# 算法说明书

本文总结了本次“甜橙金融杯”大数据建模竞赛的总算法概要，流程如图 1 所示，其中我们把整个解决方案分为数据探索、数据清洗、特征工程、模型训练与算法融合等模块，下文将分别对各模块进行阐述。

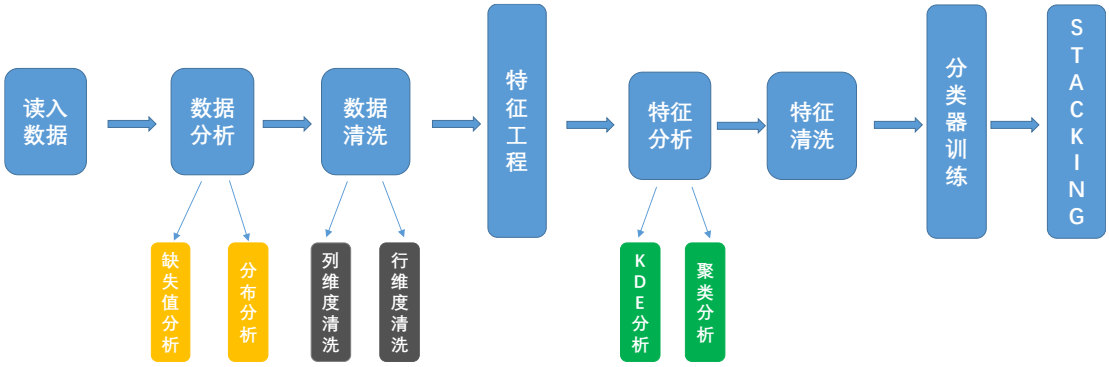


**图 1 黑产鉴别算法流程图**

## 数据探索

复赛测试集来源于真实数据，由于在实际交易与操作行为中，用户与商户的行为发生着不断变化，造成了初赛测试集正负样本分布与复赛测试集存在偏差，并且操作表与交易表中的一些初始特征分布也存在一定偏差，因此我们通过对训练集与测试集的操作、交易表中的数据进行分析后，对训练集进行了的清洗工作。以下将对我们进行的一些数据清洗的操作进行举例说明与简要分析。

**1. 数据分析**

trans 表：用户基本属性：channel（平台），trans_type1, trans_type2（交易类型），amt_src1, amt_src2（资金类型）；op 表：os（操作系统），version（版本号）等条目，其各个种类之间，黑产比率有着明显不同。下面采用 channel 字段说明：

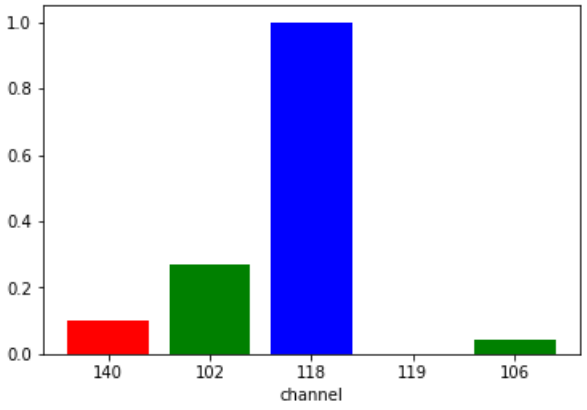


**图 2 各交易基本属性中，各类别黑产率差异很大**

**2. 数据关系**

1) 同时，可以发现，ip1 和 ip2 属性为电脑端与手机端（互补关系）。
2) device_code1, device_code2 数据存在与缺失同步，device_code3 与前者呈互补关系。
3) 电脑端与手机端的 mode 的集合完全不同
4) 不同 mode，翼支付提取的用户信息差异很大，所以 mode 值与用户的信息是否缺失有密切关系。
5) 交易表中，各平台（channel）之间用户信息差异很大，对用户信息的抽取都很不相同。如：118 平台 code1 信息为全，但其他平台很少出现
6) 各个平台的 trans_type 各不相同，完全没有重复。

**3. 复赛数据分析**

复赛数据与初赛 test 数据有着非常明显不同，首先其 channel 结构上，只有 train 集中出现平台 5 个种类中的三种，分别为 102，119 和 140。同时, code1 和 code2， ip2 和 ip2_sub 完全缺失等。

# 数据清洗

基于上述分析，我们对数据进行了如下清洗工作，保证了 train 集和 test 集的数据一致。

**1. 列维度清洗**

删除 op 表中的 ip2 与 ip2_sub 字段，trans 表中的 code1 和 code2 字段

**2. 行维度清洗：**

基于 train 和复赛 test 中某些字段种类差异很大的问题，选取用户基本属性字段：如 mode, channel 等属性，对 train 集进行数据清洗，删除 train 的两个表中，test 集中没有出现过的数据（如 op 表中，有 30 余种 test 集中没有出现过的 mode， 将这些 mode 所对应的各条数据删除）。

**3. 其他处理：**

去重处理：op 表中有很多数据重复，对数据进行去重，只保留重复数据的第一条数据

# 特征工程

由于羊毛党具有很强的团伙属性、时间属性以及设备属性，因此我们的特征侧重于以下几个方面：

围绕着多 UID 同一时间段，使用同一 IP 地址的不同时间细粒度特征群;

围绕着多 UID 同一时间段，使用同一设备的不同时间细粒度特征群;

围绕着多 UID 同一时间段，聚集在同一地区的不同时间细粒度特征群;

围绕着多付款账户等同一时间段，在同一设备上的不同时间细粒度特征群;

围绕着门店交易额，在不同时间细粒度的特征群;

围绕着 IP 在不同时间细粒度的交易额的特征群;

围绕着 UID，门店，在不同时间细粒度的交易金额特征群;

基础的设备、用户、IP、收款属性画像描述特征群;

…

我们通过对羊毛党的动作特征的了解，发现羊毛党不仅仅是分布于用户，即买家。还有的是以商家，即羊毛头子为角色出现的。羊毛党通过 QQ 群、微信群等作为聚集地，通过羊毛头子发布的羊毛信息进行套利。有些是以直接套利为动机，有些则以通过在羊毛头子开设或合作的商家，发起大量的虚假交易进行套利。另外一种，称之为软件党。即通过电脑端或移动端运行的脚本，对抢购类、优惠类进行频繁、多次的交易或是以软件对多 UID 进行操作实现获利目的。最后一类是常见的通过电脑操控的手机集群进行的群体性操作交易。此外，上述几类羊毛党直接互相有交叉。

对于以羊毛头子开设或合作的商家进行套利为代表的羊毛党，该类具有很强的冷启动特征。我们通过对店铺交易额，店铺交易用户，进行监控交易额等的波动性，该特征群的统计特征作为覆盖。

对于以电脑集群或软件操控为代表的羊毛党：通常，此类羊毛党具有很强的动作一致性以及操作频率统一的特点，而这些羊毛党在进行套利的过程中通常只会使用一个或几个真实的用户进行交易操作。因此，我们采用围绕付款信息及设备信息的特征群对此类进行覆盖。并且我们通过对设备、IP 等进行监控，同样实现对于此类的覆盖补充。

对于其他类，除此之外我们使用地理位置上的分辨，羊毛党通常出现地理位置的聚集特

征，因此，我们采用不同时间细粒度的聚集在相同区域的 UID 特征进行特征群提取。

我们对一些基础特征进行画像描述：

用户行为，如该用户在一个小时的交易次数/操作次数的统计特征，用户操作次数，交易次数等特征；

即时特征，如该设备在一小时之内出现了多少不同的用户，当前小时总的交易次数，该设备有多少不同的用户等特征；

IP 属性特征，如该 IP 被多少设备使用，多少用户使用过这个 IP，该 IP 的活跃程度，当前小时内该 IP 出现次数最多的用户，设备等特征；

收款方特征，如当前小时内与收款方有交易的用户数量，设备，收款方的活跃程度，以及收款方之前交易产生的不同用户、设备、地区的数量等特征；

除此之外，上述提及的特征之间也分别对于其他类别的羊毛党进行覆盖补充。

## 模型训练与算法融合

多个模型的集成效果往往优于单模型，本次赛题我们使用了 LightGBM、XGBoost、RF 以及 GBDT 作为 basemodel，采用 Stacking 策略计算每个模型的权重，最后采用 LR 进行模型融合。流程图如下：

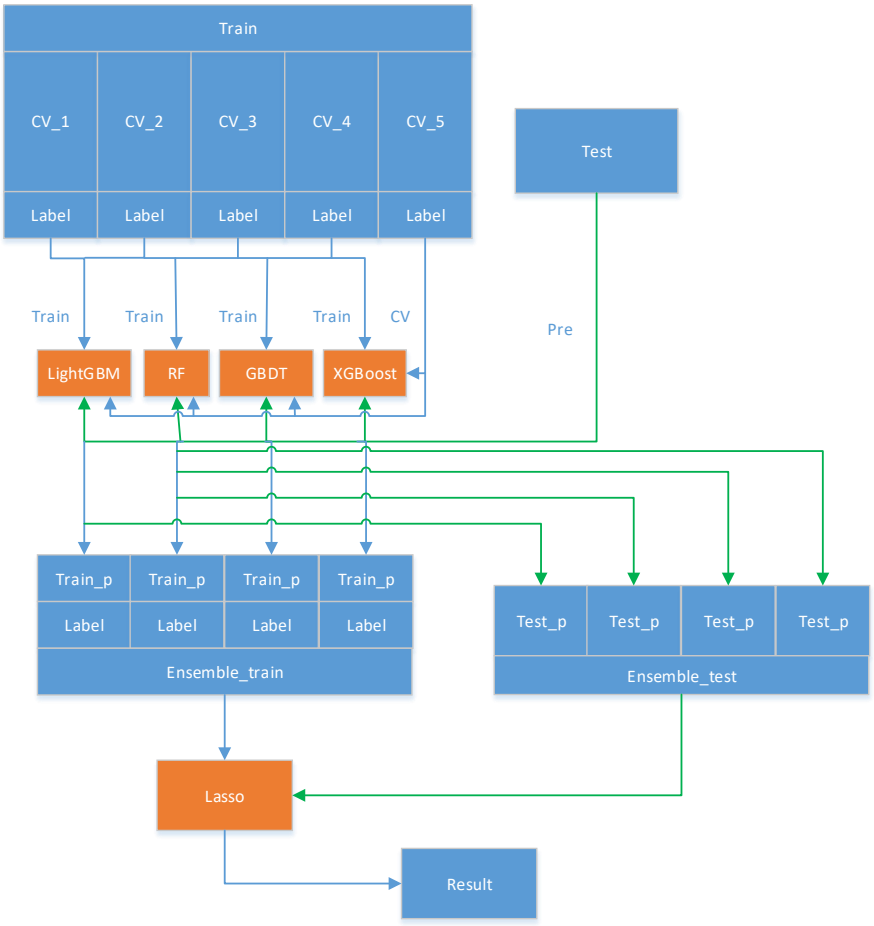


**图 3 Stacking 流程图**

由于初赛排行榜上 TOP 的成绩很高，但是切换到复赛榜单后 TOP 的成绩远远没有之前的那么高。我们判断因为由于时间点的分割长度增加，使得初赛榜上的模型过拟合造成复赛榜单成绩普遍降低。因此我们对模型进行了调整与改进。我们从原来的单模型转化到以上述四个 basemodel 为基础的 Stacking 集成中，其目的是为了让机器学习效果更好，同样也极大程度的缓解了过拟合的风险。

# 不足与展望

我们通过从信用卡套现用户在规避银联风控时的操作中学习到，即规避在交易过程中关闭手机定位与 WiFi 实现对银联风控系统中对于异地扫码套现的风控覆盖。我们在本次赛题中未能实现通过在交易时隐藏地理位置信息与 WiFi 信息来判断用户是否为羊毛党，在这个点上未来也许是一种更有意思的角度。

对于电脑控制的集群操作，我们未能很好的提取出不同时间细粒度的操作过程的序列，并计算彼此之间的操作相似度，来判断该用户是否是一个电脑操控的虚假设备。

特征处理更精细，对于空缺数据的处理，应该更精细一些，并从中挖掘出不可见信息。

由于优惠活动常常是以周为时间分割单位来进行的，未能实现以周为时间分隔的特征统计是本次赛题的一个遗憾。

FoolPepper 团队

2018-12-14