地址：[http://www.datafountain.cn/#/competitions/247/intro](http://www.datafountain.cn/%23/competitions/247/intro)

竞赛概述

京东作为中国最大的自营式电商，在保持高速发展的同时，沉淀了数亿的忠实用户，积累了海量的真实数据。如何从历史数据中找出规律，去预测用户未来的购买需求，让最合适的商品遇见最需要的人，是大数据应用在精准营销中的关键问题，也是所有电商平台在做智能化升级时所需要的核心技术。本次大赛以精准推荐中的一个具体问题为例，希望找到数据挖掘领域的顶尖人才，与我们共同奔跑在浪潮之巅。

本次大赛以京东商城真实的用户、商品和行为数据（脱敏后）为基础，参赛队伍需要通过数据挖掘的技术和机器学习的算法，构建用户购买商品的预测模型，输出高潜用户和目标商品的匹配结果，为精准营销提供高质量的目标群体。同时，希望参赛队伍能通过本次比赛，挖掘数据背后潜在的意义，为电商用户提供更简单、快捷、省心的购物体验。

数据介绍

符号定义：

S：提供的商品全集；

P：候选的商品子集，P是S的子集；

U：用户集合；

A：用户对S的行为数据集合；

C：S的评价数据。

训练数据部分：

提供2016-02-01到2016-04-15日用户集合U中的用户，对商品集合S中部分商品的行为、评价、用户数据；提供部分候选商品的数据P。

选手从数据中自行组成特征和数据格式，自由组合训练测试数据比例。

预测数据部分：

2016-04-16到2016-04-20用户是否下单P中的商品，每个用户只会下单一个商品；抽取部分下单用户数据，A榜使用50%的测试数据来计算分数；B榜使用另外50%的数据计算分数。

为保护用户的隐私和数据安全，所有数据均已进行了采样和脱敏。

数据中部分列存在空值或NULL，请参赛者自行处理。

1. 用户数据

|  |  |  |
| --- | --- | --- |
| user\_id | 用户ID | 脱敏 |
| age | 年龄段 | -1表示未知 |
| sex | 性别 | 0表示男，1表示女，2表示保密 |
| user\_lv\_cd | 用户等级 | 有顺序的级别枚举，越高级别数字越大 |
| user\_reg\_dt | 用户注册日期 | 粒度到天 |

2. 商品数据

|  |  |  |
| --- | --- | --- |
| sku\_id | 商品编号 | 脱敏 |
| attr1 | 属性1 | 枚举，-1表示未知 |
| attr2 | 属性2 | 枚举，-1表示未知 |
| attr3 | 属性3 | 枚举，-1表示未知 |
| cate | 品类ID | 脱敏 |
| brand | 品牌ID | 脱敏 |

3. 评价数据

|  |  |  |
| --- | --- | --- |
| dt | 截止到时间 | 粒度到天 |
| sku\_id | 商品编号 | 脱敏 |
| comment\_num | 累计评论数分段 | 0表示无评论，1表示有1条评论，  2表示有2-10条评论，  3表示有11-50条评论，  4表示大于50条评论 |
| has\_bad\_comment | 是否有差评 | 0表示无，1表示有 |
| bad\_comment\_rate | 差评率 | 差评数占总评论数的比重 |

4. 行为数据

|  |  |  |
| --- | --- | --- |
| user\_id | 用户编号 | 脱敏 |
| sku\_id | 商品编号 | 脱敏 |
| time | 行为时间 |  |
| model\_id | 点击模块编号，如果是点击 | 脱敏 |
| type | 1.浏览（指浏览商品详情页）；  2.加购；3.购物车删除；4.下单；5.关注；6.点击 |  |
| cate | 品类ID | 脱敏 |
| brand | 品牌ID | 脱敏 |

登录并报名参赛后方可在“数据下载”页面下载数据。

数据下载

* [数据下载（提取码：7xpy）](https://wx.jcloud.com/market/activity/2017/jdata)

地址：<https://wx.jcloud.com/market/activity/2017/jdata>

**任务描述：**

参赛者需要使用京东多个品类下商品的历史销售数据，构建算法模型，预测用户在未来5天内，对某个目标品类下商品的购买意向。对于训练集中出现的每一个用户，参赛者的模型需要预测该用户在未来5天内是否购买目标品类下的商品以及所购买商品的SKU\_ID。评测算法将针对参赛者提交的预测结果，计算加权得分。

作品要求

初赛提交CSV结果文件，进入复赛时提交源代码。

初赛提交CSV文件中包含对有购买意向的用户所购买商品的预测结果，字段如下：

user\_id：用户ID，保证唯一，请勿在一次提交的结果文件中包含重复的user\_id

sku\_id：商品集合P中的商品ID，请勿在同一行中提交多个sku\_id

对于预测出没有购买意向的用户，在提交的CSV文件中不要包含该用户的信息。

[结果文件示例](http://www.datafountain.cn/projects/download/demo.csv)

规则

评分

参赛者提交的结果文件中包含对所有用户购买意向的预测结果。对每一个用户的预测结果包括两方面：

1、该用户2016-04-16到2016-04-20是否下单P中的商品，提交的结果文件中仅包含预测为下单的用户，预测为未下单的用户，无须在结果中出现。若预测正确，则评测算法中置label=1，不正确label=0；

2、如果下单，下单的sku\_id （只需提交一个sku\_id），若sku\_id预测正确，则评测算法中置pred=1，不正确pred=0。

对于参赛者提交的结果文件，按如下公式计算得分：

Score=0.4\*F11 + 0.6\*F12

此处的F1值定义为：

F11=6\*Recall\*Precise/(5\*Recall+Precise)

F12=5\*Recall\*Precise/(2\*Recall+3\*Precise)

其中，Precise为准确率，Recall为召回率.

F11是label=1或0的F1值，F12是pred=1或0的F1值.

**参赛资格：**

大赛面向全社会开放，高等院校、科研单位、互联网企业等人员均可报名参赛。大赛主办和协办单位有机会提前接触赛题和数据的人员不得参加比赛，其他员工可以参与比赛排名，但不领取任何奖项。

**基本规则：**每支队伍最多5人，每天最多提交2次。

**附加规则：**

禁止使用外部数据；每人只能参与一个队伍；保证参赛队员报名信息准确有效，不得使用小号，否则会被取消参赛资格及奖励；保证算法可复现且不存在抄袭，对进入B榜前50名的队伍将进行代码人工检查，违反规则将取消参赛资格和奖励。

奖池奖项



\* Special offer：

社会优秀人才及应届生：通过专家评委考核即拥有offer，每支队伍至少一人确定通过；入职前boss一对一指导；充分的技术方向选择权；成为京东算法大赛代言人；如为应届博士生，直通科学家项目。

 在校非应届生：直通JD 京锐夏令营，最优秀人才的最佳实习实践。

\*大赛周冠军：

（1）周冠军共计评选7次，同一支队伍可以多次获得多周的周冠军，以每支队伍截止到下述时间点提交结果总分中最好成绩为准。

（2）周冠军评测时间为下述日期的0点，2017年4月3日、2017年4月10日、2017年4月17日、2017年4月24日、2017年5月1日、2017年5月8日、2017年5月15日；

（3）周榜单公布时间为评测日期当日的中午12点；

（4）周榜单前Top5团队需要准备最优成绩的答案结果文件及相关资料（可运行代码、运行环境说明、解题思路），大赛主办方将对TOP1团队进行作弊检验，通过检验即可获取奖励；若检验不通过，则需要Top2团队提交相关资料，通过检验即可获取奖励，以此类推；如果Top5团队全部未通过检验，则本轮周冠军轮空；

届时大赛主办方将邮件及短信通知相关资料提交事宜，请参赛选手注意查收。

（5）Top1参赛团队提交资料截止时间为每周三24点；其他团队如果候补提交资料，需要在接到邮件或短信通知后12小时之内提交，以平台邮件或短信中较早的发出时间为准，过期将视为放弃排名。

\*京东员工参赛规则详见内部宣传公告且不参与对外排名，奖金由京东内部提供，不占用大赛奖金池。

\*大赛所有奖项的奖金以队伍为单位统一在大赛结束后发放。

\*京东享有本次比赛所有模型代码及相关资料的使用权，最终解释权归京东所有。

赛程活动

**赛程说明：**

比赛赛程采用线上初赛+现场决赛答辩。

3月19日-5月29日为初赛阶段。

3月19日00:00开放赛题信息和预报名，参赛者可以注册DF用户、浏览赛题；

3月25日00:00开放数据、结果提交和评测；

5月13日24:00报名截止，并关闭组队；

5月20日24:00关闭初赛A榜；

5月21日00:00开放B榜提交，并进行小号作弊和算法有效性排查；

5月25日24:00关闭初赛B榜，并开放提交相关审核材料；

5月27日24:00提交相关审核材料截止；

5月29日18:00公布入围决赛的10支队伍名单。

QQ群：

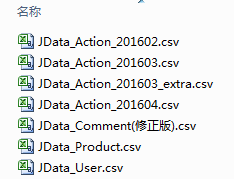
京东JData算法大赛交流群：435027670

DataFountain平台用户支持群：543820097

**赛题篇**

1、 Q：赛题数据量级

A：本次赛题数据共7个csv文件。



2、 Q：赛题数据问题件的编码格式

A：提供的数据文件编码格式为GBK。

3、 Q：评测数据中一个用户是否存在购买多个商品？

A：不存在，一个用户（user\_id）只购买一个所提供的候选商品集合中的商品（sku\_id）；用户购买候选商品集合之外的商品无需提交。

4、 Q：关于商品集合P和S

A：集合P为候选商品集合，即参赛者预测的结果中的sku需要在集合P中；集合S为行为数据中的商品全集；

5、 Q：评论数据是当天的还是累计的

A：评论数据为截止到当日的累计数据

6、 Q：关于model\_id的解释

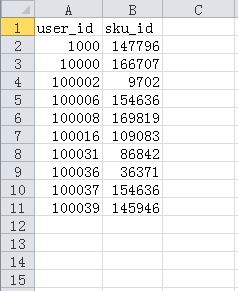
A：model\_id表明用户在页面上点击了哪一个位置；数据中可能存在一些空值，有可能是异常值，也有可能是在页面上点击了一个空白的位置产生的数据，请参赛者自行理解并处理

7、 Q：提交数据文件有什么要求？

A： 1）提交数据文件需要去重处理，即：提交结果文件中用户user\_id唯一，不存在重复；

2）预测用户下单的商品sku\_id，必须在商品子集P中；

3）提交数据文件格式为csv，编码为utf-8，第一行为表头，即：第一列为“user\_id”，第二列为“sku\_id”，提交数据文件截图如下：



8、 Q：关于日期和时间格式

A：日期格式统一为"yyyy-mm-dd"，时间格式统一为"yyyy-mm-ddhh:mi:ss"

9、 Q：关于评价数据中的有差评但是差评率为0的情况

A：这种情况是因为第一版数据中存在一些小问题，我们已经进行了紧急修正，如果是下载了第一版数据的同学，请重新下载修正版评价数据（压缩包文件名：JData\_Comment(修正版)）。

10、Q：关于3月12~3月20的行为数据缺失情况

A：目前数据已经补充完成，如果是下载了第一版数据的同学，请下载3月份行为补充数据（压缩包文件名：JData\_Action\_201603\_extra.csv）。

11、Q：关于用户表中存在10条记录，用户注册时间晚于2016年4月15日情况

A：此情况为真实生产环境中的异常数据，请选手自行理解并处理。

**奖励篇**

1、 Q：获奖参赛队伍每个成员是否都能拿到Special Offer？

A：奖项中所涉及的Special Offer均是真Special Offer，只要获奖参赛团队中成员通过专家评委的考核，就能拿到Special Offer，参赛团队成员通过几人，京东就敢发出几个Special Offer；获得冠亚季军的每个团队至少发出一个Special Offer。

以上资料来自：京东大赛官网，如有变化，请以官网为准。

<http://www.dudu2007.com/>   
友情整理:大数据矿工服务网（大数据网址导航，知识搜索，智慧学习）

**祝所有赛友都取得好成绩，找到内心的“真我”。**

**求组队QQ: 27461933**