FAQ（一）

【频繁度较高的问题】

- **跑模型的内存下限是多少？**

  A: 这个视选用的模型而异，没有特别的下限。

- **为什么广告操作数据上时间出现了 201902300000，2 月有 30 日吗？**

  A: 2 月没有 30 日，脏数据需要自己清洗。

- **婚恋状态（Status）：单身/已婚等状态，可能去多值，为什么婚恋状态的字段会有 19 种取值，而且存在组合呢？除了单身、已婚，这一类，是不是还加了别的状态类标签？**

  A: 会有多值，不用特意关注 id 的业务含义。

- **广告操作字段中创建时间为 0 表示什么意思？**

  A: 代表缺失， 这个字段可以从 ad_static_feature.dat 获得。

- **缺失字段是否是手册给的标的物类型 ？“同一个广告可能有多个不同尺寸的素材，用逗号分隔”，广告静态数据里面的素材尺寸字段是否正常，为什么数据里面在第六个 column 出现逗号分隔符？第 230021 记录的商品 id 为什么会是个列表呢？**

  A: 不同字段都可能缺失。逗号分隔都表示多值。除手册里说明合法多值的字段，其他认为是脏数据。

- **请问测试集预估 n+1 天的曝光，这里的 n+1 是指的历史曝光数据所有日志数据的后一天吗？**

  A: 新创建广告预估新建后一天的，其他预估 3 月 21 日的。（**对于其他预估日期的疑问：**除新建广告外，测试集给的是 20 日的广告状态，所以预估 3 月 21 日的曝光。新建广告包括有 20 日到 24 日新建的，所以是预估新建后次日的。）

【其他提到的问题】

理解类：

- **特定流量上采样怎么理解？**

  A： 腾讯效果广告可以被广告主选择在不同的媒体流量上（例如微信朋友圈、公众号、浏览器、手Q、qq 空间等）播放；为了简化问题，我们只挑选了特定的媒体流量上曝光的广告来预估。采样是为了降低数据规模，对历史曝光明细在用户维度进行了均匀采样。后面预估的也是采样后的结果。

- **参赛手册里面，评价指标中的 Ft，广告曝光值，是指 ECMP 还是什么？**

  A: 这里就是广告的曝光量， 不是 ECPM。

- **pctr 是什么的点击率？**

  A: pctr 是排序过程中预估的广告点击率，日志中的该列数据统一做了放大处理。

  pctr 为什么会大于 1，或者说 pctr 的原本数量级是多少？

  A: 整体放大了

- **totalECPM 就是曝光值吧？是不是根据这些特征预测 totalEcpm？**

  A: totalEcpm 是每次广告请求时，处于候选竞争队列的广告的最终得分。一般而言，totalEcpm 最高的广告会获得此次曝光，曝光日志中每条记录就是一次曝光。

- $$\text{score} = \frac{1}{n}\sum_{k=1}^{n}\frac{(imp_0 - imp_k)(bid_0 - bid_k)}{|(imp_0 - imp_k)(bid_0 - bid_k)|}$$ **这个公式的 imp，bid 是指什么？**

  A: imp 是预估曝光， bid 是出价。

- **请问可以说明一下相关性指标吗？怎么做那个相关性指标分析呢？**

  A: 简单说， 题目期望的相关性是指同一广告在其他设置不变的情况下，出价越高，预估值越高。

- **“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”这里面的窄定向是什么意思？**

  A: 广告定向的宽窄决定了覆盖人群的多少， 窄定向就是覆盖人群小的意思。

- **对于一次请求的候选广告队列来说，totalecpm 最高的广告一定胜出还是有最大概率胜出？**

  A： 大概率。实际线上曝光的结果会综合多样性、新鲜度等多种用户策略。但因为赛题数据的大小限制初赛没有提供全面的信息。

关键词：测试集

- **测试集中，用来统计曝光量依赖的用户是不是都包含在用户属性文件中？**
  A: 是的。

- **测试集的人群定向和投放时段，是创建该广告时的还是最后一次修改该广告时的？**
  A： 最后一次修改广告后。比如预估 2019 年 4.18 日的广告日曝光，给的广告设置是 4 月 17 日最后一次修改后的，且 4 月 18 日没有修改广告。

- **测试集中的广告，都是在第 N 天修改过或者创建过的广告吗？假设预测的是第 N+1 天？**
  A: 不完全是。有 N 日后新建的，也有未修改设置的广告。对于新建的广告预估新建日后一天的曝光， 其他是预估第 N+1 日的曝光。

- **测试数据中的 timestamp 转成时间格式后，有 17 年和 18 的数据，这个情况是正常的嘛？**
  A: 正常， 这里会有之前创建的老广告。

- **请问只有和测试集中同类型的广告才能得到全部的完整广告设置 ，怎么理解？这里全部完整的广告设置具体指哪些，能说下吗？**
  A： 自己理解吧。不是所有的广告都适合作为训练样本。广告设置可以参考测试集的广告设置字段。

- **广告的动态数据中的定向人群和投放时间、广告状态等信息，在测试集中都是直接给出的，可是训练集是只有对这个属性修改才会有对应的值。是让我们自行去统计的意思吗？**
  A: 是需要自己去构造

- **有的广告被修改为失效后，仍然可以曝光，也出现在测试集中，所以这个失效是什么意思呢？是广告下架了？还是其他意思？**
  A: 测试集里的广告是生效的

关键词：曝光数据

- **曝光量是曝光次数不是曝光的用户数是吧？例如一个用户一天访问了 20 次这样曝光数算 20 不能算 1 是吧？**
  A: 是广告曝光（被用户看到）的次数，被一个用户看了 20 次是 20 个曝光。

- **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。是指用户只有原来的 512 分之一吗？**
  A： 是的。

  请问你们统计测试集结果的曝光数是除以 512 取整吗。ps 问的是真实结果是否是整数，不是我提交的结果？
  A： 真实曝光肯定是整数。作为比赛题目，提供的是抽样后日志，也是预估抽样后的曝光结果。

- **请问广告的实际曝光量是一个定性的指标还是个定量的指标？它是由最终的点击次数来衡量吗？**

  A: 定量指标。 广告曝光或者称之为广告展现、播放， 只有曝光的广告才能被用户看见才有 可能点击。

- **日曝光是要我们自己根据“历史曝光日志数据文件”，统计平均每天“广告 id”的出现次数吗？ 其中日期要根据广告请求时间转化？**

  A: 是的

- **在历史曝光日志记录里的一条记录是一次曝光？，还是一次广告设置对于一个 uid 的一天的曝光效果？**

  A: 一次曝光。

- **曝光日志里面是不是也存在空数据啊**

  A：这里没有

- **在曝光数据里面 同一个 id 的广告在一天的时间里面有多个出价， 这个怎么理解？可以理解是广告主一天内多次修改出价了吗？**

  A: 测试集里的广告都是传统 cpc 出价，即不修改的情况下每次请求的价格都相同。曝光日志里面的广告包括全部类型，包括智能出价的广告，即每次请求都可能不同出价。还有 cpm 出价的广告，日志里是折算的 cpc 出价。

- **唯一请求 id 有重复，在计算曝光时需要去重吗？**

  A: 一个请求可能在不同广告位曝光多个不同广告， 如果是同一请求同一广告位同一广告多次曝光建议去重。

- **历史曝光日志文件有 100 万条左右完全重复的数据这是正常的吗？**

  A： 可以作为重复数据处理掉。该场景是同一用户多次看到一次请求曝光的同一个广告，在移动场景中会真实存在，但比例较低。

- **历史曝光日志文件每一行数据代表什么意思？如果这是代表用户点击一个广告的记录，为什么广告请求 ID 不唯一（所有数据 1 亿多条，请求 ID 只有 8 千多万条）？**

  A： 是一次曝光， 即用户看到广告。不唯一的原因是有同一次请求同时曝光多个广告位的广告。

- **您好，我问一下曝光数据中的时间戳是按照什么时区算得，比如第一条的曝光数据时间是 20190217093359 还是 20190217013359 上一条写错了？**

  A： 如无特殊说明，赛题中都按北京时间处理。ps：时间戳本身其实并没有时区的概念。

- **曝光数据的“曝光广告出价 bid”和广告操作数据的“出价”字段 是一回事，那曝光广告出价 bid 这个数值会经常变动。但是却没有用户修改广告出价，这是为什么呢？**

  A: 曝光日志里的广告包含非 cpc 计费的广告和智能出价广告， 这些出价 bid 会在曝光日志里面会经常变动。

- **比赛采用了 SMPAE 作为评分标准的一部分，这样是不是可以认为线上测试数据里面没有曝光量为 0 的样本？**

  A: 有为 0 的， 分母做了平滑处理。

- **请问一下，一个广告是在创建的时候就会上线被曝光还是说要创建完以后，修改广告状态为正常以后才会上线被曝光？**

  A: 可以理解为创建之后即生效。（但是在这个题目中因为只提供了一个时间段内的操作数据，可能创建后有暂停广告的操作在该时间段之前，建议构造样本时选择明确生效的时间段）

关键词：广告数据

- **广告数据空值率高，20%的广告无大小，这是否是个问题？然后行业 ID 数据出现了较多多值特征，如果都作为脏数据清洗，会减少很多记录。这样是否有问题？**

  A: 空值的广告从出题意图上来说就没打算让选手作为训练数据，可以清洗掉。

- **会不会出现广告请求时间正好和广告操作数据中的创建和修改时间相等的情况？如果相等是取修改前的广告属性吗？**

  A: 如果严格相等，则应该是修改时间前的广告属性。不过这种情况应该不是特别关注的重点， 预估目标是天级的曝光，对于单次请求的误差来说应该影响不大。

- **广告操作数据里面，广告修改时间有一些时分秒也都是 0，这部分也是时段缺失了用 0 填充的嘛？**

  A：时分秒是 0 的属于正常数据。

- **操作数据里面只有目标时间那一个月的数据，很多广告的初始化操作数据都不在目标时间段，所以这些广告没有任何出价，投放时间等字段，但实际上应该是已知的吧**?

  A: 新建操作会有出价、投放时间等，没有新建行为的广告实际上无法做样本

- **如果广告 ID 相同，但在广告操作数据里，进行了一些修改，还算是同一广告吗？**

  A： 同一广告 id 不同的设置显然曝光效果可能是有差异的。

- **广告操作数据中的操作后字段值-广告状态取值，这一项具体是什么含义呢？其他的字段可以理解，但是不知道这里的 正常和失效 跟其他字段是否有关系？**

  A: 广告如果失效，则不会被请求召回，更不会曝光。

- **广告的操作文件里，只有 3 万多个广告 id 的信息，为什么曝光数据和广告静态数据里有接近 60 万的广告 id。 差异就是那个公众号推文里写到的其他为了不造成干扰，不提供完整信息的广告？需要自行清洗掉？还是说差异是需要自己填充的？**

  A： 前者。3 万多个和测试集中的广告一样是简单定向的传统 cpc 广告。之所以广告静态数据里提供几乎全部的广告信息，是因为里面的信息部分也许用得到，毕竟他们是在一起竞争的。

- **数据中 bid* pctr + quality_ecpm 的值，总是大于 total_ecpm，是不是意味着 basic_ecpm 并不等于 bid* pctr？**

  A: 因为一些 bid 是非 cpc 广告折算的 cpc 出价，所以会有四舍五入的差异，导致不严格相等。

- **给出的数据中，会出现广告主的预算花光的情况吗？另外，参竞的广告，还存在其他 dsp 的情况吗？**

  A: 为简化比赛题目， 测试数据会是从预算充分的账户中挑选，不用考虑 dsp 情况。

- **“广告 ID：广告所在账户唯一标识，账户结构分为四级：账户-推广计划-广告-素材”这个四级账户在所给数据中是如何体现的呀？**

  A: 为了简化，数据中只涉及账户 id 和 广告 id

- **关于投放时段，会不会一天有两个投放时段，比如 5:00 - 12:00 16:00 - 20:00 这种，我大致看了一下数据，大多数是一天一个投放时段，所以不太确定是否有两个投放时段？**

  A： 广告可以这样设置。

- **素材尺寸和商品 ID 这两个都可以有多值和空值吗？**

  A：素材尺寸可能有多值，也会有缺失（空）。商品 id 不会是多值，文件里面这一列用-1 表示的无商品 id。

广告修改：

- **广告修改过程中涉及的广告 id 中有部分并没有创建行为。请问这要怎么处理呢？未修改字段的值怎么取？**

  A: 取不到的就无法使用。

- **请问再 ad 的 feature 被广告主修改后，是立刻生效还是到固定点再刷新？**

  A： 实际线上会有很短的延迟，可以认为是立刻生效。

- **如果一个广告在上线一段时间后修改了广告内容（例如修改了展示图片）， 这样的话是算作一只新广告吗？是对应一个新的 ID 吗？**

  A: 还是之前的 id。

- **广告 id202 只有修改记录，没有创建记录。类似于这样的 1522 条广告。这种的是因为他的创建时间很早所以没有记录进来么？还是说是异常数据需要自行清洗？**

  A： 异常数据

- **修改定向配置是有时分秒的。修改广告状态都是没有时分秒的，是因为修改广告状态只会在第二天整点生效吗?**

  A: 这么理解对最终结果影响不大。修改广告状态当然不是只有 0 点生效，不过题目中很多修改广告状态的操作是为了简化大家的处理虚拟的操作，即失效后表示该广告不会曝光（这个不一定是真的暂停操作，可能是账户没钱了或者广告审核失效了等）

- **请问预测出结果之后可以 post process 吗？**

  A: 这里没有限制，目前评分只是根据最终的预估结果。