1. **数据说明**

**1.1 训练数据**

和初赛相比，训练数据做了部分调整。最大的变化是我们在日志数据加入了竞争广告队列的信息，即除了曝光的广告外，和它一起竞争的其他未曝光广告的信息。另一方面我们的预估对象也做了调整，一是除了传统cpc广告还引入了cpm广告和ocpa广告的预估，二是我们提供了测试当天的用户请求队列，以及待预估广告被哪些请求召回。这样降低了部分业务门槛，不用像之前通过定向和用户匹配、广告位和素材匹配等规则去估算广告的覆盖范围，更聚焦于对广告竞争力本身的预估。

具体文件说明如下：

**历史日志数据文件（track\_log.zip, 解压后里面是分天的每个文件）**

和初赛一样的uv采样，但为了控制数据规模，广告位做了更小范围的筛选。 各字段使用制表符(\t)分隔，每列的具体含义如下（列顺序和下面顺序相同）：

1. **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）
2. **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
3. **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
4. **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
5. **竞价广告信息（不同广告间用分号分隔，同一广告各属性逗号分隔）**
6. **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
7. **竞价bid：**这里按计费类型不同对值有差异，cpc类型广告为cpc出价，对于cpm广告是cpm出价。对于ocpa广告，可以近似认为竞价cpc\_bid = target\_cpa(目标出价） \* pcvr（预估转化率）。由此可知同一广告不同请求时参与竞争排序的cpc出价时动态调整的。
8. **pctr：**预估的pctr， 和cpc\_bid相乘得到basic\_ecpm
9. **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。
10. **totalEcpm**: 广告排序的分数依据，由basic\_ecpm和quality\_ecpm相加得到。
11. **是否被策略过滤： 0/1** ,一般有多样性、新鲜度等策略
12. **是否曝光：0/1**

**用户特征属性文件**， **user\_data.zip**

该文件和初赛格式一致，不再详细赘述

* **用户id：** 此处和上面曝光日志文件中的用户id关联。
* **年龄（Age）：**随机映射id。
* **性别(Gender)：** 男/女/未知
* **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，可能取多值，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

**广告数据文件（**和初赛一样分静态和操作两个文件。）

1. **广告静态数据（map\_ad\_static.out)**

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（final\_map\_bid\_opt.out)**

因为测试集中提供了广告实际触发的请求和对应的竞争者队列，所以复赛不用再关注广告定向、时段的变化情况，所以只提供了出价相关的操作。格式稍有修改。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
* **新的目标转化类型：** 0为传统竞价广告，大于0代表ocpa广告的不同转化目标（如下载激活还是销售等），一般因推广目标不同而差异
* **操作后新的计费类型：** 1- cpm计费按前次曝光计费， 2-cpc，按点击计费。
* **操作后新的出价值**： 整数（单位分）。Ocpa广告代表目标转化成本（target\_cpa）， 传统竞价广告即为每次参与的竞价值。

**1.2 测试数据**

**本次测试集分为三部分。**

1. **测试当日的请求日志和竞价队列 ， atest\_tracklog.zip**
   1. （格式和训练集相同，差别是隐去了是否曝光的标识，以及待预估广告的信息）
2. **待预估广告对应的请求集合 final\_select\_test\_request.out**
   1. 第一列广告id
   2. 第二列是请求集合（每个请求用请求id和广告位id标识，逗号分隔 ）不同请求使用|分隔
3. **广告的出价信息文件，test\_sample\_bid.out**

各列由制表符(\t)分隔， 含义说明如下：

* 样本id
* 广告id
* 目标转化类型（同操作文件中字段含义）
* 计费类型（同操作文件中字段含义）
* 出价（单位分）

和初赛一样，为了评估模型的出价单调性，每个广告设置了不同出价生成多个样本。预估目标是不同出价设置下的广告对应改日触发请求下的总曝光量。