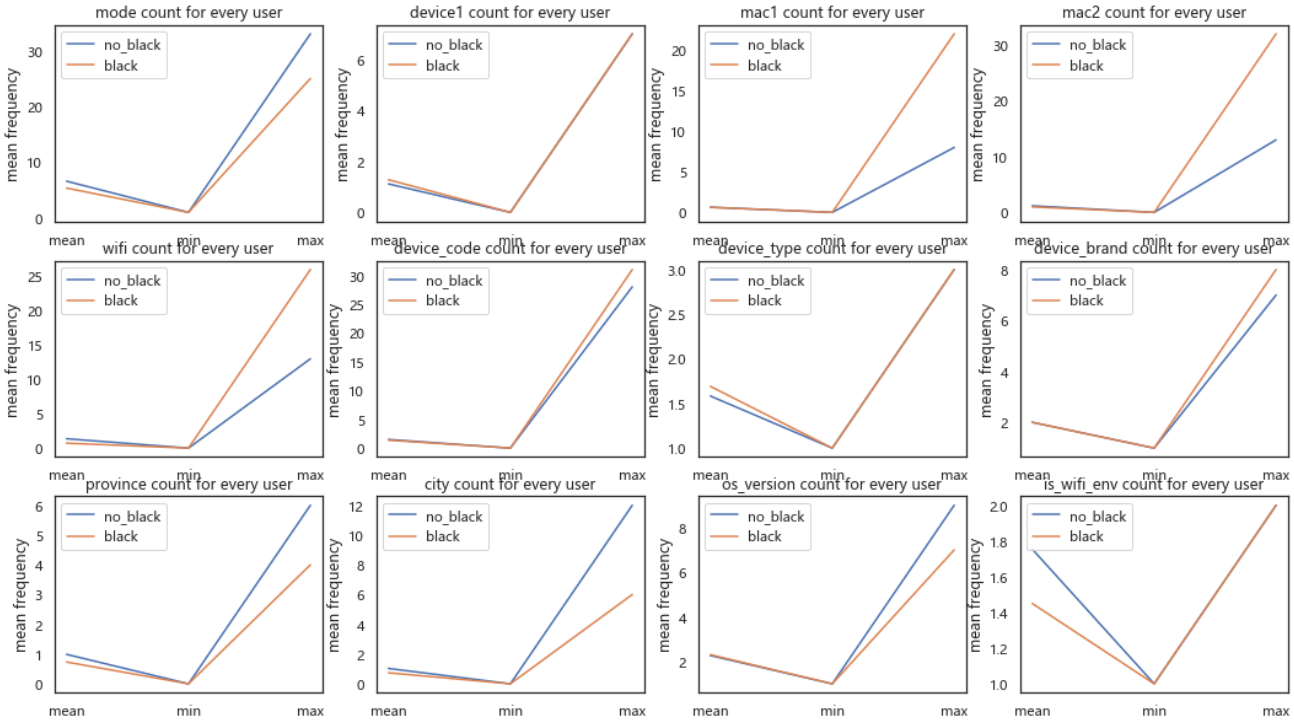








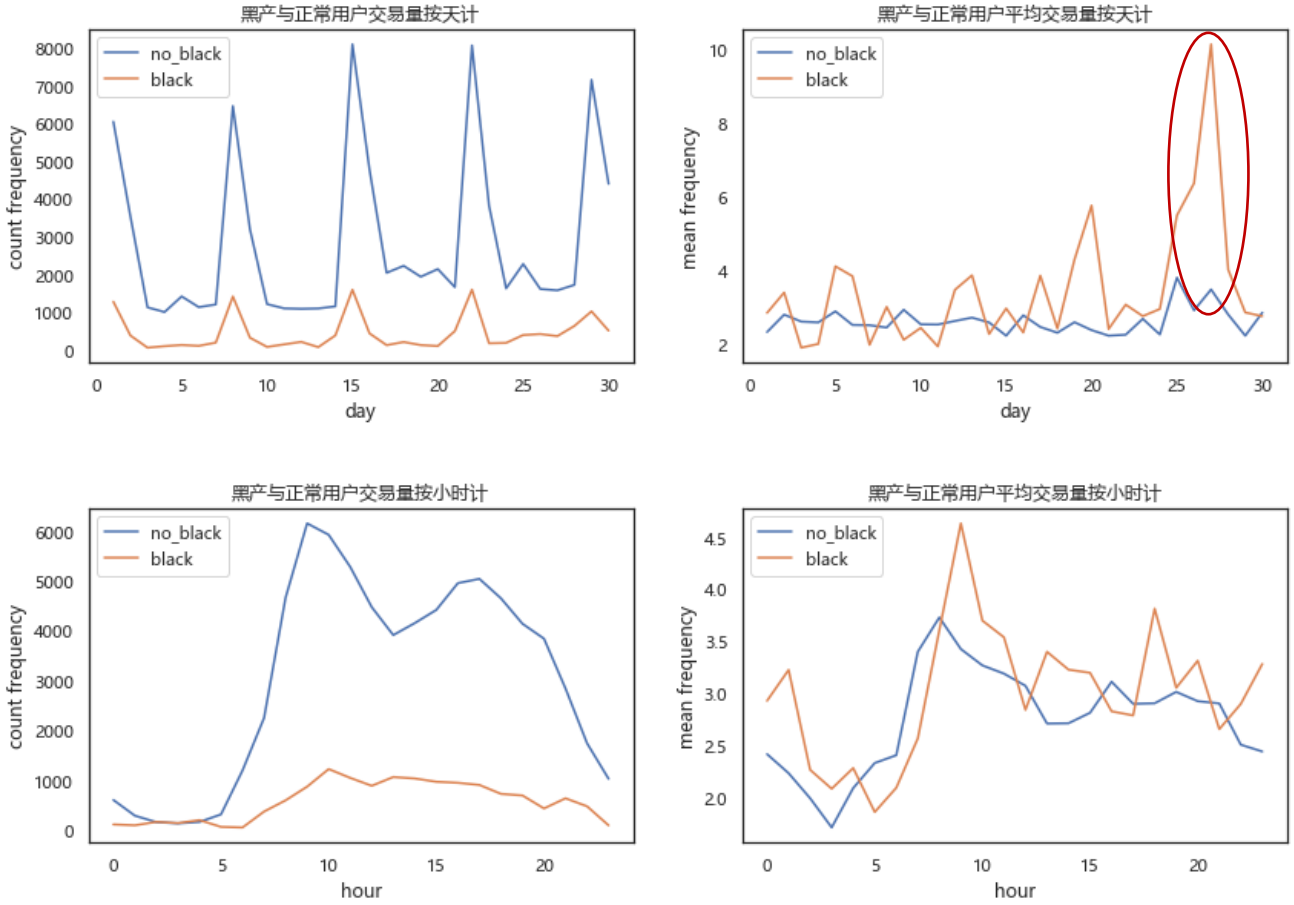
3. 统计每个用户 mode, device1 等特征唯一值的个数，可以看到羊毛党和正常用户在 mac1、mac2、wifi、city 等特征的最大值存在较大区别。



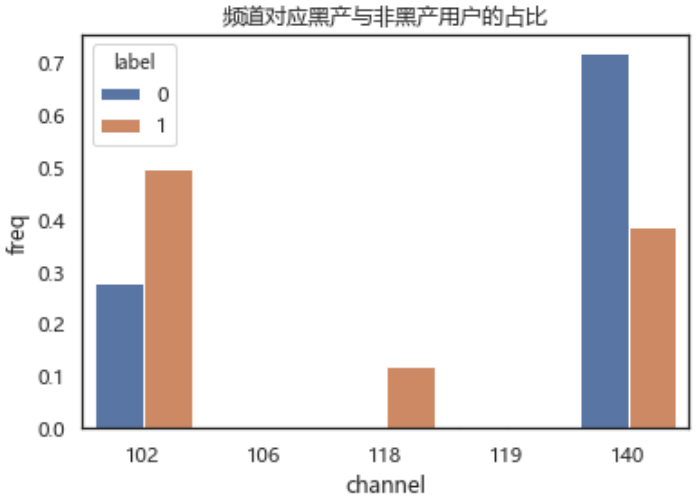
4. 分别统计黑产和正常用户在 mode 特征中最具热度的前 5 个操作类型，并计算两者排前 5 的操作类型共中有几个相同。同时对'device1','mac1','mac2','wifi','device_code','device_type','device_brand','os_version'也做相同处理。其中，特征 mac1、mac2、wifi、device_code 共有的数量为 1、1、0、0，两者区别较大。

3.2transaction 数据集

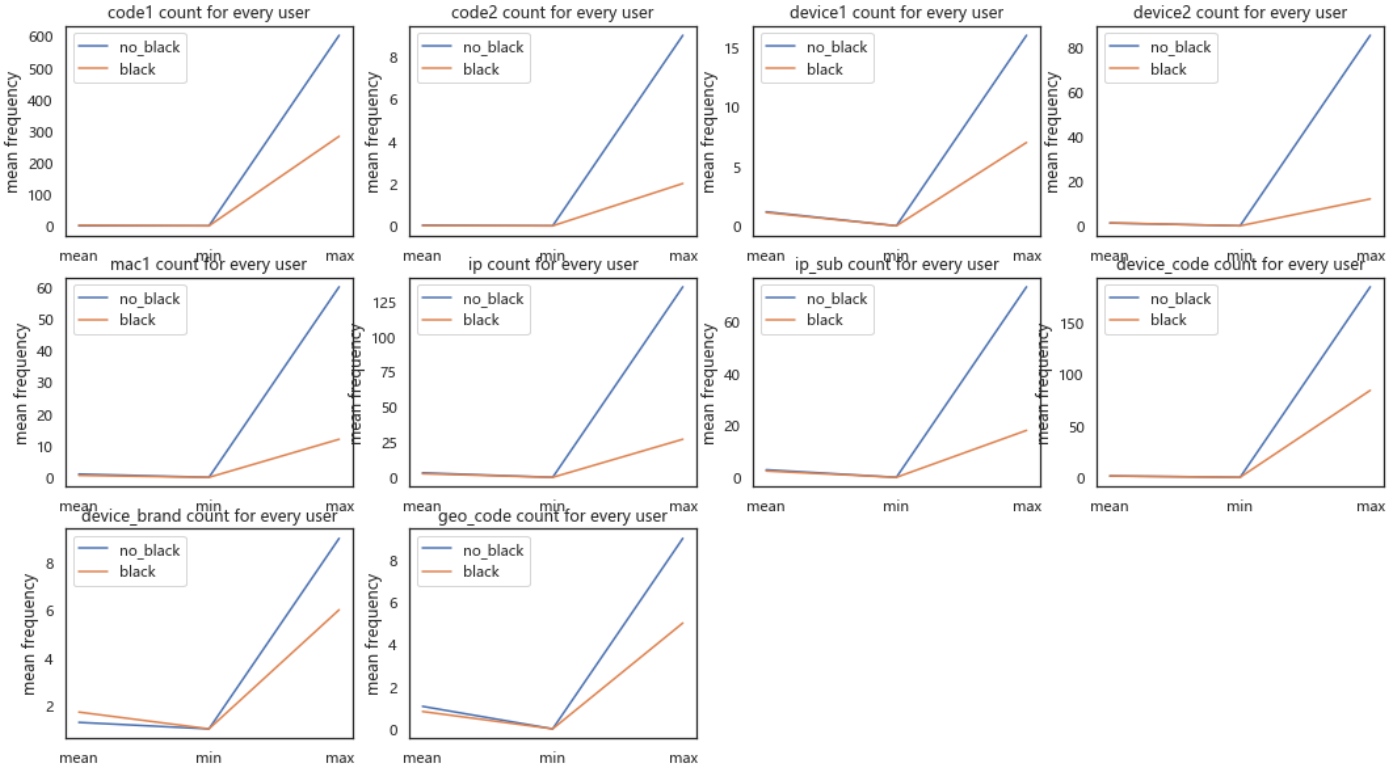
1. 黑产用户平均交易次数为 11 次，而正常用户平均交易次数为 8 次，羊毛党的交易频率略高于正常用户。
2. 分天和小时来看，黑产用户的平均交易量大部分时间也高于正常用户。值得注意的一点是，在大概 25-30 天之间的时间段黑产用户的平均交易量和平均操作量都显著增高，黑产用户可能集中在某个时间段操作，参加完活动可能就不再登录。



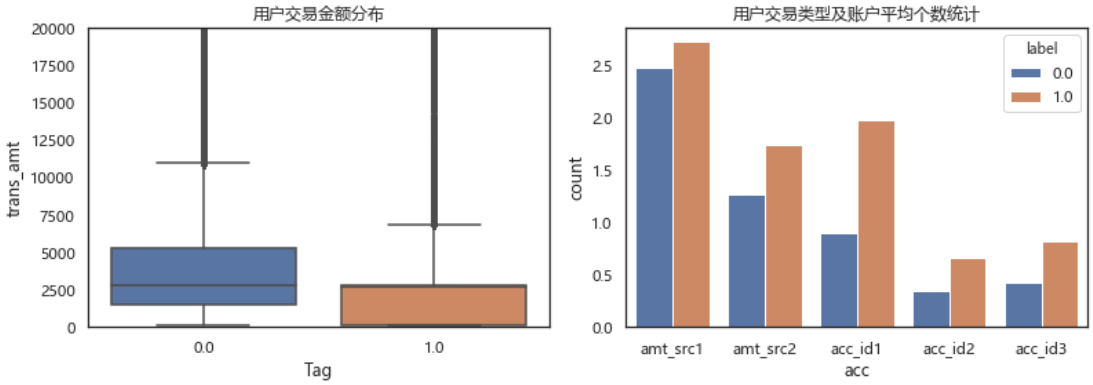
3. 黑产用户 channel 对应的平均种类为 1.725，正常用户为 1.3，黑产用户对应的的频道数量明显高于正常用户。细分到具体渠道，可以看到在 102 渠道下，黑产用户显著高于正常用户，而 106 和 119 渠道都不怎么活跃，而 118 渠道下，正常用户占比几乎为 0，而黑产客户较多，在 140 频道活跃显著低于正常用户。



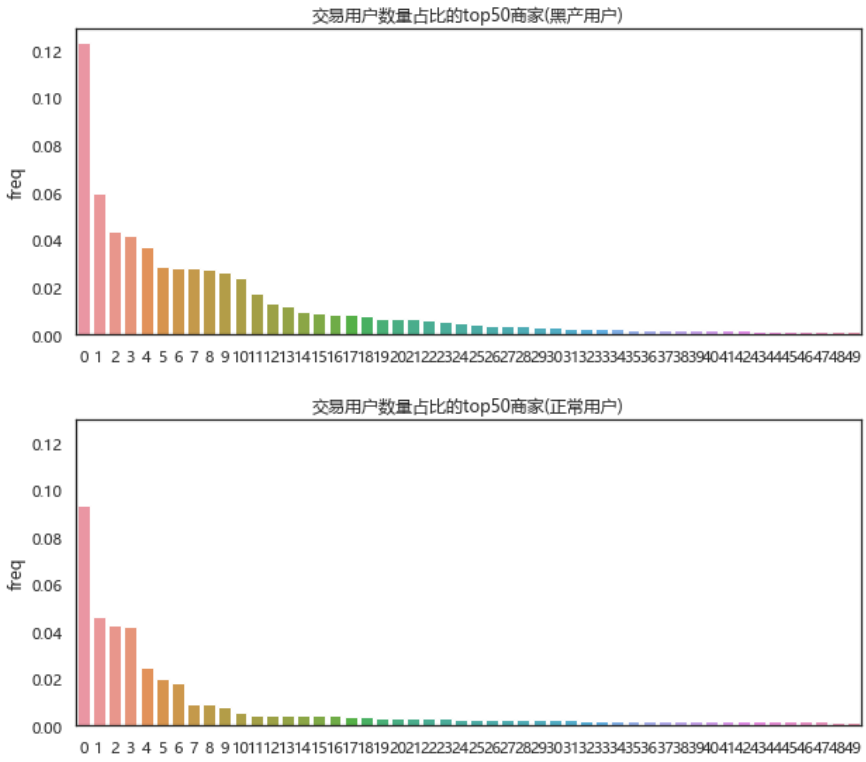
4. 统计每个用户 mode, device1 等特征唯一值的个数，可以看到羊毛党和正常用户在这些特征的最大值都存在较大区别。



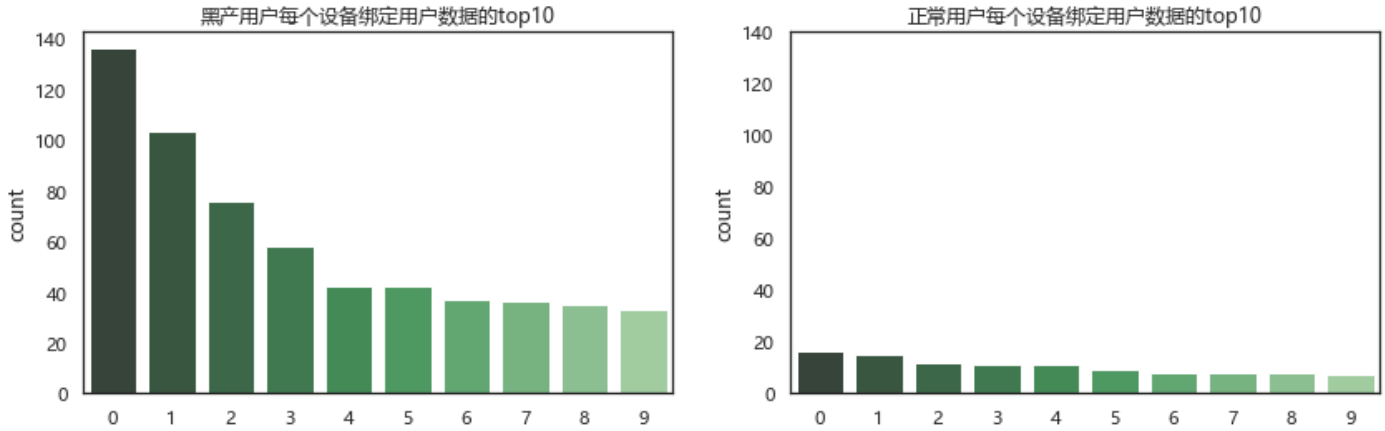
5. 羊毛党交易金额额度显著低于正常用户且交易金额容易固定。同时黑产用户的交易类型(amt_src1\ amt_src2)、交易账户(acc_id1)以及转出转入账户(acc_id2\acc_id3)的平均个数都显著高于正常用户。



6. 黑厂用户对应的商家更为集中，平均账户余额更少。



7. 黑产用户一个设备可能绑定多个用户 id。



四、 特征工程

4.1basemodel 特征衍生

4.1.1 基本特征衍生

operation	transaction
每个用户总交易次数、操作+交易次数、交易次数/操作次数、是否有操作或者交易记录	
用户操作成功次数、失败次数、成功失败次数差值、成功/失败占比，交易次数占比和成功/失败占比的比值	
用户各操作/交易特征的种类个数(大部分特征都统计一遍)	
用户整体使用设备/环境种类差值/重合度 设备变量：'device1', 'device2', 'device_code', 'mac1', 'device_brand','ip', 'ip_sub' 环境变量： 'geo_code', 'nation', 'city', 'district'	
设备/环境缺失个数平均值、最大值	
用户行为地点发生在中国境内的频次及频率	
用户当前行为是否为常用设备以及是否常在省份城市,统计频次及频率	
用户第一次操作/交易时间，最后一次操作/交易时间	
用户从操作到交易时间间隔小于 100 秒的次数	
设备、环境的热度(mean,max,min,skew,std,sum)，以及两两特征/三个特征之间交叉的热度(sum)	
设备被用户操作的时间间隔	

-	用户的交易总金额
-	用户当前交易行为是否使用常用的资金来源、交易方式、交易商家、账户
-	用户使用的资金来源、交易方式、交易商家、交易金额、账户、转出账户、转入账户、营销活动种类个数

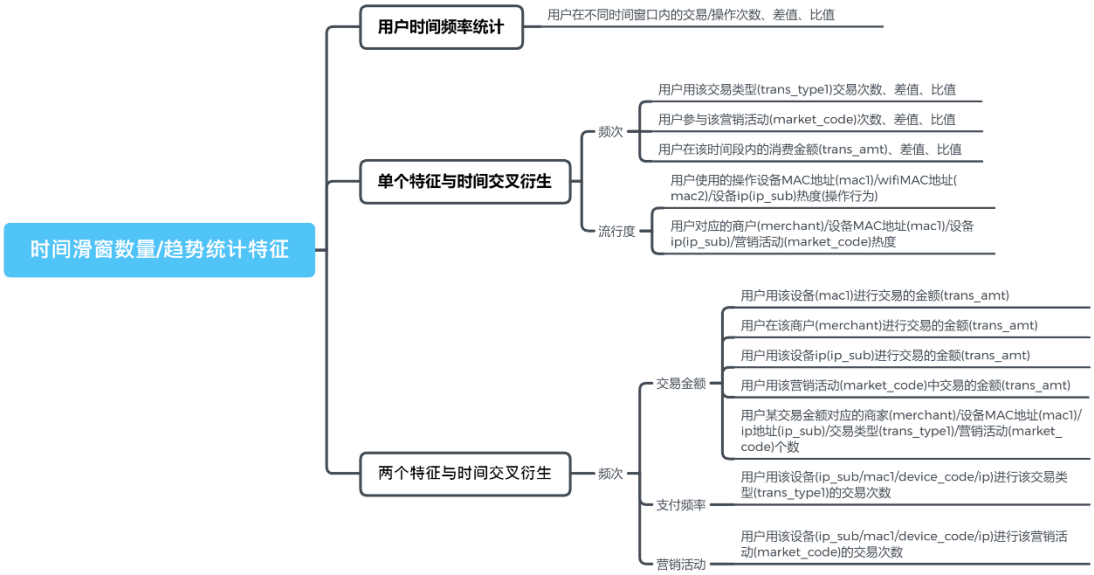
4.1.2 时间滑窗特征统计

我们从时间维度提取借款人在不同时间点的特征以衡量用户动态风险。我们将时间窗口分为每天/每三天/每七天/每小时/半夜/早上/下午/晚上进行时间滑窗特征统计，其中早上时间为[6,12)，下午为(12,18]，晚上为(18,23]，半夜为[0,6)。

1. 时间滑窗数量统计类特征
   - 用户时间频率统计。用户在不同时间窗口内的交易/操作次数。
   - 非时间特征(单个)与时间交叉衍生。
     - ✓ 流行度统计：非时间特征以操作设备地址 mac1 为例，一个用户在时间窗口内可能有多个 mac1 类别，统计用户在时间窗口内对应的 mac1 类别流行度的平均值、最大值、最小值和方差作为新的衍生特征。其中流行度为不同的操作设备地址在时间窗口内对应的操作/交易次数。
     - ✓ 频次统计：以交易类型 trans_type1 为例，统计用户在不同时间窗口，用该交易类别(消费/退款…)进行交易的次数。
   - 非时间特征(两个)与时间交叉衍生。同单个特征与时间交叉的频次统计类似。例如用户在该 ip 地址中用某种交易方式(trans_type1)进行交易的次数。
2. 时间滑窗趋势统计类特征。由于一个人的行为是会动态变化的，因此我们差值和比值来衡量这种动态变化。以交易类型 trans_type1 为例，时间窗口以天计：

差值=当天某交易方式次数-前一天交易方式次数

比值=当天某交易方式次数/前一天交易方式次数

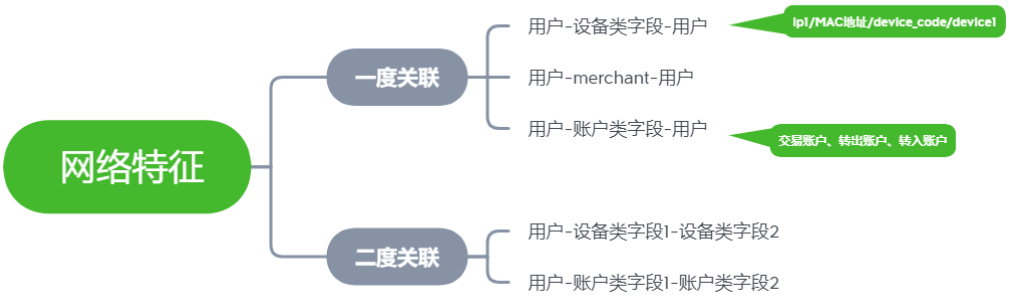


4.2 关系网络模型特征衍生

4.2.1 网络特征提取

- **异构网络构建**

我们把网络中的节点分为两类，一类是用户本身，另一类是用户交易操作涉及到的信息，如账户类字段，通过 graph embedding 构建异构网络。



- **Node2Vec 提取特征**

我们采用 Node2Vec(控制模型更倾向于进行广度还是深度优先搜索的概率)网络表示学习方法来进行特征提取。使用 Node2Vec 网络表示方式来训练表达一阶邻近度，由于网络中的一阶近邻非常稀疏，因此采用 PCA 进行降维处理。

4.2.2 非网络特征

同 4.1 部分的特征衍生方法类似 basemodel 特征，主要围绕设备环境、商户交易，构造用户画像、商户画像和设备环境画像。

五、 特征筛选

basemodel 特征衍生部分，我们构造了一系列的组合特征和时间滑窗特征，加起来近 2200 维特征(不包含关系网络提取特征)，这么多维特征一方面可能会导致维数灾难，另一方面很容易导致过拟合，可以采用降维或特征选择来降低特征维度。这里我们采用基于模型的特征排序的方法来降低特征维度，这种方法有一个好处：模型学习的过程和特征选择的过程同时进行，因此我们采用这种方法。基于 xgboost 来做特征选择，xgboost 模型训练完成后可以输出特征的重要性，据此我们保留了特征重要性大于 0 的特征，最终选出 500 多维特征。

六、 模型训练

LightGBM 作为业界公认优秀的集成树模型，具有非常强的非线性拟合能力，非常适合对我们构造的大量连续型特征进行建模。对比同样强大的 xgboost，它具有更快的速度，方便我们线下测试模型效果。同时，lgb 相较于传统的树模型，加入了 boosting 和 bagging 的集成思想，以及正则项，可有效控制过拟合。通过模型训练及贝叶斯优化得到 basemodel 线下 TPR 为 0.77759，关系网络模型线下 TPR 为 0.64693，而模型融合后线下 TPR 为 0.783571，相对 basemodel 有百分位的提升。

代码：https://github.com/Susanna333/black-industry-recognition-based-on-user-behavior