2021“微派·种子杯”创新性软件编程 PK 赛赛题

(社交游戏 app-会玩 未成年玩家识别)

一、大赛背景

随着国内对未成年保护越来越重视，市面主流 app 都新增了未成年防沉迷功能和实名认证的要求，进而营造出有益未成年身心健康的网络环境，目前来看取得了一定效果。但与此同时也面临诸多挑战，一方面，未成年会用成年长辈的信息来完成实名注册，进而绕过平台限制;另一方面租号买号日益猖獗，获号成本极低。具体到本场景，会玩 app 对未成年用户的游戏时间和充值金额做了限制，但存在一部分未成年用户冒用成年人身份证实名认证，因此从所有用户中找出冒用了成年身份实名的未成年用户是本赛题的目标。由于风控场景本身的特殊性，收集大量的有标签数据往往面对较大的困难，在某些场景甚至不可能；仅利用小样本量的数据来进行机器学习建模已经成为机器学习的发展中一个十分重要的课题，这也是本次比赛最大的挑战；

二、大赛数据

本次大赛提供源自真实业务的脱敏数据做为训练集，主要包含如下几个维度的数据：用户基础属性，用户历史统计特征，用户行为轨迹，部分数据可能存在缺失需要自行处理，最后参赛队伍基于训练集（包含标签：是否为未成年）构建未成年预测模型；初赛阶段发布测试集 A（不可下载，数据不可见，仅供评测）进行模型评估和排名；复赛阶段公布测试集 A 的标签供选手扩充数据集进行模型优化，并发布测试集 B（不可下载，数据不可见，仅供评测）进行模型评估和排

名；复赛 B 榜 Top10 进入决赛，与此同时进入决赛的队伍需要提交代码和 ppt 文档到微派-会玩算法团队进行审核，若代码无法复现，需要给出说明，否则取消成绩。

初赛阶段公布数据如下：

- 表 1:user_base_info (数据量-37045)

字段	类型	含义
col_1	bigint	编号 id
col_2	int	认证年龄
col_3	int	意向社交用户类型
col_4	int	性别
col_6	int	游玩卡等级
col_7	int	游玩卡积分
col_8	int	奖券数量
col_10	int	是否打开消息推送
col_11	int	导流渠道
col_12	int	注册天数
col_13	int	app 版本
col_14	int	设备品牌
col_14	int	设备型号
label	int	类别标签：1-未成年，0-成年 (训练集有值，测试集为空)

- 表 2:user_his_features(数据量-37045)

字段	类型	含义
col_1	bigint	编号 id
col_2	int	历史加好友数量
col_3	int	历史加未成年用户数量
col_4	int	历史加 30 岁以下用户数量
col_6	int	历史加社交意向为未成年好友数量
col_7	int	历史浏览好友数量
col_8	int	历史浏览未成年用户数量
col_10	int	历史浏览 30 岁以下用户数量
col_11	int	历史浏览社交意向为未成年用户数量
col_12	int	历史私聊好友数量
col_13	int	历史私聊未成年用户数量

col_14	int	历史私聊 30 岁以下用户数量
col_15	int	历史私聊社交意向为未成年用户数量
col_16	int	历史送礼物好友数量
col_17	int	历史送礼物未成年用户数量
col_18	int	历史送礼物 30 岁以下用户数量
col_19	int	历史送礼物意向为未成年用户数量

- 表 3: user_track(数据量-2568838)

近 3 个月用户登录 app 行为，起始日期记为 1，后续依次增加；例：近 3 个月最早的那一天记为 1，第二天记为 2

字段	类型	含义
col_1	bigint	编号 id
col_2	int	登录 app 日期
col_3	int	登录日期类型 0-工作日，1-周末
col_4	int	当日最早登录时刻
col_5	int	当日最晚登录时刻

- 表 4:submit_sample

测评提交数据仅包含一个字段，**不带列名**，提供预测的类别即可（样例可参考 submit/submit_sample_a.txt）

字段	类型	含义
label	int	预测标签：1-未成年，0-成年

- 表 5:test_a（用作评测，不公开 label, 数据量-365）

测评数据包含两个字段，带列名，分别为用户编号 id 和真实的类别

字段	类型	含义
col_1	bigint	编号 id
label	int	真实标签：1-未成年，0-成年

复赛阶段公布数据如下：

- 表 6:test_a（公开，数据量-365）

字段	类型	含义
col_1	bigint	编号 id
label	int	真实标签：1-未成年，0-成年

- 表 7:test_b（用作评测，不公开 label，数据量-456）

字段	类型	含义
col_1	bigint	编号 id
label	int	真实标签：1-未成年，0-成年

三、评估方式

选手需要提交结果表到组委会，只包含**一列数据**即预测标签（0 | 1），**标签顺序与给定测试集相对应**，并且**不含列名**，示例见`submit/submit_sample_a.txt`，**请注意上传时最后一行不能包含空行**。初赛提交表名（见表 4）submit_sample_a.txt，复赛提交表名 submit_sample_b.txt。

具体评估规则如下：

定义成年人为负样本，未成年为正样本；准确率 Precision 记为 P，召回率 Recall 记为 R，其中：

P = 预测为正的样本中真正为正的数量/预测为正的样本总量

R = 预测为正的样本中真正为正的数量/真实正样本总量

比如 10000 条数据，其中 7000 条为负样本，3000 条为正样本，参赛者一共将 2000 条判断为正样本，其中只有 1500 条真正是正样本，那么 P=1500/2000 = 0.75；R=1500/3000=0.5 参赛队伍最终得

分 F = 5PR/(2P+3R)=0.6250 最终排名按照 F 值(保留 4 位小数)评判，F 值越大，代表结果越优，排名越靠前。