[2017“贝贝网·种子杯” 复赛试题]

【赛题简介】

贝贝网是国内最大的母婴特卖网站，以品牌正品、独家折扣、限时抢购为特色。贝贝网定位于分众电商，针对妈妈群体提供细致深入的产品与服务。

购买预测是电商场景中非常重要的课题，它所延伸出的技术、算法能广泛应用到大数据营销几乎所有场景中，如搜索、推荐和营销服务。购买预测算法竞赛将为参赛者提供过于一段时间贝贝网部分活跃用户在贝贝app上产生的商品浏览、收藏、加购和购买等脱敏行为数据，几十万级别的贝贝部分品类下的商品文本和属性数据。期待参赛者能从以上行为、文本数据中挖掘出用户购物模型，为用户提供优质的、专业的个性化推荐方案。

【赛制说明】

1. 选手下载数据，在本地调试算法，在线提交结果测试评分；
2. 每天提供2次的评测和排名机会，两小时内更新排行榜；按照评测指标得分从高到低排序；排行榜将选择历史最优成绩进行展示；
3. 得分最高的10支队伍将进入决赛。

【赛题数据】

赛题数据包含3部分，分别为商品基本信息数据、用户基本信息数据和用户历史行为数据。

赛题数据下载：competition.zip

商品基本信息数据

文件名：product_info.txt 共6列数据，TAB键分隔。

列	说明
1	商品ID，整型
2	商家ID，整型
3	品牌ID，整型
4	类目ID，整型
5	商品标题的分词，空格分割的字符串，顺序已打乱
6	商品价格，单位：分，整型

用户基本信息数据

文件名：user_info.txt 共8列数据，TAB键分隔。

列	说明
1	用户ID，整型
2	用户会员等级：1普通，2铜牌，3铁牌，4金牌，5铂金，6钻石
3	用户性别: 0默认，1男，2女,-99 未知
4	用户生日
5	用户年龄
6	宝宝生日
7	宝宝年龄月份
8	宝宝性别: 0默认，1王子，2公主，3孕育中,-99 未知

用户行为数据

文件名：behavior_info.txt 共4列数据，TAB键分隔。

列	说明
1	用户ID，整型
2	商品ID，整型
3	行为时间戳，the number of seconds from unix epoch (1970-01-01 00:00:00 UTC)
4	行为类型，1:浏览；2：收藏；3：加购；4：购买

用户行为数据包含了6万多贝贝活跃用户从2017年7月26日到2017年8月25日的系统交互行为，参赛者需要从这些行为数据中挖掘出影响用户购买决策的有效特征，并基于这些特征构建模型预测用户未来三天（2017年8月26日到2017年8月28日）的购物行为。

比赛测试说明

选手的答案内容存储在文本文件中，共2列数据，TAB健分隔。标准zip压缩后提交，压缩后的文件大小不能超过10M，压缩包解压后应能够直接得到答案文件，**不能有目录结构，文件数量也不能多于1个**。测试文件名不做要求。

列	说明
1	用户ID，整型
2	商品ID，整型

答案文件中的每一行表示一组记录，表示模型预测出的2017年8月26日到2017年8月28日期间该用户购买了该商品。只有预测出的购买行为才需要提交。

测试文件提交至：http://dian.hust.edu.cn/seedcup2017/ （10月5日开放测试）

测试截止时间：2017年10月15日22:00

【评测指标】

参赛选手提交的答案将和训练数据中的用户在2017年8月26日到2017年8月28日期间的真实购物行为数据比对，计算F1分数作为最终的成绩排名。

具体地，假设选手提交的答案文件中的购物pair组成的集合为A；系统保留的真实的购物pair组成的集合为B，则

精确率(precision)=（A ∩ B）/ A

召回率(recall)=(A ∩ B) / B

F1分数的计算方式：

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\mathrm{recall}} + \frac{1}{\mathrm{precision}}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

【最终提交说明】

提交内容:最高分数测试样例文件、代码以及比赛报告，格式如下：

- seedCup2017 复赛-xxx 队.zip
  - answer.txt(要求测试最优的结果)
  - src (源文件目录)
  - xxx.pdf
- 比赛报告内容
  - 使用语言以及运行环境。
  - 提供代码相应的接口并指明运行需要用到的变量含义，以便裁判组进行测试。
  - 数据特征提取思路。
  - 预测模型选取（包括对于规则的描述和最终模型的选择）。
  - 对于模型参数的选择与优化思路。
  - 报告内容不限于以上所述内容。

提交截止时间：2017年10月15日22:00， 未提交最终内容的队伍视为放弃比赛。

比赛报告提交至大赛公邮 seedcup@dian.org.cn ， 邮件标题以队名命名。

【注意事项】

1. 本次比赛不允许使用外部数据，但可以使用开源的已有算法和工具。
2. 如果发现参赛队伍有造假、作弊、雷同等行为，将取消该队伍的参赛资格及奖励。
3. 复赛 / 决赛过程中，大赛评审组可能会根据比赛情况，对比赛内容和评分标准进行调整。
4. 比赛最终解释权归大赛评审组所有。

【参考文档】

1. Command line tools for Machine learning
2. scikit-learn
3. tensorflow