解决方案

比赛：CCF-Human or Robot?
队伍名称：Data-Guerrilla

赛题背景与任务

赛题背景：2016 年第一季度 Facebook 发文称，其 Atlas DSP 平台半年的流量质量测试结果显示，由机器人模拟和黑 IP 等手段导致的非人恶意流量高达 75% ． 仅 2016 上半年，AdMaster 反作弊解决方案认定平均每天能有高达 28% 的作弊流量。低质量虚假流量的问题一直存在，这也是过去十年间数字营销行业一直在博弈的问题。基于 AdMaster 海量监测数据，50%以上的项目均存在作弊嫌疑；不同项目中，作弊流量占广告投放 5%到 95%不等；其中垂直类和网盟类媒体的作弊流量占比最高；PC 端作弊流量比例显著高于移动端和智能电视平台。

赛题任务：需要基于给定的数据，建立一个模型来识别和标记作弊流量，区分正常用户曝光记录与作弊行为记录，并进行标记。

我们将这作为一个典型的二分类问题来考虑。

数据预处理

1.由于硬件条件的制约以及从时间连续性考虑，面对上亿的数据样本，我们选择只用最后一天的样本，约 4900 万条数据。
2.考虑到整型相对于字符型占用空间更少,我们将字符型特征转整型,如 Cookie, idfa, mobile_mac, mobile_openudid, imei, android_id, mobile_type, mobile_app_key, mobile_app_name, placementid, os_type.
3.将 ccf_media_info.csv 中的字段由中文字符串转为整型，然后通过 mediaid 汇总到主表。
4.将 timestamps 转成标准时间，并从中提取 1 分 / 30 分 / 60 分的时间窗信息。

特征工程基本框架

我们将原始特征分成如下三类：
1. 用户 / 设备（U）：cookie，f，born_time，idfa，mobile_mac，mobile_openudid，imei，android_id，mobile_os， mobile，type，mobile_app_key，mobile_app_name，os_type，useragent
2. 广告（A）：camp，play，channel，creativeid，placementid，global_mediaid，media_info
3. 时间（T）：dt，timestamps

用户/设备（U）	cookie,f,born_time,idfa...
广告（A）	camp,play,channel,creativeid...
时间（T）	dt,timestamps

我们将基于这三个分类来设计特征工程的基本框架

1. 单特征：

U 计数特征，例如 cookie_cnt, f_cnt imei_cnt
A 计数特征，例如 camp_cnt, global_mediaid_cnt

2. 多特征：

U&A&T 组合计数特征
例如：f_cookie_cnt, placementid_cookie_cnt, f_hour_cnt, camp_hour_cnt

	U	**A**	**T(Time)**
U	U&U (f_cookie_cnt)	U&A (placementid_cookie_cnt)	U&T (f_hour_cnt)
A			A&T (camp_hour_cnt)

这里补充说明两点，1.组合特征不仅包括两两特征之间的组合还包括多特征之间的组合；2.与时间特征相关的组合特征使用不同的时间窗。

3. 二次统计特征

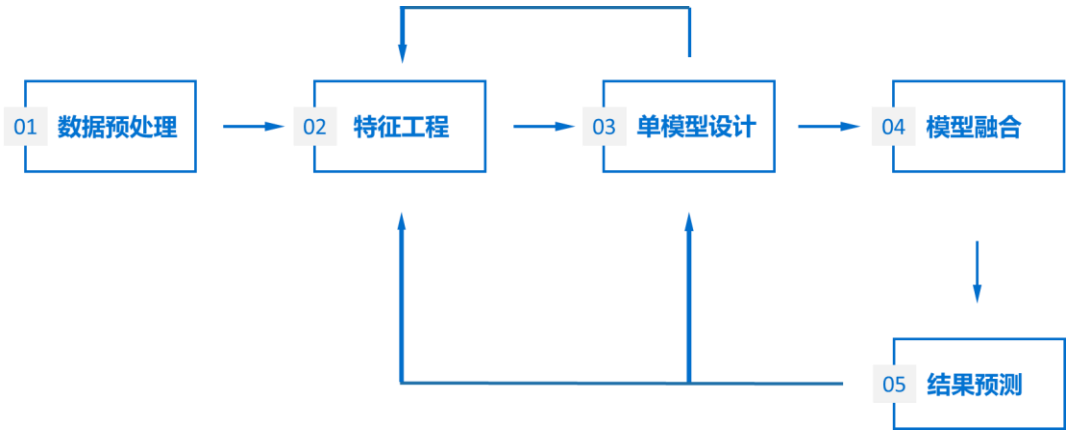
计数特征的均值和方差，例如：f_minute_cnt_mean, f_minute_cnt_std

4. One-Hot 特征

1. 与移动设备相关的特征存在缺失值，将是否为缺失值作为 one-hot 特征
2. 根据 media_info 中的 category，firstType，secondType，tag，4 个特征做 one-hot 特征，例如：firstType_cz, firstType_sp

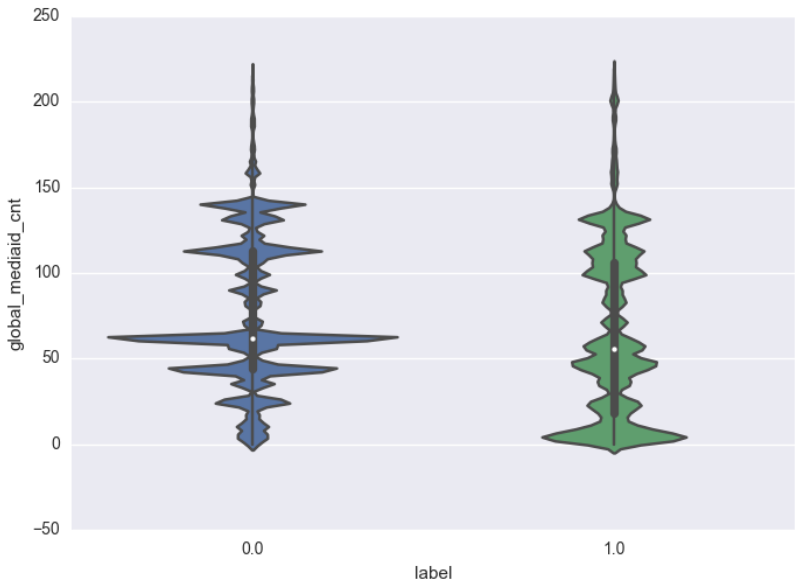
特征工程

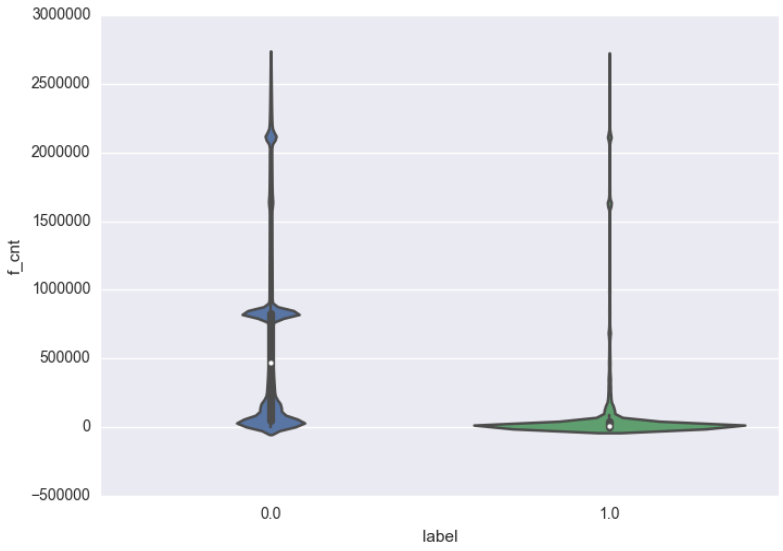
整体流程图

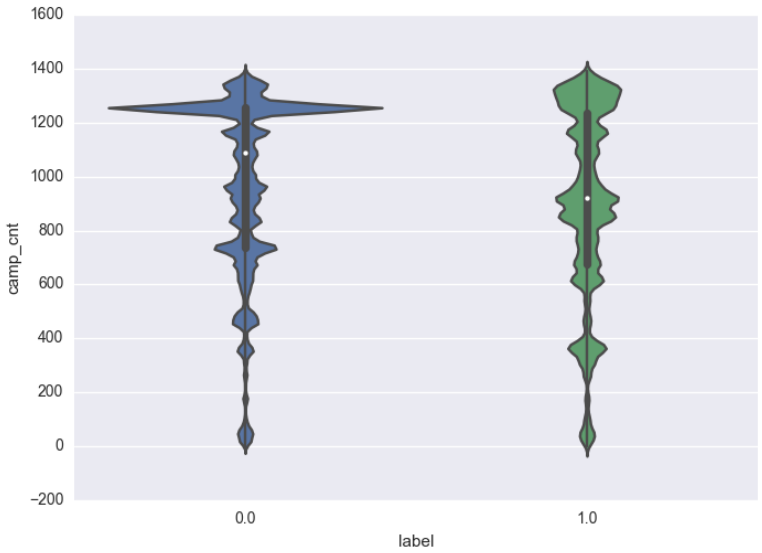


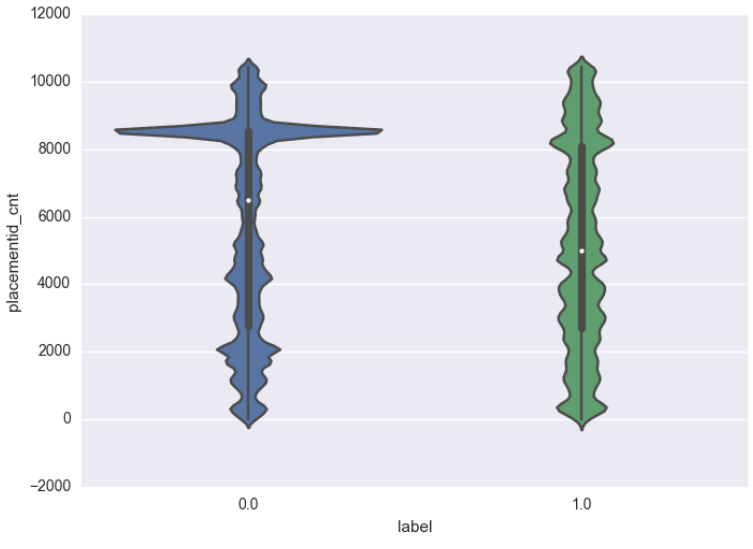
1. 根据小提琴图筛选核心特征

由于硬件原因，无法使用特征工程基本框架中的所有特征进行模型训练。我们绘制小提琴图描述特征与标签的关系，得到 global_meidiaid,f,camp,placement,另外加上描述用户标识特征的 cookie，将这 5 个特征作为核心特征，基于这 5 个核心特征设计其他特征。




3000000
2500000
2000000
1500000
1000000
500000
0
−500000
f_cnt
0.0
1.0
label


1600
1400
1200
1000
800
600
400
200
0
−200
camp_cnt
0.0
1.0
label


12000
10000
8000
6000
4000
2000
0
−2000
placementid_cnt
0.0
1.0
label

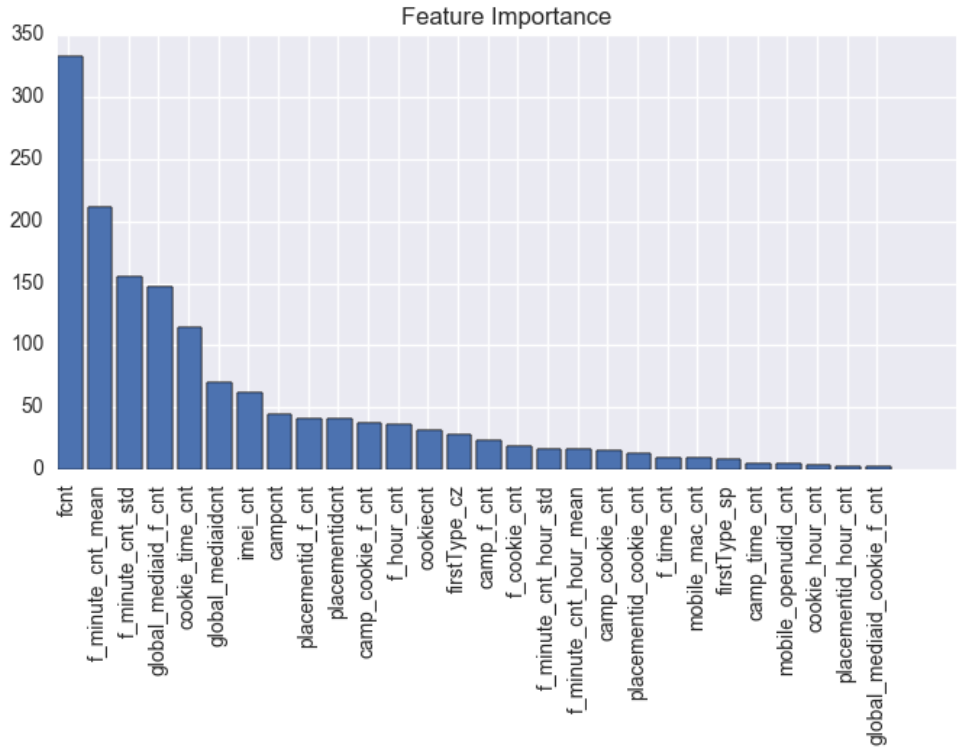
2. 根据模型输出的特征重要性筛选特征

通过每次模型输出的特征重要性，删除特征重要性较低的特征，保留特征重要性较高的特征。主要基于两方面考虑：1.由于数据量较大，在有限硬件条件下，能使用进行训练的特征个数有限，2.删除部分冗余特征。

3. 根据模型在验证集的评测结果进行调整

在验证集上输出多种评测标准的结果（包括 F1，precision，recall，AUC）。通过对不同评测标准综合分析，对特征工程进行调整。比如，前期我们未选用有关移动设备的特征，验证集上出现 precision 低 recall 高的情况，说明我们将一部分 label 为 0 的样本错误预测成了 label 为 1，而在赛题背景中有提及 PC 端的作弊流量显著高于移动端，故我们加入了与移动设备相关的特征，成绩有明显提升。

最终选用的 28 个特征以及其特征重要性



模型介绍

LightGBM

基本介绍：Light Gradient Boosting Machine 为微软最新开源的机器学习项目。其基于决策树算法的快速的、分布式的、高性能的框架，可以被用于分类、回归等机器学习任务中。并且在最近提供了 Python 接口。

特点：1.更快的训练速度更低的内存使用；2.更高的准确率；3.支持并行学习；4.处理大规模数据的能力。

模型融合

考虑到该数据集的规模，我们使用了简化版模型融合，即将 28 个特征根据特征重要性排序，根据重要性权重，每次选择 70%-80%的特征进行模型拟合，最后将三个模型进行线性加权得到最终结果，该模型融合增强了模型的鲁棒性和泛化性。

模型调试过程

下表为模型主要的调试过程

0	Feature	Model	F1
1	U	Xgboost	0.842/0.853
2	U&T	Xgboost	0.857/0.867
3	U&T	LightGBM	0.855
4	U&A&T	LightGBM	0.903/0.906
5	U&A&T	Ensenble	0.916/0.921

改进方向

如果不考虑到数据规模和硬件制约，后续计划的算法改进方向如下，

1. 多时间窗数据集：充分利用训练集的所有数据，以 1 天、2 天、3 天为时间窗切分数据训练模型，得到时间多样性的差异性模型。

2. 多特征群：构建更多简单有效的特征，如排序特征(Rank)，按照时间窗口、广告项目等组合，对样本进行排序，根据以往的经验，这将是有效提升成绩的突破口。同时划分更多的时间窗口，捕捉不同作弊流量的行为习惯，更好的提升效果。

3. 多样性模型组合成的模型：采用 RF, GBDT, Xgboost, LightGBM, LR, SVM 等多种模型进行训练，比较结果的相似性，挑选出结果关联性不强但成绩相近的模型作为 Level1 模型，进而根据 Level1 模型输出的预测值作为 Level2 模型的输入，继续训练若干模型，最后线性加权得到最终结果。