**2018阿里妈妈广告算法大赛调研方案**

2018-11-08

# 目录

[1. 赛题描述 - 1 -](#_Toc529441237)

[2. 评价指标（Logloss） - 1 -](#_Toc529441238)

[3. 数据格式 - 1 -](#_Toc529441239)

[4. 输入 - 2 -](#_Toc529441240)

[5. 输出 - 2 -](#_Toc529441241)

[6. 数据分析 - 2 -](#_Toc529441242)

[7. Top1&2团队算法 - 4 -](#_Toc529441243)

[8. Top1团队模型&特征&主代码介绍 - 5 -](#_Toc529441244)

[8.1 Top1模型设计 - 5 -](#_Toc529441245)

[8.2 Top1特征简单介绍 - 5 -](#_Toc529441246)

[8.2.1Embedding特征 - 5 -](#_Toc529441247)

[8.2.2统计特征 - 6 -](#_Toc529441248)

[8.2.3时差特征 - 6 -](#_Toc529441249)

[8.2.4排序特征 - 6 -](#_Toc529441250)

[8.3 Top1主代码&特征重要度 - 7 -](#_Toc529441251)

[8.3.1主代码展示 - 7 -](#_Toc529441252)

[8.3.2特征重要度结果展示 - 7 -](#_Toc529441253)

[9. Top2团队模型&特征介绍 - 8 -](#_Toc529441254)

[9.1 Top2模型设计 - 8 -](#_Toc529441255)

[9.1.1数据划分 - 8 -](#_Toc529441256)

[9.1.2模型设计 - 8 -](#_Toc529441257)

[9.2 Top2特征详细介绍 - 9 -](#_Toc529441258)

[9.2.1第一部分：基础特征群（3+14=17维） - 9 -](#_Toc529441259)

[9.2.2第二部分：查询交互，用户交互，竞争特征（60+52+16=128维） - 10 -](#_Toc529441260)

[9.2.3第三部分：统计转化率特征，转化率排名特征（149+63=212维） - 11 -](#_Toc529441261)

[9.2.4第四部分：统计点击特征，点击数量占比特征（36+198+48=282维） - 14 -](#_Toc529441262)

[9.2.5第五部分：不购买\_购买特征，趋势特征，一次性购买特征，出现\_购买特征，item\_shop属性变化特征（7+225+12+3+40=287维） - 16 -](#_Toc529441263)

[9.2.6第六部分：强制cross特征（100\*99/2=4950维） - 17 -](#_Toc529441264)

[9.3 Top2特征重要度 - 17 -](#_Toc529441265)

[9.3.1特征群重要度结果展示 - 17 -](#_Toc529441266)

[9.3.2特征重要度结果展示 - 17 -](#_Toc529441267)

# 赛题描述

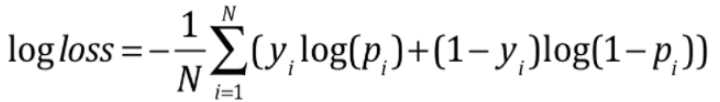
以阿里电商广告为研究对象，给定广告点击相关的用户（user）、广告商品（ad）、检索词（query）、上下文内容（context）、商店（shop）等信息的条件下预测广告产生购买行为的概率（pCVR）。形式化定义为：pCVR=P(conversion=1 | query, user, ad, context, shop)。属于2分类问题。

初赛和决赛分别为：

（1）日常的转化率预估；

（2）特殊日期的转化率预估；

# 评价指标（Logloss）



# 数据格式

|  |  |
| --- | --- |
| **字段** | **字段解释** |
| instance\_id | 序列编号 |
| item\_id | 广告商品编号，Long类型 |
| item\_category\_list | 广告商品的的类目列表，String类型；从根类目（最粗略的一级类目）向叶子类目（最精细的类目）依次排列，数据拼接格式为 "category\_0;category\_1;category\_2"，其中 category\_1 是 category\_0 的子类目，category\_2 是 category\_1 的子类目 |
| item\_property\_list | 广告商品的属性列表，String类型；数据拼接格式为 "property\_0;property\_1;property\_2"，各个属性没有从属关系 |
| item\_brand\_id | 广告商品的品牌编号，Long类型 |
| item\_city\_id | 广告商品的城市编号，Long类型 |
| item\_price\_level | 广告商品的价格等级，Int类型；取值从0开始，数值越大表示价格越高 |
| item\_sales\_level | 广告商品的销量等级，Int类型；取值从0开始，数值越大表示销量越大 |
| item\_collected\_level | 广告商品被收藏次数的等级，Int类型；取值从0开始，数值越大表示被收藏次数越大 |
| item\_pv\_level | 广告商品被展示次数的等级，Int类型；取值从0开始，数值越大表示被展示次数越大 |
| user\_id | 用户的编号，Long类型 |
| user\_gender\_id | 用户的预测性别编号，Int类型；0表示女性用户，1表示男性用户，2表示家庭用户 |
| user\_age\_level | 用户的预测年龄等级，Int类型；数值越大表示年龄越大 |
| user\_occupation\_id | 用户的预测职业编号，Int类型 |
| user\_star\_level | 用户的星级编号，Int类型；数值越大表示用户的星级越高 |
| context\_id | 上下文信息的编号，Long类型 |
| context\_timestamp | 广告商品的展示时间，Long类型；取值是以秒为单位的Unix时间戳，以1天为单位对时间戳进行了偏移 |
| context\_page\_id | 广告商品的展示页面编号，Int类型；取值从1开始，依次增加；在一次搜索的展示结果中第一屏的编号为1，第二屏的编号为2 |
| predict\_category\_property | 根据查询词预测的类目属性列表，String类型；数据拼接格式为 “category\_A:property\_A\_1,property\_A\_2,property\_A\_3;category\_B:-1;category\_C:property\_C\_1,property\_C\_2” ，其中 category\_A、category\_B、category\_C 是预测的三个类目；property\_B 取值为-1，表示预测的第二个类目category\_B没有对应的预测属性 |
| shop\_id | 店铺的编号，Long类型 |
| shop\_review\_num\_level | 店铺的评价数量等级，Int类型；取值从0开始，数值越大表示评价数量越多 |
| shop\_review\_positive\_rate | 店铺的好评率，Double类型；取值在0到1之间，数值越大表示好评率越高 |
| shop\_star\_level | 店铺的星级编号，Int类型；取值从0开始，数值越大表示店铺的星级越高 |
| shop\_score\_service | 店铺的服务态度评分，Double类型；取值在0到1之间，数值越大表示评分越高 |
| shop\_score\_delivery | 店铺的物流服务评分，Double类型；取值在0到1之间，数值越大表示评分越高 |
| shop\_score\_description | 店铺的描述相符评分，Double类型；取值在0到1之间，数值越大表示评分越高 |

# 输入

见“数据格式”；

# 输出

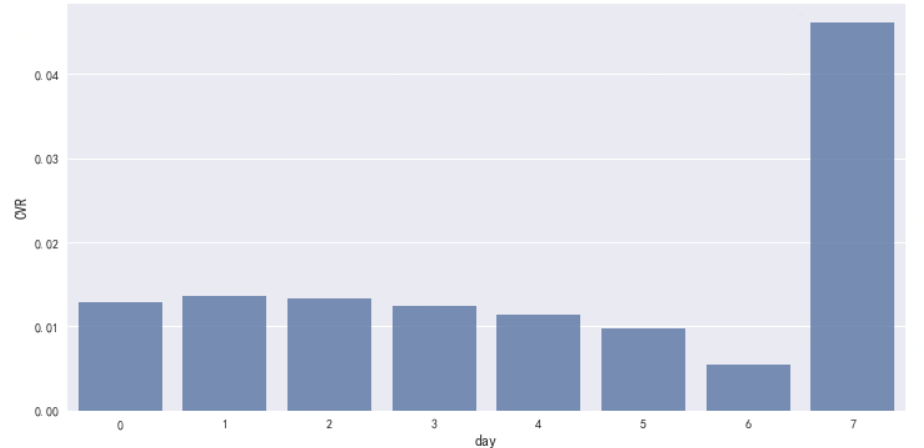
转化概率；

示例：

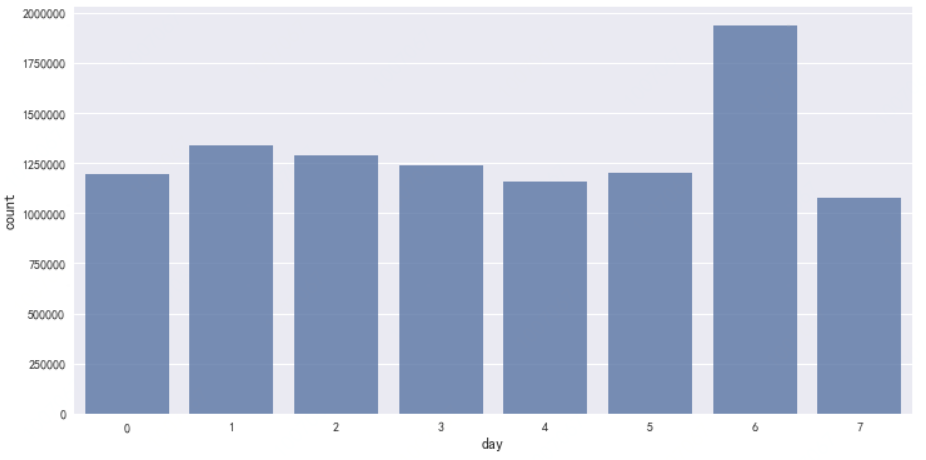
|  |  |
| --- | --- |
| instance\_id | predicted\_score |
| 2475218615076601065 | 0.9 |

# 数据分析

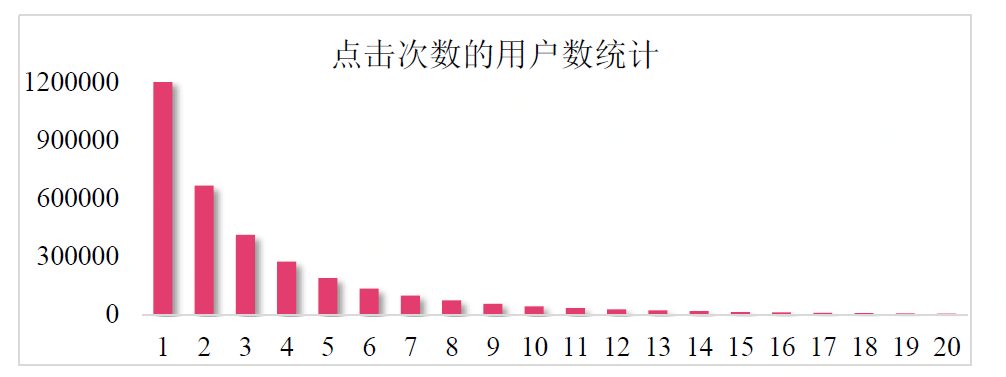
1. 正负样本比例：1:70（151,210 / 10,432,037）；
2. User: 2,958,506；
3. Item: 84,678；
4. 每天的转化率情况；



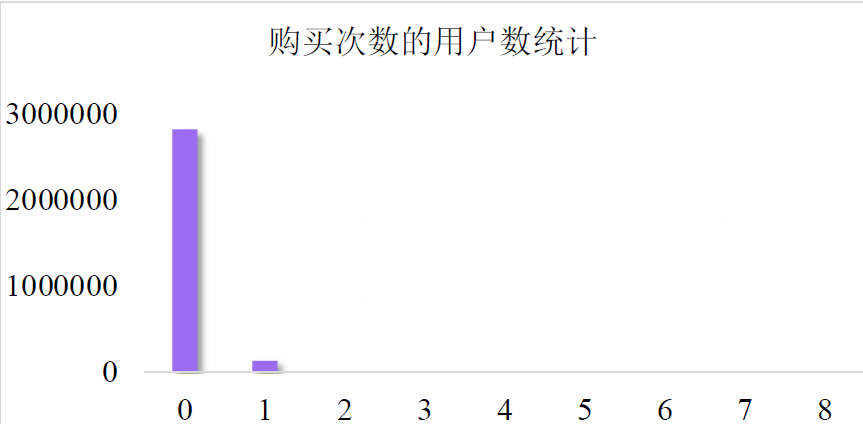
1. 每天的点击次数；



1. 用户点击行为次数分布；



1. 用户购买行为次数分布；



1. 数据情况总结：
2. 发现每个人的平均点击及购买次数较少，所以和通常的根据用户的历史数据预测未来是否购买问题是不一样的，没必要进行滑窗统计user\_id的特征；
3. 是一个低频诉求的场景，及具有长尾分布形式；

# Top1&2团队算法

第一名：lightgbm；

第二名：lightgbm；

第三名：

Lightgbm

Xgboost

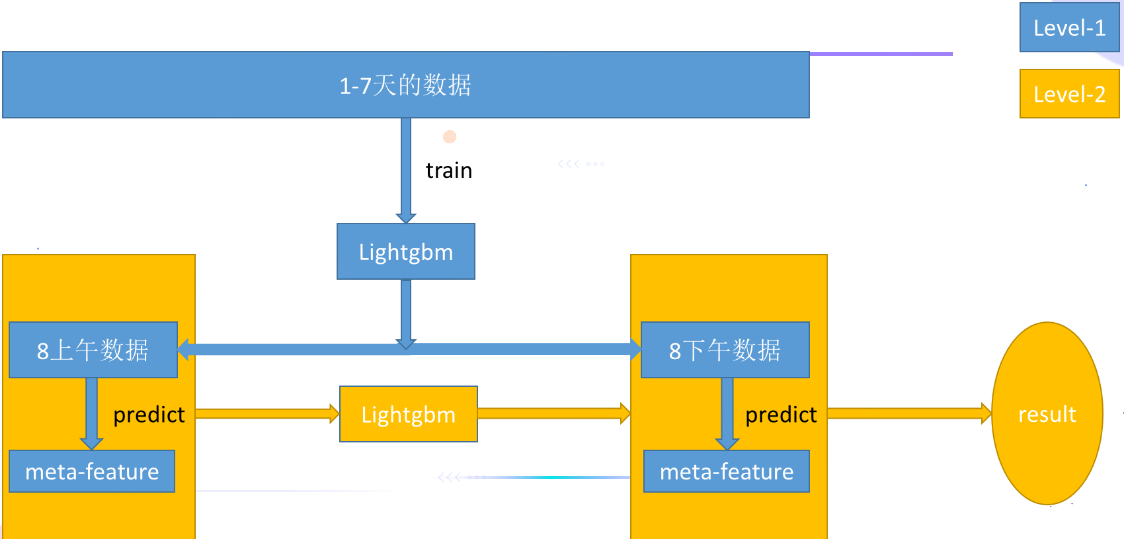
Catboost

GBDT+LR

NN (DeepFFM,DeepFM,FNN)

# Top1团队模型&特征&主代码介绍

## 8.1 Top1模型设计



1. Level-1 将使用预热期的所有数据对这一时期的购物行为建模，并将它在购物节当天的预测作为第二个模型的输入；
2. 单模型：lightgbm；

## 8.2 Top1特征简单介绍

### 8.2.1 Embedding特征

根据原数据提供的“item\_property\_list”和“predict\_category\_property”字段来分析。根据原始数据提供的item特征及搜索item特征，描述user属性偏好，然后统计有这些属性偏好的user和所搜索并点击的item之间的关系。即：用属性来表征user\_id, 用user\_id来表征item\_id；

**（1）Sample Embedding**

sample\_emb\_x=[x1,x2,x3,x4,...,xn]

解释：xn为第n个property在不在predict\_category\_property中；

sample\_emb\_y=[y1,y2,y3,y4,...,yn]

解释：yn为第n个property在不在item\_property\_list中；

**（2）User Embedding**

user\_emb\_x=mean([sample\_emb\_x\_1,sample\_emb\_x\_2,...,sample\_emb\_x\_k])

解释：sample\_emb\_x\_k为该user的第k条样本的sample\_emb；

user\_emb\_y=mean([sample\_emb\_y\_1,sample\_emb\_y\_2,...,sample\_emb\_y\_k])

解释：sample\_emb\_y\_k为该user的第k条样本的sample\_emb；

通过这种对所有样本的sample\_emb做mean操作来对user做embedding。

**（3）Item Embedding**

item\_emb\_x=mean([user\_emb\_x\_1,user\_emb\_x\_2,...,user\_emb\_x\_k])

解释：user\_emb\_x\_k为该item的第k条样本的user\_emb；

item\_emb\_y=mean([user\_emb\_y\_1,user\_emb\_y\_2,...,user\_emb\_y\_k])

解释：user\_emb\_y\_k为该item的第k条样本的user\_emb；

通过这种对所有样本的use\_emb做mean操作来对item做embedding。

该部分得到了6\*n个特征，n的大小视情况而定，这里取count\_top100的property来做embedding，所以总共6\*100个特征。【示例可参考“[参考链接（2）](https://github.com/plantsgo/ijcai-2018/blob/master/eda_solve.ipynb)”】

### 8.2.2 统计特征

|  |
| --- |
| user点击item个数 |
| user最后一次搜索时间 |
| user浏览展示页面的最大页数 |
| user搜索的小时平均 |
| user和item最后一次交互时间 |
| user点击item所属品类的个数 |
| … |

### 8.2.3 时差特征

|  |
| --- |
| user距离上次时长，距离下次时长 |
| user与商品item交互距离上次时长，距离下次时长 |
| user与商品品类item\_category交互距离上次时长，距离下次时长 |
| user与商品品牌item\_brand\_id交互距离上次时长，距离下次时长 |
| … |

### 8.2.4 排序特征

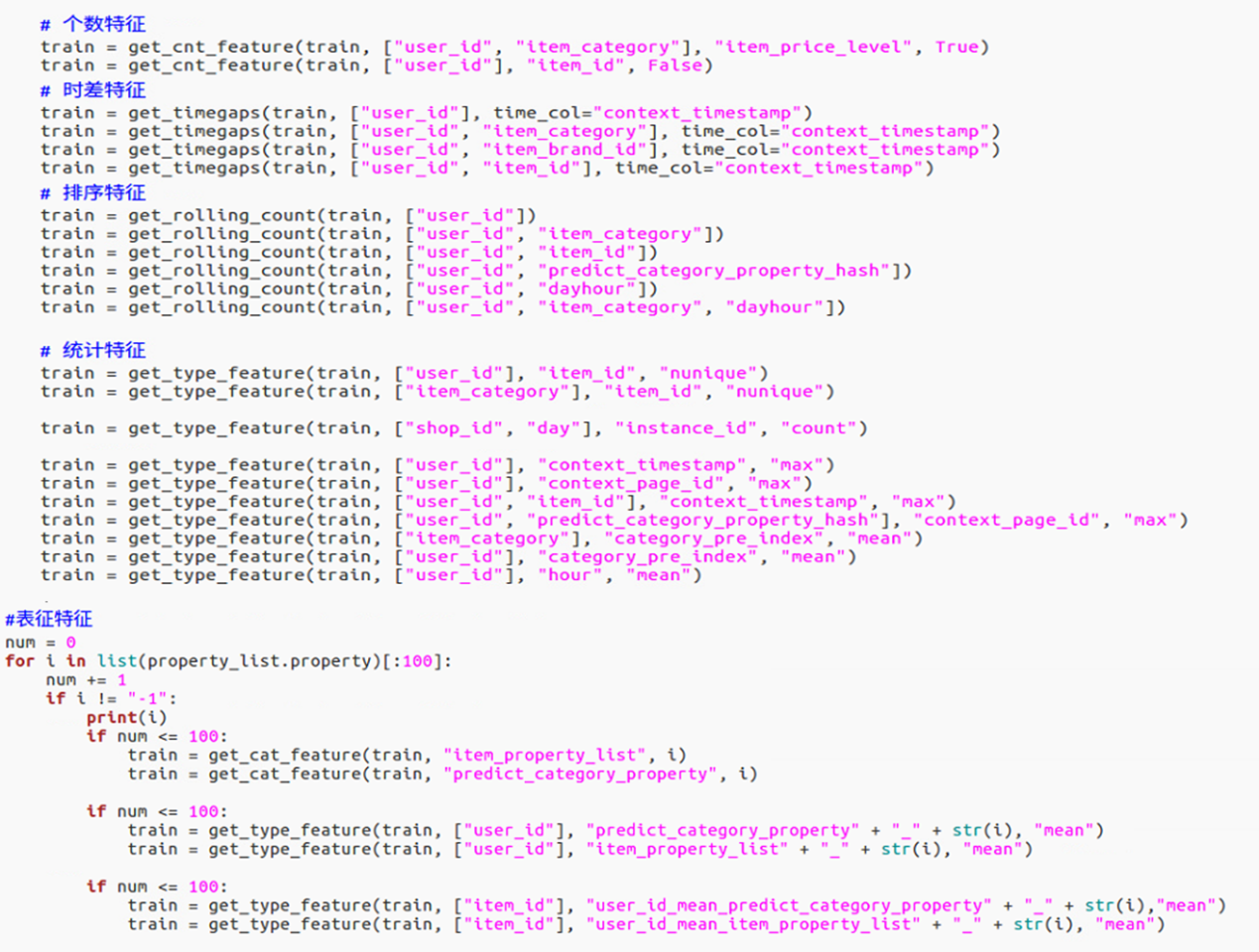
|  |
| --- |
| User的第几次交互，倒数第几次交互 |
| User与商品item的第几次交互，倒数第几次交互 |
| user与商品品类item\_category第几次交互，倒数第几次交互 |
| user与商品品牌item\_brand\_id第几次交互，倒数第几次交互 |
| … |

注：

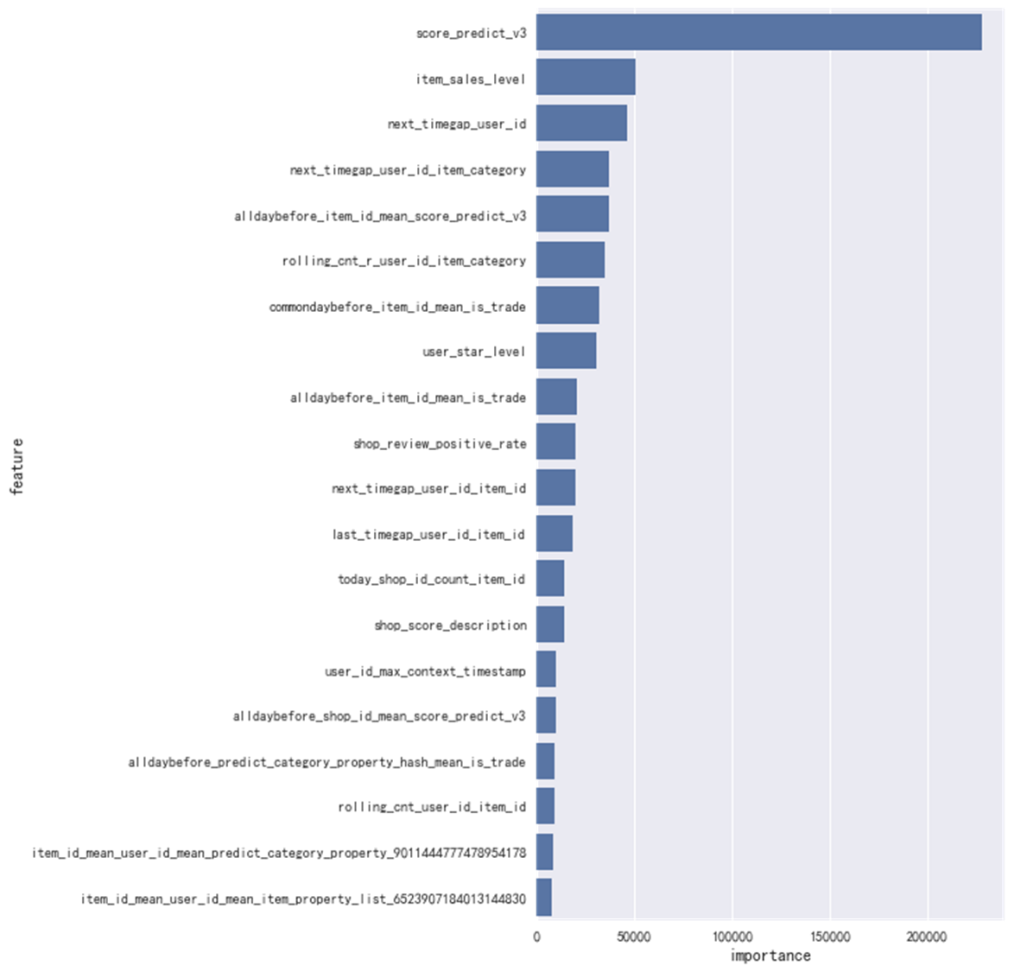
（1）代码中只展示了8.2.1的示例demo；

## 8.3 Top1主代码&特征重要度

### 8.3.1 主代码展示



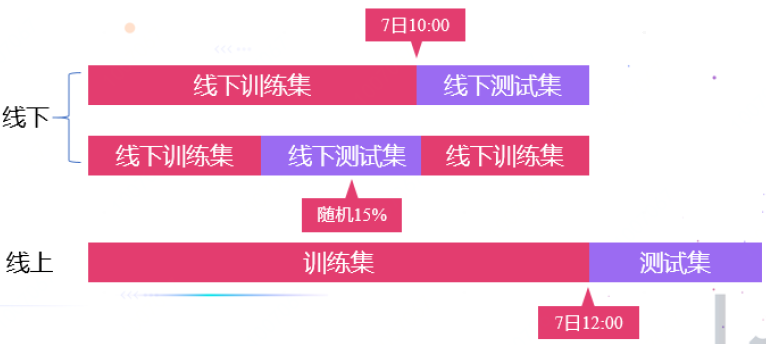
### 8.3.2 特征重要度结果展示



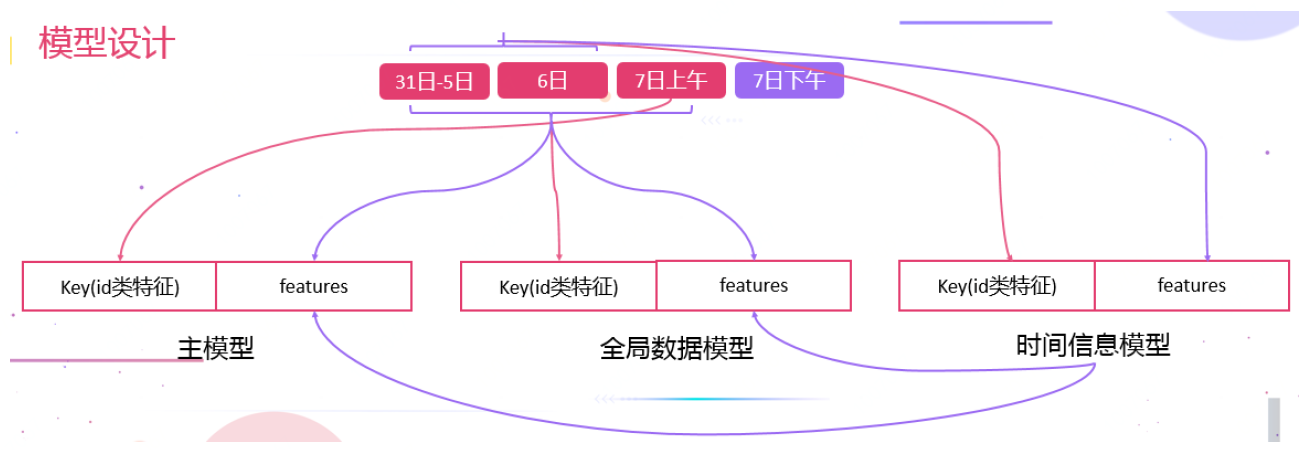
# Top2团队模型&特征介绍

## 9.1 Top2模型设计

### 9.1.1 数据划分



### 9.1.2 模型设计



采用三种方式训练模型：（1）主模型，使用7号上午的数据作为训练样本，对31-5号，6号，7号数据提取特征；（2）全局数据模型，使用全部带标签的样本作为训练数据，使用全部数据提取特征；（3）时间信息模型，使用31-6号的数据作为训练样本，对31-6号的数据提取特征。时间信息模型对7号当天的样本进行预测，将预测结果作为新的特征添加到（1）和（2）模型中，弥补前面模型对时间刻画的缺失。

## 9.2 Top2特征详细介绍

### 9.2.1 第一部分：基础特征群（3+14=17维）

1. **转化时间戳，提取时间特征：天、小时；（3维）**

（1.1）天；（1维）

（1.2）小时维度序号\_1h；（1维）

（1.3）30min维度序号\_30min；（1维）

1. **统计特征：item\_category\_list，item\_property\_list，predict\_category\_property字段的类别与属性；（14维）**

（2.1）item\_category\_list与predict\_category\_property相同的category数量；（1维）

（2.2）item\_property\_list与predict\_category\_property相同的property数量；（1维）

（2.3）item\_category \_list中的第二个类别（下称‘cate’，id类，可理解为2级品类。数据中的item\_category \_list字段第一个类别全部相同）；（1维）

（2.4）item\_property\_list字段，统计属性的数量；（1维）

（2.5）predict\_category\_property字段，统计类别的数量；（1维）

（2.6）predict\_category\_property字段，统计属性的数量；（1维）

（2.7）predict\_category\_property字段，提取第一个类别（下称‘query1’，id类）；（1维）

（2.8）predict\_category\_property字段，提取全部类别（下称‘query’，以‘\_’连接，形成id类特征）；（1维）

（2.9）item\_property\_list字段，提取top1属性；（解释：使用全部点击数据，统计property 出现的次数，根据次数从大到小对当前样本item\_property\_list字段中的property 排序，提取top1 的property， id类）（1维）

（2.10）item\_property\_list字段，提取top2属性-top5属性，top10属性；（解释：使用全部点击数据，统计property 出现的次数，根据次数从大到小对当前样本item\_property\_list字段中的property 排序，取top2/top3/top4/top5/top10 的property，以‘\_’连接，id类）（5维）

注：

（1）item\_category\_list，item\_property\_list，predict\_category\_property信息见“数据格式表”；

（2）后续代码中使用query指代“查询操作后的预测类别”，有3种情况，分别为：

（1.1）query：predict\_category\_property 字段的全部类别，以‘\_’连接，id类特征；

（1.2）query1：predict\_category\_property 字段的第一个类别， id类特征；

（1.3）predict\_category\_property：predict\_category\_property字段本身，id类特征；

（3）top1示例：2636395404473730413；

（4）top2示例：2636395404473730413\_9148482949976129397；

（5）将id类特征转化为序号，使用的是LabelEncoder().fit\_transform函数。

### 9.2.2 第二部分：查询交互，用户交互，竞争特征（60+52+16=128维）

1. **查询交互特征；【时间泄露】（同一user，当前样本的时间戳，60维）**

（1.1）当前样本之前（全部数据开始时间戳-当前时间戳） / 之后（当前时间戳-全部数据结束时间戳），该user点击相同的query / query1 / predict\_category\_property次数；（2\*3=6维）

（1.2）当前样本之前 / 之后，该user点击相同的query / query1 / predict\_category\_property、相同item / shop / brand / city / item\_category\_list / context\_page\_id 的次数；（2\*3\*6=36维）

（1.3）当前样本之前 / 之后，该user点击了除当前query / query1 / predict\_category\_property之外的unique(query) / unique(query1) / unique(predict\_category\_property) 数量；（2\*3=6维）

（1.4）查询该user在当前时间戳的query / query1 / predict\_category\_property在当天数据中第一次及最后一次出现的时间戳，统计user在最小时间戳之前 / 最大时间戳之后点击unique(shop) / unique(item)数量；（3\*2\*2=12维）

注：

（1）item / shop / brand / city / item\_category\_list / context\_page\_id信息见“数据格式表”；

1. **用户交互特征；【时间泄露】（同一user，当天，52维）**

（2.1）user上下两次点击之间，最大时间间隔（当天，时间单位：秒）；（1维）

（2.2）user上下两次点击之间，最小时间间隔（当天，时间单位：秒）；（1维）

（2.3）user上下两次点击之间，平均时间间隔（当天，时间单位：秒）；（1维）

（2.4）user上下两次点击之间，中值时间间隔（当天，时间单位：秒）；（1维）

（2.5）当前样本时间戳，距离当天第一次点击时间间隔（时间单位：秒）；（1维）

（2.6）当前样本时间戳，距离当天最后一次点击时间间隔（时间单位：秒）；（1维）

（2.7）当前样本时间戳，距离当天上次点击时间间隔（时间单位：秒）；（1维）

（2.8）当前样本时间戳，距离当天下次点击时间间隔（时间单位：秒）；（1维）

（2.9）当前样本时间戳，user之前 / 之后点击unique(query) / unique(query1) / unique(predict\_category\_property)的占比（前 / 后占比相加等于1）；（2\*3=6维）

（2.10）当前样本时间戳，user之前 / 之后点击unique(item) / unique(shop) / unique(brand) / unique(city) 的占比（前 / 后占比相加等于1）；（2\*4=8维）

（2.11）当前样本时间戳，user之前 / 之后点击与该样本相同的item / shop / brand / city / item\_category\_list的unique(query) / unique(query1) / unique(predict\_category\_property)的数量；（2\*5\*3=30维）

1. **竞争特征；（同一user，当天，16维）**

（3.1）当前样本时间戳，user之前 / 之后点击了价格更低的unique(item)数量；（2维）

（3.2）当前样本时间戳，user之前 / 之后点击了销量更高的unique(item)数量；（2维）

（3.3）当前样本时间戳，user之前 / 之后点击了评价数量更多的unique(shop)数量；（2维）

（3.4）当前样本时间戳，user之前 / 之后点击了好评率更高的unique(shop)数量；（2维）

（3.5）当前样本时间戳，user之前 / 之后点击了星级更高的unique(shop)数量；（2维）

（3.6）当前样本时间戳，user之前 / 之后点击了服务态度更高的unique(shop)数量；（2维）

（3.7）当前样本时间戳，user之前 / 之后点击了物流更好的unique(shop)数量；（2维）

（3.8）当前样本时间戳，user之前 / 之后点击了描述平均分更高的unique(shop)数量；（2维）

### 9.2.3 第三部分：统计转化率特征，转化率排名特征（149+63=212维）

1. **统计类特征；（10+36+2+10+36+10+45 = 149维）**

**（1.1）前6天（包括第6天）的统计单特征，作为第7天的特征；（10维）**

（1.1.1）前6天（包括第6天）user的总转化率；（平滑处理）（1维）

（1.1.2）前6天（包括第6天）item / brand / shop / city / item\_category\_list / context\_page\_id / query / query1 / predict\_category\_property的总转化率；（9维）

**（1.2）前6天（包括第6天）的统计交叉特征，作为第7天的特征；（36维）**

id 特征列表如下：

['item',

'brand',

'shop',

'item\_category\_list',

'city',

'predict\_category\_property',

'query1',

'query',

'context\_page\_id']

（1.2.1）前6天（包括第6天），item与其之后8个id特征，交叉后的转化率；（8维）

（1.2.2）前6天（包括第6天），brand与其之后7个id特征，交叉后的转化率；（7维）

（1.2.3）前6天（包括第6天），shop与其之后6个id特征，交叉后的转化率；（6维）

（1.2.4）前6天（包括第6天），item\_category\_list与其之后5个id特征，交叉后的转化率；（5维）

（1.2.5）前6天（包括第6天），city与其之后4个id特征，交叉后的转化率；（4维）

（1.2.6）前6天（包括第6天），predict\_category\_property与其之后3个id特征，交叉后的转化率；（3维）

（1.2.7）前6天（包括第6天），query1与其之后2个id特征，交叉后的转化率；（2维）

（1.2.8）前6天（包括第6天），query与其之后1个id特征，交叉后的转化率；（1维）

**（1.3）前6天和第6天的统计特征（用户行为编码），作为第7天的特征；（2维）**

（1.3.1）前6天（包括第6天），user的‘点击次数\_购买次数’（id类特征）；（1维）

（1.3.2）第6天，user的‘点击次数\_购买次数’（id类特征）；（1维）

**（1.4）第6天的统计单特征，作为第7天的特征；（10维）**

（1.4.1）第6天，user的总转化率；（平滑处理）（1维）

（1.4.2）第6天，item / brand / shop / city / item\_category\_list / context\_page\_id / query / query1 / predict\_category\_property的总转化率；（9维）

**（1.5）第6天的统计交叉特征，作为第7天的特征；（36维）**

id 特征列表如下：

['item',

'brand',

'shop',

'item\_category\_list',

'city',

'predict\_category\_property',

'query1',

'query',

'context\_page\_id']

（1.5.1）第6天，item与其之后8个id特征，交叉后的转化率；（8维）

（1.5.2）第6天，brand与其之后7个id特征，交叉后的转化率；（7维）

（1.5.3）第6天，shop与其之后6个id特征，交叉后的转化率；（6维）

（1.5.4）第6天，item\_category\_list与其之后5个id特征，交叉后的转化率；（5维）

（1.5.5）第6天，city与其之后4个id特征，交叉后的转化率；（4维）

（1.5.6）第6天，predict\_category\_property与其之后3个id特征，交叉后的转化率；（3维）

（1.5.7）第6天，query1与其之后2个id特征，交叉后的转化率；（2维）

（1.5.8）第6天，query与其之后1个id特征，交叉后的转化率；（1维）

**（1.6）第7天上午的统计单特征，作为第7天下午的特征；（存在过拟合现象）（10维）**

（1.6.1）第7天上午，user的转化率；（1维）

（1.6.2）第7天上午，item / brand / shop / city / item\_category\_list / context\_page\_id / query / query1 / predict\_category\_property的交易率（mean）；（9维）

**（1.7）第7天上午的统计交叉特征，作为第7天下午的特征；（存在过拟合现象）（45维）**

id 特征列表如下：

['user',

'item',

'brand',

'shop',

'item\_category\_list',

'city',

'predict\_category\_property',

'query1',

'query',

'context\_page\_id']

（1.7.1）第7天上午，user与其之后9个id特征，交叉后的交易率（mean）；（9维）

（1.7.2）第7天上午，item与其之后8个id特征，交叉后的交易率（mean）；（8维）

（1.7.3）第7天上午，brand与其之后7个id特征，交叉后的交易率（mean）；（7维）

（1.7.4）第7天上午，shop与其之后6个id特征，交叉后的交易率（mean）；（6维）

（1.7.5）第7天上午，item\_category\_list与其之后5个id特征，交叉后的交易率（mean）；（5维）

（1.7.6）第7天上午，city与其之后4个id特征，交叉后的交易率（mean）；（4维）

（1.7.7）第7天上午，predict\_category\_property与其之后3个id特征，交叉后的交易率（mean）；（3维）

（1.7.8）第7天上午，query1与其之后2个id特征，交叉后的转化率；（2维）

（1.7.9）第7天上午，query与其之后1个id特征，交叉后的转化率；（1维）

1. **排名特征；（12+18+15+12+6=63维）**

（2.1）前6天 / 第6天 / 第7天上午，user转化率在brand / shop / item\_category\_list / city下面的排名；（3\*4=12维）

（2.2）前6天 / 第6天 / 第7天上午，item转化率在brand / shop / item\_category\_list / city / query1 / query下面的排名；（3\*6=18维）

（2.3）前6天 / 第6天 / 第7天上午，shop转化率在brand / item\_category\_list / city / query1 / query下面的排名；（3\*5=15维）

（2.4）前6天 / 第6天 / 第7天上午，brand转化率在shop / city / query1 / query下面的排名；（3\*4=12维）

（2.5）前6天 / 第6天 / 第7天上午，item\_category\_list转化率在query1 / query下面的排名；（3\*2=6维）

### 9.2.4 第四部分：统计点击特征，点击数量占比特征（36+198+48=282维）

**（1）前6天 / 第6天 / all天，统计单id类点击数特征；（36维）**

（1.1）前6天 / 第6天 / all days，user / item / brand / shop / item\_category\_list / city / cate/ top10\_property / predict\_category\_property / context\_page\_id / query / query1点击数；（3\*12=36维）

**（2）前6天 / 第6天 / all天，统计id特征交叉特征的点击数量（3\* 11\*12/2 =198维）**

id列表如下：

['user',

'item',

'brand',

'shop',

'item\_category\_list',

'city',

'cate',

'top10\_property',

'predict\_category\_property',

'context\_page\_id',

'query1',

'query']

（2.1）前6天 / 第6天 / all days，user与其之后11个id特征，交叉后的点击数量；（11维）

（2.2）前6天 / 第6天 / all days，item与其之后10个id特征，交叉后的点击数量；（10维）

（2.3）前6天 / 第6天 / all days，brand与其之后9个id特征，交叉后的点击数量；（9维）

（2.4）前6天 / 第6天 / all days，shop与其之后8个id特征，交叉后的点击数量；（8维）

（2.5）前6天 / 第6天 / all days，item\_category\_list与其之后7个id特征，交叉后的点击数量；（7维）

（2.6）前6天 / 第6天 / all days，city与其之后6个id特征，交叉后的点击数量；（6维）

（2.7）前6天 / 第6天 / all days，cate与其之后5个id特征，交叉后的点击数量；（5维）

（2.8）前6天 / 第6天 / all days，predict\_category\_property与其之后4个id特征，交叉后的点击数量；（4维）

（2.9）前6天 / 第6天 / all days，context\_page\_id与其之后3个id特征，交叉后的点击数量；（3维）

（2.10）前6天 / 第6天 / all days，query1与其之后2个id特征，交叉后的点击数量；（2维）

（2.11）前6天 / 第6天 / all days，query与其之后1个id特征，交叉后的点击数量；（1维）

**（3）前6天 / 第6天 / all天，分别统计指定的16个交叉点击数量，计算占比特征；（3\*16=48维）**

**（说明：自己指定，未全部列举）**

[count(['user','query']) / count(query),

count(['user','query1']) / count(query1),

count(['user','shop']) / count(shop),

count(['user','item']) / count(item),

count(['item','shop']) / count(shop),

count(['item', 'brand']) / count(brand),

count(['item','city']) / count(city),

count(['item','cate']) / count(cate),

count(['item','top10\_property']) / count(top10\_property),

count(['item','context\_page\_id']) / count(context\_page\_id),

count(['item','query1']) / count(query1),

count(['item','item\_category\_list']) / count(item\_category\_list),

count(['item','query']) / count(query),

count(['brand', 'shop']) / count(shop),

count(['shop','city']) / count(city),

count([ 'shop','context\_page\_id']) / count(context\_page\_id)]

### 9.2.5 第五部分：不购买\_购买特征，趋势特征，一次性购买特征，出现\_购买特征，item\_shop属性变化特征（7+225+12+3+40=287维）

**（1）不购买\_购买特征；（7维）**

（1.1）统计user前6天连续点击行为（未购买）最大次数；

（1.2）统计user第6天连续点击行为（未购买）最大次数；

（1.3）统计user第6天购买的unique(item)个数；

（1.4）统计user第6天点击未购买unique(item)个数；

（1.5）统计user第6天购买item所属的unique(shop)个数；

（1.6）统计user第6天点击未购买item所属的unique(shop)个数；

（1.7）计算user第6天 (1.4) / (1.6)；

**（2）趋势特征；（54+63+108=225维）**

（2.1）统计过去第1,2,3,4,5,7天（共6天，第6天的点击数据已经被统计）item / brand / shop / item\_category\_list / city / predict\_category\_property / context\_page\_id / query1 / query点击数量；（6\*9=54维）

（2.2）统计过去每天（共7天）item / brand / shop / item\_category\_list / city / predict\_category\_property / context\_page\_id / query1 / query购买数量；（7\*9=63维）

（2.3）统计过去6天的item / brand / shop / item\_category\_list / city / predict\_category\_property / context\_page\_id / query1 / query，点击 / 购买趋势（后一天的数据除以当前天的数据），比上一天高为1，否则为0；（6 \*9\*2维=108维）；

**（3）一次性购买特征；（12维）**

解释：一次性购买，即转化率=1。

（3.1）前6天 / 第6天 / 第7天上午，[user, item]一次性购买的次数/item购买次数（3\*1维）；

（3.2）前6天 / 第6天 / 第7天上午，[user, shop]一次性购买的次数/shop购买次数（3\*1维）；

（3.3）前6天 / 第6天 / 第7天上午，[user, query]一次性购买的次数/query购买次数（3\*1维）；

（3.4）前6天 / 第6天 / 第7天上午，[user, query1]一次性购买的次数/ query1购买次数（3\*1维）；

**（4）出现\_购买特征；（3维）**

（4.1）第6天，item\_id / query / query1第一次出现到购买的时间差（时间单位：秒）；（3维）

**（5）item\_shop属性变化特征（40维）**

（5.1）item\_price\_level / item\_sales\_level / item\_collected\_level / item\_pv\_level

计算4个属性：前6天的方差、均值；第7天均值与前6天均值的差值；第7天均值与第6天均值的差值；（4\*2+4+4=16维）

（5.2）shop\_review\_num\_level / shop\_review\_positive\_rate / shop\_star\_level / shop\_score\_service / shop\_score\_delivery / shop\_score\_description

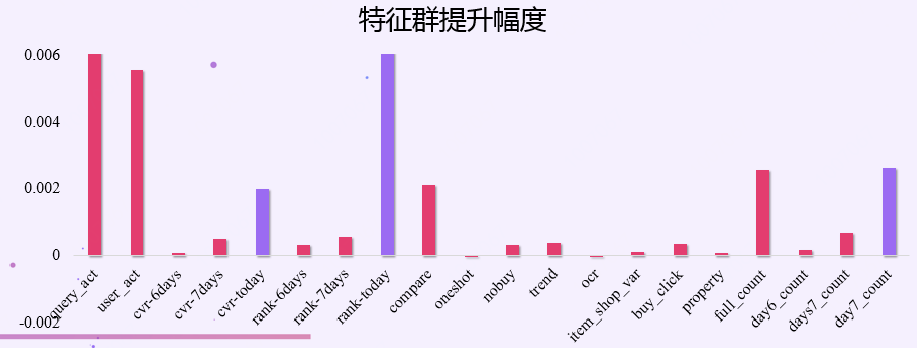
计算6个属性：前6天的方差 / 均值；第7天均值与前6天均值的差值；第7天均值与第6天均值的差值；（6\*2+6+6=24维）

### 9.2.6 第六部分：强制cross特征（100\*99/2=4950维）

**（1）使用上面特征训练模型，选择top100的特征强制相除，组合成新的特征；**

## 9.3 Top2特征重要度

### 9.3.1 特征群重要度结果展示



query（查询）交互特征；

user交互特征；

当天转化率；

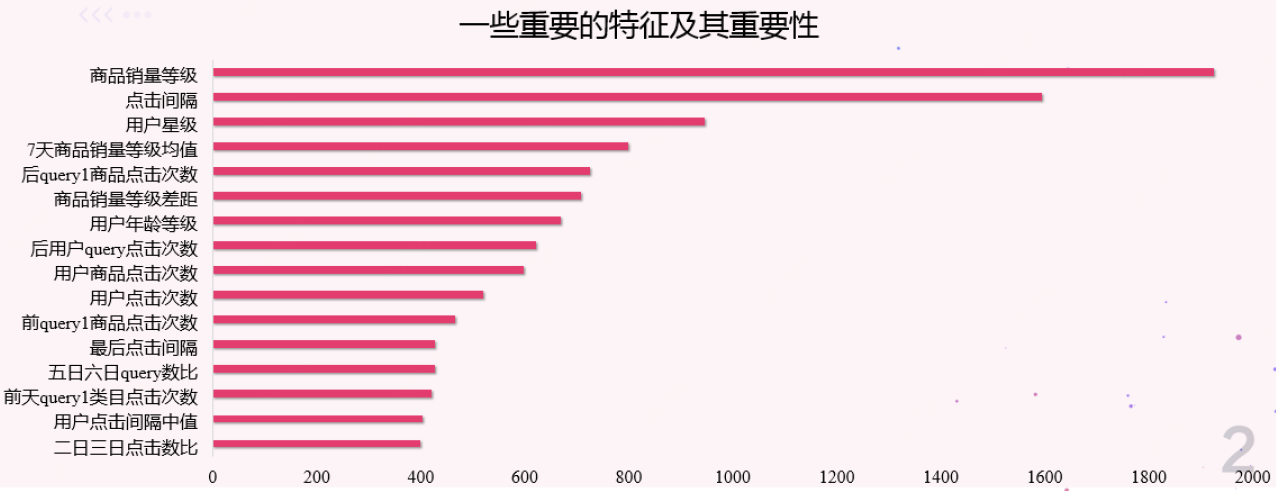
转化率排名特征；

竞争特征；

统计点击特征&占比特征；

第7天点击统计特征；

### 9.3.2 特征重要度结果展示



# 参考链接

（1）官方网址：

<https://tianchi.aliyun.com/competition/information.htm?spm=5176.100067.5678.2.5b0e2163b2Tv1G&raceId=231647>

（2）第一名github地址：<https://github.com/plantsgo/ijcai-2018/blob/master/eda_solve.ipynb>

（3）第二名github地址：<https://github.com/YouChouNoBB/ijcai-18-top2-single-mole-solution>

（4）第二名思路讲解：<https://tianchi.aliyun.com/forum/videoStream.html#postsId=5531>

（5）第三名github地址：<https://github.com/luoda888/2018-IJCAI-top3>