加油站营销推荐算法赛题

1、 场景介绍

随着互联网的发展，用户对隐私越来越重视，国家也出台了各种法律法规来规范商家对数据的使用。为了保护用户隐私，商家需要在隐私保护的数据上寻找特定用户群体，制定对应的营销策略，以及在合适的场地推广。本赛题场景为：有一个加油站计划给来加油的用户制定优惠券发放策略，根据各方面的信息综合判断是否给用户发放优惠券，发放券额，等等。银行拥有不同车主的个人信息和消费信息，车载智能系统厂商拥有不同车主的行程信息，某加油站(优惠卷发放方)拥有不同车主过往使用优惠卷的信息以及当天的环境等信息。假设上述几方都是互相不信任的，且不会合谋的前提下，通过综合使用多方数据，在保护用户隐私的情况下进行优惠券发放模型的构建。

2、 赛题任务

- 计算参与方 :甲方(银行）、乙方（车载智能系统）、丙方（某加油站）共三方。
- 数据输入：
    - 为参赛队伍提供银行方数据集（A）、车载智能系统数据集（B）和某加油站数据集(C)，供解决方案开发使用。数据特征见附录一。
    - 数据源之间用户已经完成样本对齐
    - 数据中存在缺失值
    - 推荐目标：用户是否有兴趣使用优惠卷
- 目标输出 :基于三方数据安全训练模型，验证模型效果。
- 技术要求 :完全使用多方安全计算（MPC）保护建模全程，不能使用联邦学习、差分隐私等其他隐私保护技术；可以采用3PC或者2PC的各种MPC协议。
- 安全性要求:128bits安全性、30bits统计安全性；符合MPC安全定义。
- 安全假设：
    - 半诚实模型假设
- 隐私保护目标：
    - 三方之间不能交互明文数据；建模中间计算结果不能暴露。

- 评测原则：
  - 解决方案评测时，会使用另一套未公开的数据集进行训练和测试。训练数据集样本量4000，测试数据集样本量2000。
  - 结果的正确性(训练好的模型在测试数据集上AUC）。
  - 联合计算所需的总时间（包括预计算、在线计算）。
  - 方案先按AUC排名，AUC相差不超过0.01的方案间，综合按联合计算耗时和网络总流量排名。
- 评测环境 :提供相对高配的计算型云主机配置，暂定通过KVM部署，KVM计算环境的暂定参考配置为 :8核心CPU、32 GB 内存、500GB硬盘、带宽1G。
- 解决方案提交要求 :
  - 运行于3方的MPC Training程序
    - 三方输入：训练数据集，按格式，带有label，通过多方安全计算（MPC）完成计算。
    - 丙方输出：解密后的model
  - 运行于丙方的明文计算Testing程序
    - 丙方输入：model；测试数据集，按格式，不带label
    - 丙方输出：推荐结果
  - 启动training和testing的必要脚本
  - 详细的Readme和参赛方案设计文档pdf
  - 参赛选手提交到平台方案需要在12小时内完成运行
  - 解决方案提交方式、日志和接口格式详见附录二

附录一：三个数据集特征如下

银行	车载系统特征	加油站特征
ID	ID	ID
性别	目的地	天气
年龄	车上乘客	时间
婚姻状况	到优惠券餐厅/酒吧的车程大于15分钟	优惠卷
教育	到优惠券餐厅/酒吧的车程大于25分钟	有效期
职业	餐厅/酒吧是否与您当前的目的地在同一方向	是否接受优惠券
收入	餐厅/酒吧是否与您当前的目的地在相反方向	
每月去酒吧次数		
每月去咖啡店次数		
每月点外卖次数		
每月去餐馆平均每人消费低于120元次数		
每月去餐馆平均每人消费低于120～300元次数		

附录二：输出格式

为了更好的评估所构建的模型性能，对模型的输出进行规范，具体要求如下：

1. testing程序将测试集的预测结果保存为**csv**格式，文件共**2**列，包含**ID**和预测概率**Y_prob**。
2. Y_prob字段的取值范围为**[0,1]**，取值越接近1表示该用户有优惠券的概率越高。
3. 在写入结果文件时，请保留表头信息，即ID和Y_prob

   保存文件的格式示例如下：

ID	Y_prob
0	0.12
1	0.88

2	0.75

4. 具体的提交方式等待后续通知。