2020 腾讯广告算法大赛参赛手册

题目：广告受众基础属性预估

大赛简介

作为国内领先的营销平台，腾讯广告汇聚腾讯公司全量的应用场景，拥有核心商业数据、营销技术与专业服务能力。同时，腾讯广告依托丰富的社交场景和数以亿计的优质用户，凭借自身在大数据、算法和人工智能领域的技术优势，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。应用场景的复杂性、广告形态的多样性和人群的庞大规模，是实现这一美好目标所必须面对的挑战。为此，腾讯广告也在不断寻找更为优秀的数据挖掘和机器学习算法。

腾讯广告算法大赛步入第四年，已经为来自海内外的企业和研究人员提供了富有研究价值和应用价值的议题，有效地推动了产学研的交流与融合。本届算法大赛的题目“广告受众基础属性预估”兼具实用性和趣味性，从广告行业的经典假设出发，逆向验证这一命题的科学性。参赛者需要探索来自真实业务的海量脱敏数据，综合运用机器学习领域的各种技术，实现更准确的预估。

赛制说明

本次大赛分为初赛、复赛和答辩三个环节。

初赛阶段时间为 5 月 7 日 12:00:00-6 月 22 日 11:59:59。每天（中午 12 点开始的 24 小时内，如无特殊约定，涉及的时间均为北京时间，二十四小时制）限提交 3 次结果，系统将实时计算得到此次提交结果的得分，并在个人信息页展示。

初赛开始后，系统将每天进行一次排名。排名基于每天 12:00 前各队伍提交的结果，并按照参赛队伍当前赛事阶段的历史最优成绩从高到低依次排序，最后一天除外。排行榜将于每天 15:00 更新，此排行榜仅供参考，不作为最终排名结果。

初赛 6 月 22 日 12:00:00 结束，当天 15:00 根据参赛队伍最后一天的最佳成绩更新排行榜（参见“评估方式”）。成绩排名前 10%（原则上最多不超过 100 支队伍，但大赛举办方有权根据报名情况等确定最终数量）的队伍进入复赛。

复赛阶段为 6 月 23 日 12:00:00-7 月 22 日 11:59:59。在复赛阶段，大赛会提供更多的训练数据。各参赛队伍提交结果的方式和排行榜更新的方式与初赛阶段保持一致。复赛结束时，成绩排名在前的 10 支队伍（含并列，大赛举办方有权根据复赛情况等确定最终数量）进入最终答辩环节。

本次大赛将对复赛阶段成绩、答辩成绩和代码进行综合评估，作为最终的比赛成绩。

赛题描述

本届算法大赛的题目来源于一个重要且有趣的问题。众所周知，像用户年龄和性别这样的人口统计学特征是各类推荐系统的重要输入特征，其中自然也包括了广告平台。这背后的假设是，用户对广告的偏好会随着其年龄和性别的不同而有所区别。许多行业的实践者已经多次验证了这一假设。然而，大多数验证所采用的方式都是以人口统计学属性作为输入来产生推荐结果，然后离线或者在线地对比用与不用这些输入的情况下的推荐性能。本届大赛的题目尝试从另一个方向来验证这个假设，即以用户在广告系统中的交互行为作为输入来预测用户的人口统计学属性。

我们认为这一赛题的“逆向思考”本身具有其研究价值和趣味性，此外也有实用价值和挑战性。例如，对于缺乏用户信息的实践者来说，基于其自有系统的数据来推断用户属性，可以帮助其在更广的人群上实现智能定向或者受众保护。与此同时，参赛者需要综合运用机器学习领域的各种技术来实现更准确的预估。

具体而言，在比赛期间，我们将为参赛者提供一组用户在长度为 91 天（3 个月）的时间窗口内的广告点击历史记录作为训练数据集。每条记录中包含了日期（从 1 到 91）、用户信息（年龄，性别），被点击的广告的信息（素材 id、广告 id、产品 id、产品类目 id、广告主 id、广告主行业 id 等），以及该用户当天点击该广告的次数。测试数据集将会是另一组用户的广告点击历史记录。提供给参赛者的测试数据集中不会包含这些用户的年龄和性别信息。本赛题要求参赛者预测测试数据集中出现的用户的年龄和性别，并以约定的格式提交预测结果。大赛官网后台将会自动计算得分并排名。详情参见【评估方式】和【提交方式】。

初赛和复赛除了所提供的训练数据集的量级有所不同之外，其他设置均相同。

数据说明

测试集：https://tesla-ap-shanghai-1256322946.cos.ap-shanghai.myqcloud.com/cephfs/tesla_common/deeplearning/dataset/algo_contest/test.zip

训练集：https://tesla-ap-shanghai-1256322946.cos.ap-shanghai.myqcloud.com/cephfs/tesla_common/deeplearning/dataset/algo_contest/train_preliminary.zip

bucket: tesla-ap-shanghai

region: ap-shanghai

path: /cephfs/tesla_common/deeplearning/dataset/algo_contest/train_preliminary.zip

sdk：cos://tesla-ap-shanghai/cephfs/tesla_common/deeplearning/dataset/algo_contest/train_preliminary.zip

notebook: https://tesla-ap-shanghai-1256322946.cos.ap-shanghai.myqcloud.com/cephfs/tesla_common/deeplearning/dataset/algo_contest/train_preliminary.zip

训练数据中包含了一组用户 91 天的广告点击日志，被组织为三张表，以带标题行的 CSV 文件的格式提供（编码采用无 BOM 的 UTF-8），分别是：click_log.csv，user.csv，ad.csv。测试数据包含了另一组用户 91 天的广告点击日志，组织方式与训练数据相同，但不包含 user.csv。各表的详细格式描述如下（字段顺序以下面的描述为准，各字段用逗号‘，’分隔，中间无空格）：

click_log.csv:

- time: 天粒度的时间，整数值，取值范围[1, 91]。
- user_id: 从 1 到 N 随机编号生成的不重复的加密的用户 id，其中 N 为用户总数目（训练集和测试集）。
- creative_id: 用户点击的广告素材的 id，采用类似于 user_id 的方式生成。
- click_times: 当天该用户点击该广告素材的次数。

user.csv:

- user_id
- age: 分段表示的用户年龄，取值范围[1-10]。
- gender：用户性别，取值范围[1,2]。

ad.csv:

- creative_id
- ad_id: 该素材所归属的广告的 id，采用类似于 user_id 的方式生成。每个广告可能包含多个可展示的素材。
- product_id: 该广告中所宣传的产品的 id，采用类似于 user_id 的方式生成。
- product_category: 该广告中所宣传的产品的类别 id，采用类似于 user_id 的方式生成。
- advertiser_id: 广告主的 id，采用类似于 user_id 的方式生成。
- industry: 广告主所属行业的 id，采用类似于 user_id 的方式生成。

提交方式

参赛者提交的结果为一个带标题行的 submission.csv 文件，编码采用无 BOM 的 UTF-8，具体格式如下（字段顺序以下面的描述为准，各字段用逗号分隔，中间无空格）：

- user_id
- predicted_age: 预测的用户年龄分段，取值范围[1,10]。
- predicted_gender: 预测的用户性别，取值范围[1,2]。

测试数据集中每个用户均应在 submission.csv 文件中对应有且仅有一行预测结果。各用户的预测结果在该文件中的出现顺序与评估结果无关。

评估方式

大赛会根据参赛者提交的结果计算预测的准确率（accuracy）。年龄预测和性别预测将分别评估准确率，两者之和将被用作参赛者的打分。

测试数据集会和训练数据集一起提供给参赛者。大赛会将测试数据集中出现的用户划分为两组，具体的划分方式对参赛者不可见。其中一组用户将被用于初赛和复赛阶段除最后一天之外的排行榜打分计算，另一组则用于初赛和复赛阶段最后一天的排行榜打分计算，以及最后的胜出队伍选择。