一、大赛介绍

作为国内领先的大数据营销平台，全新升级的腾讯广告，以更强大的全景连接、更全链的数字智慧、更友好的人本体验等三大核心能力，构建品牌与用户的智慧连接，助力广告主高效实现商业增长。而复杂的社交场景，多样的广告形态，以及庞大的人群数据，给实现这一目标带来了不小的挑战。为攻克这些挑战，腾讯广告也在不断地寻找更为优秀的数据挖掘方式和机器学习算法。

本次算法大赛的题目是源于腾讯广告业务中一个面向广告主服务的真实业务产品 ——广告曝光预估。广告曝光预估的目的是在广告主创建新广告和修改广告设置时，为广告主提供未来的广告曝光效果参考。通过这个预估参考，广告主能避免盲目的优化尝试，有效缩短广告的优化周期，降低试错成本，使广告效果尽快达到广告主的预期范围。比赛中使用的数据经过脱敏处理，通过本次大赛，我们旨在挑选出更为优秀的曝光预估算法以及遴选出杰出的社交广告算法达人。

二、大赛流程

本次大赛分为初赛、复赛和答辩三个环节。

初赛：2019 年 4 月 18 日 ~ 5 月 23 日

- 初赛分为初赛 A 和初赛 B 两个阶段。初赛 A 阶段为北京时间 4 月 18 日 12:00:00 - 5 月 16 日 11:59:59，初赛 B 阶段为北京时间 5 月 16 日 12:00:00 - 5 月 23 日 11:59:59。AB 阶段的训练集相同，测试集不同，每天（北京时间中午 12 点开始的 24 小时内）限提交 3 次结果，系统将实时计算得到此次结果的评分，并在个人主页上显示。最终成绩排行榜将以初赛 B 阶段各参赛队伍的历史最好成绩进行排名。

- 初赛开始后，系统每天进行一次排名。每天基于北京时间 12:00 前提交的结果，按照参赛队伍的历史最优成绩从高到低依次排序，并于北京时间 15:00 更新排行榜，此排行榜成绩不作最终排名计算。

- 初赛最终结果提交时间为北京时间 5 月 23 日 12:00，并于北京时间 15:00 更新排行榜。结束时，成绩排名前 20%（最多不超过 200 支，以大赛官网解释为准）的队伍进入复赛。

复赛：2019 年 5 月 24 日 ~ 6 月 14 日

- 复赛将会更换数据集，分为复赛 A 和复赛 B 两个阶段。复赛 A 阶段为北京时间 5 月 24 日 12:00:00 - 6 月 7 日 11:59:59，复赛 B 阶段为北京时间 6 月 7 日 12:00:00 - 6 月 14 日 11:59:59。AB 阶段的训练集相同，测试集不同，每天（北京时间中午 12 点开始的 24 小时内）限提交 3 次结果，系统将实时计算得到此次结果的评分，并在个人主页上显示，最终成绩排行榜将以复赛 B 阶段各参赛队的历史最好成绩进行排名。

- 复赛开始后，系统每天进行一次排名。每天基于北京时间 12:00 前提交的结果，按照选手的历史最优成绩从高到低依次排序，于 15:00 更新排行榜，此排行榜成绩不作最终排名计算。

- 复赛最终结果提交时间为北京时间 6 月 14 日 12:00 ，并于 15:00 更新排行榜。结束时，主办方将按照成绩排名顺序向参赛队伍发起代码提交请求，对代码进行检查和复现后，确认前 10 名（含并列）的队伍进入答辩。

答辩及颁奖：7 月

- 现场根据复赛成绩、代码和答辩成绩，评出前 20 支队伍的最终排名并颁奖，未到现场视为放弃比赛。

- 参与答辩的队伍需提前准备答辩材料（含答辩 PPT，算法详细说明，团队分工，以及代码，现场核验有效证件） 。

三、赛题介绍

腾讯效果广告采用的是 GSP（Generalized Second-Price）竞价机制，广告的实际曝光取决于广告的流量覆盖大小和在竞争广告中的相对竞争力水平。其中广告的流量覆盖取决于广告的人群定向（匹配对应特征的用户数量）、广告素材尺寸（匹配的广告位）以及投放时段、预算等设置项。而影响广告竞争力的主要有出价、广告质量等因素（如 pctr/pcvr 等），以及对用户体验的控制策略。通常来说，基本竞争力可以用千次曝光收益 ecpm = 1000 * cpc_bid * pctr = 1000 * cpa_bid * pctr * pcvr （cpc, cpa 分别代表按点击付费模式和按转化付费模式)。综上，其中前者决定广告能参与竞争的次数以及竞争对象，后者决定在每次竞争中的胜出概率。二者最终决定广告每天的曝光量。

本次竞赛将提供历史 n 天的曝光广告的数据（特定流量上采样），包括对应每次曝光的流量特征（用户属性和广告位等时空信息）以及曝光广告的设置和竞争力分数；测试集是新的一批广告设置（有完全新的广告 id，也有老的广告 id 修改了设置），要求预估这批广告的日曝光 。（出于业务数据安全保证的考虑，所有数据均为脱敏处理后的数据。）

四、数据说明

4.1 训练数据

如之前所述，训练数据包含广告 n 天的曝光历史数据、曝光用户的属性数据，广告设置和操作数据，按具体的维度可以分为如下几部分提供。

1）**历史曝光日志数据文件**

为避免数据量过大，历史曝光数据选择了在用户维度（uv）上按 512 分之一进行均匀采样。各字段使用制表符(\t)分隔，每列的具体含义如下：

- **广告请求 id**：唯一标识每次请求（每个请求对应一个用户某一时刻，可能多个

广告位）

- **广告请求时间**：该字段为时间戳，即 1970 纪元后经过的浮点秒数
- **广告位 id**：加密后无业务含义，只区分不同广告位，每个广告位只能曝光特定素材尺寸的广告
- **用户 id**（即看广告的人）：加密后无业务含义，只区分不同用户，可和后面的用户特征数据中 id 相关联
- **曝光广告 id**：加密后无业务含义，只区分不同广告，可以和广告特征文件中的广告 id 关联
- **曝光广告素材尺寸**：枚举型取值，不同广告位对素材的尺寸要求不同，同一个广告位可能适配多个不同尺寸的素材
- **曝光广告出价 bid**：这里只记录 cpc 出价，非 cpc 广告此处记录折算后的 cpc 价格
- **曝光广告 pctr**：预估的 pctr，和 bid 相乘得到 basic_ecpm
- **曝光广告 quality_ecpm** 将广告质量和用户体验等因素折算成 ecpm 的分数，主要影响因素有 pctr/pcvr/窄定向等
- **曝光广告 totalEcpm** :广告排序的分数依据 ,由 basic_ecpm 和 quality_ecpm 相加得到

2）**用户特征属性文件**

注：用户特征字段未知的均使用 0 表示。每列特征取值都用加密后的 id 表示，均为随机映射。不同列的 id 取值区间会重复。各字段使用制表符(\t)分隔，每列的具体业务含义如下：

- **用户 id**：此处和上面曝光日志文件中的用户 id 关联

- **年龄（Age）**：每个取值随机映射为[1-N]的唯一 id
- **性别(Gender)**：男/女
- **地域(area)**：每个省/市用唯一 id 标识，可能多标签，使用逗号分隔不同 id
- **婚恋状态（Status）**：单身/已婚等状态，可能去多值，使用逗号分隔
- **学历(Education)**：博士/硕士/本科/高中/初中/小学
- **消费能力（ConsuptionAbility）**：高/低
- **设备（device）**：IOS/Android，不区分版本号
- **工作状态（work）**：在校大学生/商旅人士/政府公职人员/科研教育者/IT 互联网工作者/医护工作者，可能取多值，逗号分隔
- **连接类型(ConnectionType)**：无线/2G/3G/4G
- **行为兴趣(behavior)**：每个兴趣点一个 id，可多值，逗号分隔

3）广告数据文件

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）以及广告出价等信息。这里将广告的数据分为两部分，一部分是广告静态数据（如所属账户等），另一部分是动态数据，即修改操作数据，主要包括广告状态、出价、定向等可修改的设置，该部分数据对应一个时间戳，多次修改的同一条广告会对应多条记录。

a. **广告静态数据**

该类广告属性一般从广告创建后无法修改。所有 id 类数据均为加密后随机映射。各列用制表符分隔，含义如下：

- **广告 id**：和曝光日志中的广告 id 相关联
- **创建时间**：广告创建时的时间戳
- **广告账户 id**：广告所在账户的唯一标识，账户结构分为四级：账户——推广计划——广告——素材
- **商品 id**：广告推广目标的唯一标识，若推广目标是落地页，则该字段为空
- **商品类型**：广告推广目标的类型，枚举型
- **广告行业 id**：广告所属的行业类别标识
- **素材尺寸**：不同广告位对素材的尺寸要求不同，同一个广告可能有多个不同尺寸的素材，用逗号分隔

b．广告操作数据

记录广告操作的流水数据，以及操作后的属性值。各列使用制表符分隔，含义如下：

- **广告 id**（同上）
- **创建/修改时间**：即广告创建或者修改设置的时间
- **操作类型**：1-修改，2-新建
- **修改字段**：1-广告状态，2-出价，3-人群定向，4-广告时段设置
- **操作后的字段值：**
  - **广告状态取值**：1- 正常，0-失效
  - **出价**：整数（单位分）
  - **投放时段**：字符串。包含 7 个 64 位无符号整型数字（逗号分隔），每个整数分别代表周一到周日的投放时段。该整数转为 2 进制后从低到高 48 位 bit 代表全天各时段（半小时为一时间窗口）是否投放，1-投放，0-不投。举例说明，17179865088= 11111111111111111111110000000000，

代表投放时段为 6：00-17：00；

281474976710655=

111111111111111111111111111111111111111111111111 ,代表全天投放。

- **人群定向**：字符串。格式如下：

feature_name1:feature_value1,feature_value2|feature_name2:feature_value3,feature_value4|... 此处 feature_name 取值同用户属性文件中的各列属性名，feature_value 取值 id 同用户属性文件中的定义，不同 feature 用“|”分隔，不同 feature 取值用逗号分隔。广告通过人群定向的设置来召回对应的用户请求，对应的人群规则：不同 feature_name 是求交集，同一 featurename 下不同的 value 求并集，未定义的 feature_name 则表示该维度不限。举例如：定向设置为 age:51,62,73,84|gender:1|area:1,3,5；则表示该广告能被“**（年龄 id 为 51 或 62 或 73 或 84）且（性别取值为 1）且（地域取值为 1 或 3 或 5）**”的用户召回（即在这些用户上有曝光机会）。

4.2 测试数据

第 n+1 天的待预估广告以及对应的设置。各列由制表符(\t)分隔， 含义说明如下：

- 样本 id
- 广告 id
- 创建时间
- 素材尺寸
- 广告行业 id

- 商品类型
- 商品 id
- 广告账户 id
- 投放时段
- 人群定向
- 出价（单位分）

以上各字段的格式均和训练数据中的广告数据格式一致。

由于要评估出价相关性，故测试数据中一条广告 id 会对应多条不同出价的样本。

为简化问题，测试数据也要求预估指定广告样本在下一个自然日的曝光量。（此处的曝光量是和训练数据一样 uv 采样后的结果）。为了更准确的评估，此处的测试广告已经剔除掉下一个自然日期间再次修改的广告（如暂停、修改出价、定向等），即认为在预估周期内该广告设置保持不变。

五、评估方式

1）准确性指标

$$\mathbf{SMAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{(F_t + A_t)/2}$$

其中，Ft 为预估的广告曝光值，At 为真实的曝光值。易知，SMAPE 越小越好。

注：测试数据中某个广告只有一个样本的出价是对应线上真实出价，对该样本的预估结果对比线上真实曝光（采样后）来评估准确性。

2）出价单调相关性指标

由于竞价机制的特性，在广告其他特征不变的前提下，随着出价的提升，预估曝光值也单调提升才符合业务直觉。所以，定义出价单调相关性评估指标如下：

对待预估广告 ad，除出价 bid 外其他设置不变，任意变化 n 个 bid 取值，得到对应的 n 个曝光预估值，计算如下该广告出价单调性得分如下：

$$\text{score} = \frac{1}{n}\sum_{k=1}^{n}\frac{(imp_0 - imp_k)(bid_0 - bid_k)}{|(imp_0 - imp_k)(bid_0 - bid_k)|}$$

对所有待预估广告，计算单调性得分均值如下（易知，该值越大越好）

$$\text{MonoScore} = \frac{1}{m}\sum_{t=1}^{m} score_t$$

3）最终得分 Totalscore

上述两个指标值域和趋势都不同，为了比赛评分简便，会将上述两个指标各自归一化后再加权求和得到一个最终得分。形式如下：

$$\text{TotalScore} = w_1 * \left(1 - \frac{SMPAE}{2}\right) + w_2 * \frac{MonoScore + 1}{2}$$

六、提交方式

参赛者提交结果为一个 submission.csv 文件，编码采用无 BOM 的 UTF-8，格式如下：样本 id（和提供的测试样本对应），预估日曝光（保留整数）。用逗号分隔字段，无空格，无表头。

示例如下：

1，230

2，35

七、建议使用的计算资源

单机运行内存不超过 128G，CPU 不超过 24 核。