### 移动App广告转化率预估

### 题目描述

计算广告是互联网最重要的商业模式之一，广告投放效果通常通过曝光、点击和转化各环节来衡量，大多数广告系统受广告效果数据回流的限制只能通过曝光或点击作为投放效果的衡量标准开展优化。腾讯社交广告([http://ads.tencent.com](http://ads.tencent.com/" \t "http://algo.tpai.qq.com/home/information/_blank))发挥特有的用户识别和转化跟踪数据能力，帮助广告主跟踪广告投放后的转化效果，基于广告转化数据训练转化率预估模型(pCVR，Predicted Conversion Rate)，在广告排序中引入pCVR因子优化广告投放效果，提升ROI。

本题目以移动App广告为研究对象，预测App广告点击后被激活的概率：pCVR=P(conversion=1 | Ad,User,Context)，即给定广告、用户和上下文情况下广告被点击后发生激活的概率。

### 测试数据

从训练数据时段随后1天(即第31天)的广告日志中按照与训练数据同样的采样方式抽取得到，测试数据文件(test.csv)每行代表一个测试样本，各字段之间由逗号分隔，顺序依次为：“instanceID，-1，clickTime，creativeID，userID，positionID，connectionType，telecomsOperator”。其中，instanceID唯一标识一个样本，-1代表label占位使用，表示待预测。

| 文件名 | 文件大小 | 操作 |
| --- | --- | --- |
| 初赛数据集 | 216M | [下载数据](http://spabigdata-1253211098.file.myqcloud.com/pre.zip" \t "http://algo.tpai.qq.com/home/information/_blank) |

### 训练数据

从腾讯社交广告系统中某一连续两周的日志中按照推广中的App和用户维度随机采样。

每一条训练样本即为一条广告点击日志(点击时间用clickTime表示)，样本label取值0或1，其中0表示点击后没有发生转化，1表示点击后有发生转化，如果label为1，还会提供转化回流时间(conversionTime，定义详见“补充说明”)。给定特征集如下：

| **特征** | **分类** | **描述** |
| --- | --- | --- |
| 广告特征 | 账户ID(*advertiserID*) | 腾讯社交广告的账户结构分为四级：账户——推广计划——广告——素材，账户对应一家特定的广告主。 |
| 推广计划ID(*campaignID*) | 推广计划是广告的集合，类似电脑文件夹功能。广告主可以将推广平台、预算限额、是否匀速投放等条件相同的广告放在同一个推广计划中，方便管理。 |
| 广告ID(*adID*) | 腾讯社交广告管理平台中的广告是指广告主创建的广告创意(或称广告素材)及广告展示相关设置，包含广告的基本信息(广告名称，投放时间等)，广告的推广目标，投放平台，投放的广告规格，所投放的广告创意，广告的受众(即广告的定向设置)，广告出价等信息。单个推广计划下的广告数不设上限。 |
| 素材ID(*creativeID*) | 展示给用户直接看到的广告内容，一条广告下可以有多组素材。 |
| AppID(*appID*) | 广告推广的目标页面链接地址，即点击后想要展示给用户的页面，此处页面特指具体的App。多个推广计划或广告可以同时推广同一个App。 |
| App分类(*appCategory*) | App开发者设定的App类目标签，类目标签有两层，使用3位数字编码，百位数表示一级类目，十位个位数表示二级类目，如“210”表示一级类目编号为2，二级类目编号为10，类目未知或者无法获取时，标记为0。 |
| App平台(*appPlatform*) | App所属操作系统平台，取值为Android，iOS，未知。同一个appID只会属于一个平台。 |
| 用户特征 | 用户ID(*userID*) | 唯一标识一个用户 |
| 年龄(*age*) | 取值范围[0, 80]，其中0表示未知。 |
| 性别(*gender*) | 取值包括男，女，未知。 |
| 学历(*education*) | 用户当前最高学历，不区分在读生和毕业生，取值包括小学，初中，高中，专科，本科，硕士，博士，未知 |
| 婚恋状态(*marriageStatus*) | 用户当前感情状况，取值包括单身，新婚，已婚，未知。 |
| 育儿状态(*haveBaby*) | 用户当前孕育宝宝状态，取值包括孕育中，宝宝0~6个月，宝宝6~12个月，宝宝1~2岁，宝宝2~3岁，育儿但宝宝年龄未知，未知。 |
| 家乡/籍贯(*hometown*) | 用户出生地，取值具体到市级城市，使用二级编码，千位百位数表示省份，十位个位数表示省内城市，如1806表示省份编号为18，城市编号是省内的6号，编号0表示未知。 |
| 常住地(*residence*) | 最近一段时间用户长期居住的地方，取值具体到市级城市，编码方式与家乡相同。 |
| App安装列表(*appInstallList*) | 截止到某一时间点用户全部的App安装列表(appID)，已过滤高频和低频App。 |
| App安装流水 | 最近一段时间内用户安装App行为流水，包括appID，行为发生时间(installTime)和app类别(appCategory)，已过滤高频和低频App。 |
| 注：2~8基于用户个人注册资料和算法自动修正得到，9~10基于用户行为日志统计得到。 | |
| 上下文特征 | 广告位ID(*positionID*) | 广告曝光的具体位置，如QQ空间Feeds广告位。 |
| 站点集合ID(*sitesetID*) | 多个广告位的聚合，如QQ空间 |
| 广告位类型(*positionType*) | 对于某些站点，人工定义的一套广告位规格分类，如Banner广告位。 |
| 联网方式(*connectionType*) | 移动设备当前使用的联网方式，取值包括2G，3G，4G，WIFI，未知 |
| 运营商(*telecomsOperator*) | 移动设备当前使用的运营商，取值包括中国移动，中国联通，中国电信，未知 |

特别的，出于数据安全的考虑，对于userID，appID，特征，以及时间字段，我们不提供原始数据，按照如下方式加密处理：

| **字段** | **解释** |
| --- | --- |
| userID | 对每个用户ID，随机生成一个不重复的userID(假设用户数为100w，将所有用户随机打散排列，将其序号作为userID，取值范围是[1, 100w])，使用加密后的userID作为用户标识。 |
| appID | 参考用户ID的加密方式，生成加密后的appID。 |
| 特征 | 参考用户ID的加密方式，生成加密后的特征ID。特别的，我们使用“0”表示特征值未知。 |
| 时间 | 包括 clickTime，conversionTime，installTime，格式均为DDHHMM，其中DD代表第几天，HH代表小时，MM代表分钟。 |

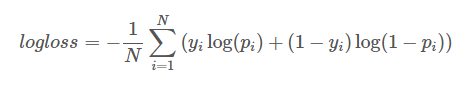
**训练数据文件(train.csv)**

每行代表一个训练样本，各字段之间由逗号分隔，顺序依次为：“label，clickTime，conversionTime，creativeID，userID，positionID，connectionType，telecomsOperator”。当label=0时，conversionTime字段为空字符串。特别的，训练数据时间范围为第17天0点到第31天0点(定义详见下面的“补充说明”)。为了节省存储空间，用户、App、广告和广告位相关信息以独立文件提供(训练数据和测试数据共用)，具体如下：

| **文件类型** | **描述** |
| --- | --- |
| 用户基础特征文件(*user.csv*) | 每行代表一个用户，各字段之间由逗号分隔，顺序依次为：“userID，age，gender，education，marriageStatus，haveBaby，hometown，residence”。 |
| 用户App安装列表文件(*user\_installedapps.csv*) | 每行代表一个用户安装的单个App，各字段之间由逗号分隔，顺序依次为：“userID，appID”。特别的，我们提供了截止到第1天0点用户全部的App安装列表。 |
| 用户App安装流水文件(*user\_app\_actions.csv*) | 每行代表一个用户的单个App操作流水，各字段之间由逗号分隔，顺序依次为：“userID，installTime，appID”。特别的，我们提供了训练数据开始时间之前16天开始连续30天的操作流水，即第1天0点到第31天0点。 |
| App特征文件(*app\_categories.csv*) | 每行代表一个App，各字段之间由逗号分隔，顺序依次为：“appID，appCategory”。 |
| 广告特征文件(*ad.csv*) | 每行描述一条广告素材，各字段之间由逗号分隔，顺序依次为“creativeID，adID，camgaignID，advertiserID，appID，appPlatform”。 |
| 广告位特征文件(*position.csv*) | 每行描述一个广告位，各字段之间由逗号分隔，顺序依次为：“positionID，sitesetID，positionType”。 |
| 注：若字段取值为0或空字符串均代表未知。 | |

### 评估方式

通过Logarithmic Loss评估(越小越好)，公式如下：



其中，N是测试样本总数，y**i**是二值变量，取值0或1，表示第i个样本的label，p**i**为模型预测第i个样本 label为1的概率。

**示例代码如下(Python语言)：**

import scipy as spdef logloss(act, pred):

epsilon = 1e-15

pred = sp.maximum(epsilon, pred)

pred = sp.minimum(1-epsilon, pred)

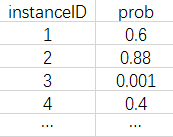
ll = sum(act\*sp.log(pred) + sp.subtract(1,act)\*sp.log(sp.subtract(1,pred)))

ll = ll \* -1.0/len(act)

return ll

### 提交方式

模型预估结果以zip压缩文件方式提交，内部文件名是submission.csv。每行代表一个测试样本，第一行为header，可以记录本文件相关关键信息，评测时会忽略，从第二行开始各字段之间由逗号分隔，顺序依次为：“instanceID, prob”，其中，instanceID唯一标识一个测试样本，必须升序排列，prob为模型预估的广告转化概率。示例如下：



### 补充说明

**1.什么是转化回流时间？**

App 的激活定义为用户下载后启动了该App，即发生激活行为。从用户点击广告到广告系统得知用户激活了App(如果有)，通常会有较长的时间间隔，主要由以下两方面原因导致：

1) 用户可能在下载之后过了很久才启动App；

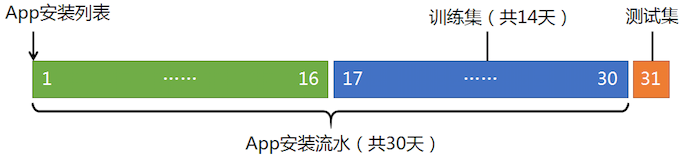
2) 用户启动App的行为需要广告主上报回传给广告系统，通常会有一定的延时。

这里回流时间表示了广告主把App激活数据上报给广告系统的时间，回流时间超过5天的数据会被系统忽略。

值得注意的是，本次竞赛的训练数据提供的截止第31天0点的广告日志，因此，对于最后几天的训练数据，某些label=0并不够准确，可能广告系统会在第31天之后得知label实际上为1。

**2.原始时间信息抹掉了，关键数据源时间关系是什么？**

为了直观起见，如下图所示：



**3.初赛和复赛数据是如何选取的？**

从复赛数据中根据appID随机采样得到初赛数据。

**4.用户个人资料和App安装数据来源？**

数据来自腾讯业务产品历史上一段时间真实数据集合的小量随机采样，关键字段均经过加密处理，只提供ID字段表示，不提供具体含义。数据处理上为比赛用途单独设计，不考虑、也不作为其他用途。