机器人检测

1 题目背景

背景：GAMMA交易所是一家繁忙的交易平台，其旗下的交易网站提供了一个广受欢迎的拍卖服务。然而，最近，工作人员注意到一个令人不安的现象：拍卖活动中充斥着大量的机器人参与者。这些机器人的存在严重影响了人类用户的拍卖体验，因为人类用户发现自己很难在交易中胜过这些高速、高效的机器人。这个问题已经开始对用户的参与热情产生负面影响，甚至导致了用户流失。

为了解决这个问题，保护用户的利益并恢复他们对平台的信心，GAMMA交易所现在正在寻找有效的解决方案，以便能够准确地检测和识别出交易中的机器人行为。

2 数据

为了用户隐私，一些敏感数据经过了特殊转化处理。

对于用户文件user.csv，描述每个用户的一些基本信息。在train_user.csv中包含outcome信息，test_user.csv中不包含outcome信息。

字段	含义
bidder_id	投标人的id
payment_account	投标人账户
address	投标人地址
outcome	标签,1代表为机器人，0为非机器人

对于交易文件trade.csv，描述每场拍卖会的每次交易信息，一件竞拍物品可能有来自不同/相同竞拍人的多次出价。

字段	含义
bid_id	此次出价的id
bidder_id	投标人id
auction	此次竞拍id
merchandise	拍卖的商品
device	设备类型
time	出价时间
country	IP地址所属国家
ip	出价的ip地址
url	投标人来源的网址

3 任务目标

根据用户的基本信息，以及每次拍卖交易的记录的交易信息，预测每个用户是否为机器人的概率。

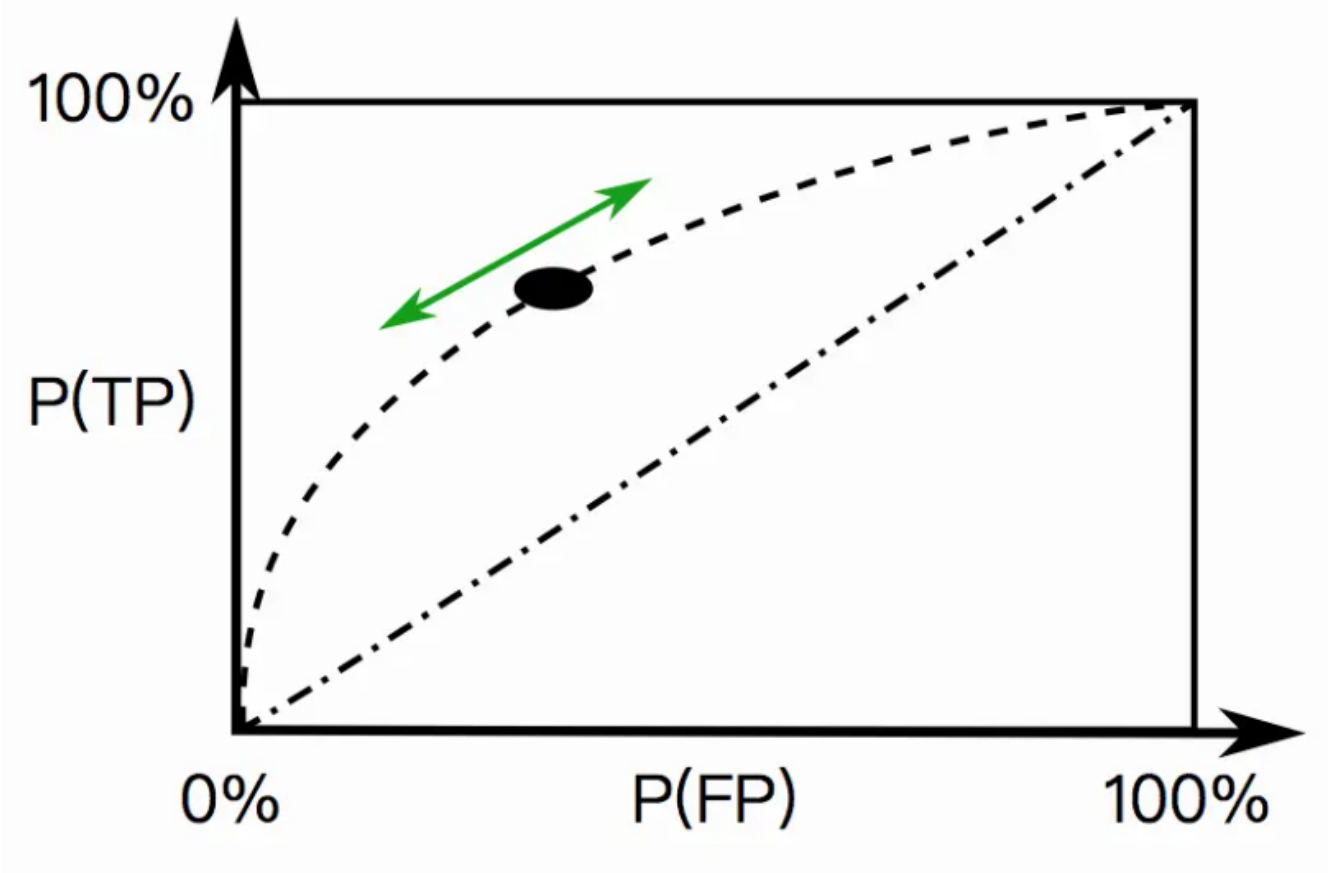
4 输出格式

一个文件，每行代表一个用户是机器人的概率，概率越接近1表示越可能是机器人。

```
0.660345999585404
0.4010444851453697
0.8161498880818824
0.8435329973479264
0.6466601122551972
0.6928764107873429
0.4854858208856211
0.11522604580884699
0.8742830361941845
0.31178306749154505
0.5543307673421333
0.41176355431263556
```

5 评估指标

AUC（Area Under the ROC Curve）。AUC是ROC曲线下的面积，ROC曲线是以假阳性率（False Positive Rate）为横坐标，真阳性率（True Positive Rate）为纵坐标画出的曲线。AUC值越大，说明模型的预测性能越好。AUC对样本的分布和阈值选择不敏感，因此常被用于评价不平衡数据的分类器性能。



6 分组

2人为一组组队完成，落单可以组3人队

7 评分标准

- 排名得分(60%)：
  - rk1:100
  - rk2-3:95
  - rk4以后：从92开始按照位次递减（96 - your_rk）
  - 如果最终测试rk低的组效果也不错，会酌情减少递减的幅度
- 报告得分（30%):
  - 数据处理过程，包含分析
  - 记录实验的结果并保存
  - 如果rk分太低，但是尝试了多种方法，酌情加分
  - 线下报告
- 代码得分（10%):
  - 简洁，美观
  - 注释完整

8 提交文件

- 如上输出格式的csv格式文件，请严格按照给定的user_test的顺序输出每个用户的预测值。
- 报告中要包含小组成员和分工明细
- 代码、报告和使用pip freeze > requirements.txt导出你的python环境包依赖，打包为zip压缩包，注明如何运行。命名方式为**${你的组号}组.zip**，例如1组.zip。

9 一些建议

不要手动修改预测标签（方案和结果对不上的我会复现结果）

不要抄袭

不要脚本爆破服务器（提交次数限制32）

10 截止日期

6.9(15周)