[品类流失预警](https://cf.jd.com/pages/viewpage.action?pageId=437087251)

目标：预测未来7天、15天用户在品类下的流失概率，基于预测概率锁定预流失用户进行营销干预，实现品类用户降流失

样本：进入品类流失时间上下限范围内的用户及对应品类为基础数据池，每日总样本量约2-3亿。其中未来7天、15天内在品类下未下单标记为正样本，正负样本比例为10:1，采样后正负样本比例为2:1。

特征：

1）用户基本特征：包含用户微群类型、生命周期类型、身份类型、京享值、性别，年龄，城市线级，促销敏感度，评价敏感度等

2）用户行为特征：包含用户在品类下搜索、浏览、收藏、加购、下单、咨询等基本行为

3）用户统计特征数据：包含用户在品类下的faster、fast得分，近期（离线1、3、7天，15天，30天，60天，90天）行为次数及时间间隔。

评估指标：

验证集：7天AUC=0.806，15天AUC = 0.809

# [登端模型](https://cf.jd.com/pages/viewpage.action?pageId=435300960&src=contextnavpagetreemode)

预测用户在未来7天内是否登端（app)

# 2. 建模方式

1. 样本
   1. t日的全生命周期营销用户(9亿)为样本，其中在[t+1,t+7]时间段内有登陆的用户为正样本，样本比例为1.1亿 ：7.9亿，生命周期分布如下

|  |  |  |  |
| --- | --- | --- | --- |
| 生命周期 | 人数 | 登陆率 | 人数占比 |
| 潜在期 | 218246130 | 0.029028 | 0.24260852 |
| 引入期 | 68977819 | 0.163726 | 0.076677678 |
| 衰退期 | 50044432 | 0.126746 | 0.055630794 |
| 成熟期 | 144720831 | 0.40823 | 0.160875735 |
| 流失期 | 350307287 | 0.03188 | 0.389411406 |
| 成长期 | 67284976 | 0.251203 | 0.074795867 |

* 1. 训练集验证集比例8：2，最终训练集量级为4.05亿，正负样本比3.5：1

1. 特征
   1. 基础属性
      1. 年龄、性别、生命周期、京享值、城市线级等
   2. 近期行为
      1. 浏览、点击、加购、下单
   3. 近期触达
      1. push发送、成功触达、打开次数
      2. sms发送、打开次数
   4. 近期权益
      1. 红包领、用
      2. 优惠券领、用

# 3.模型效果

1. auc
   1. 整体：0.963233
2. recall, precision, f1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| threshold | recall\_score | precision\_score | accuracy\_score | f1\_score | pred1\_ratio |
| 0 | 1 | 0.219917346 | 0.219917346 | 0.360544666 | 1 |
| 0.1 | 0.965951901 | 0.610471388 | 0.856965505 | 0.748131548 | 0.347976306 |
| 0.2 | 0.938532826 | 0.673373459 | 0.886366085 | 0.78414372 | 0.306515865 |
| 0.3 | 0.903729757 | 0.720065589 | 0.901563594 | 0.801510734 | 0.276010759 |
| 0.4 | 0.861144433 | 0.758302868 | 0.90910116 | 0.80645822 | 0.24974269 |
| 0.5 | 0.806730282 | 0.793525426 | 0.911333684 | 0.800073373 | 0.223576935 |
| 0.6 | 0.731729567 | 0.829951698 | 0.908031872 | 0.777751786 | 0.193890831 |
| 0.7 | 0.626939752 | 0.869620601 | 0.897286439 | 0.728603725 | 0.158546067 |
| 0.8 | 0.484264458 | 0.912303867 | 0.876343566 | 0.632688469 | 0.116735397 |
| 0.9 | 0.293799096 | 0.957292625 | 0.84181168 | 0.449610053 | 0.067494009 |
| 1 | 0 | 0 | 0.780082654 | 0 | 0 |

## 4. 模型更新

### 1. V0.2

1. 样本  
   t日的老用户(3.1亿)为样本，样本比例为0.82亿 ：2.3亿，训练集：验证集=8：2，最终训练集量级为2.5亿，未进行采样
2. 效果
   1. 分数段分布

|  |  |
| --- | --- |
| 分数分段 | 人数占比 |
| 0 | 48.91% |
| 1 | 7.93% |
| 2 | 5.95% |
| 3 | 5.31% |
| 4 | 5.06% |
| 5 | 5.48% |
| 6 | 5.36% |
| 7 | 5.11% |
| 8 | 4.94% |
| 9 | 5.96% |

* 1. auc: 0.928
  2. recall, precision,f1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| threshold | recall\_score | precision\_score | accuracy\_score | f1\_score | pred1\_ratio |
| 0 | 1 | 0.28219043 | 0.28219043 | 0.440169297 | 1 |
| 0.1 | 0.971525444 | 0.5333956 | 0.752138975 | 0.688683834 | 0.513980961 |
| 0.2 | 0.930333419 | 0.603652126 | 0.807967178 | 0.732207351 | 0.434904768 |
| 0.3 | 0.876792058 | 0.65867469 | 0.837017615 | 0.752241281 | 0.375636611 |
| 0.4 | 0.810028331 | 0.708161554 | 0.852191434 | 0.755677438 | 0.322782623 |
| 0.5 | 0.729166009 | 0.755523694 | 0.856991156 | 0.742110888 | 0.272345754 |
| 0.6 | 0.622902754 | 0.807721659 | 0.851742961 | 0.703374053 | 0.217621001 |
| 0.7 | 0.499618169 | 0.860033813 | 0.835852033 | 0.632056621 | 0.163932469 |
| 0.8 | 0.36317395 | 0.9105397 | 0.810224734 | 0.519244336 | 0.112553262 |
| 0.9 | 0.212556465 | 0.958677209 | 0.775205534 | 0.34796308 | 0.062566836 |
| 1 | 0 | 0 | 0.71780957 | 0 | 0 |

## 5. 模型实验效果

1. 结果  
   春节期间使用对该模型进行测试，选取2021-02-08至2021-02-11的空白组用户计算模型效果（未来7天是否登录app端）
   1. 从4天整体效果来看，模型效果不佳，在阈值0.5情况下，f1值为0.38，recall为0.37，precision为0.39，准确率略高于全预测为0，为0.922（在阈值0.9情况下）
   2. 4天auc均值为0.76
2. 原因分析
   1. 训练数据与测试数据分布不一致  
      训练数据为日常时期数据，测试数据为春节时期，**训练模型时因特征原因未采用去年数据进行训练**，致使模型对春节时期数据泛化能力降低
3. 效果明细数据
   1. auc

|  |  |
| --- | --- |
| dt | auc |
| 2021/2/8 | 0.79225 |
| 2021/2/9 | 0.769431 |
| 2021/2/10 | 0.752873 |
| 2021/2/11 | 0.738046 |
|  |  |

# [实时意图识别](https://cf.jd.com/pages/viewpage.action?pageId=435812077)

1. 背景   
   解决在什么时候，给什么样的用户发放权益；圈出由目标决定的人群。如目标是圈出权益发放的人群，模型需要找出那些犹豫不觉用户，再用合适的权益‘临门一脚’促成订单。

此外，意图识别可以设置不同目标，如，不同的行为概率（来访，点击，收藏，加购，购买），或者未来一段时间内的偏好（如，价格偏好，商品偏好等）。

1. 目标  
   预测用户未来n分钟内是否有消费行为。  
     
   目前主要通过实时意图识别来分发用户离站5分钟后的红包权益。  
   首先通过模型预测出未来半小时内的购买意图，再通过筛选购买概率的阈值，判断红包权益是否分发。  
   阈值的设置依据于，有无权益分发后，订单增量的对比。对于高概率和低概率人群，无论是否有权益（好的权益，如红包）发放，对成单都没有太大帮助。此时，对人群的筛选，已经起到优化红包发放的作用，所以在选定阈值后，需要对比随机发放与阈值内发放后，红包使用转化率对比。且需要确认阈值内人群规模，与红包预算的匹配程度。  
   后续，针对红包面额进行优化，根据浏览商品的价格分布，用户使用红包习惯等特征，判断红包的额度。需要根据全天的预算，与时段内的预算预估，合理发放红包。
2. 特征
   * 用户短期（30分钟以内）的行为特征
     + 加购次数，加购商品的价格分布
     + 浏览商详次数，浏览商品（三级）品类统计值（如三级品类数量，及每个品类浏览次数），浏览商品价格分布，浏览停留时间。
     + 搜索次数，搜索关键词
     + 浏览评论次数与停留时间（暂无）
     + 下单商品数量，金额。
   * 时间类特征
     + 周几，时段，日期等。
   * 用户近3日的行为特征
     + 加购次数，加购商品的价格分布
     + 浏览商详次数，浏览商品（三级）品类统计值（如三级品类数量，及每个品类浏览次数），浏览商品价格分布，浏览停留时间。
     + 搜索次数，搜索关键词
     + 浏览评论次数与停留时间（暂无）
     + 下单商品数量，金额。
   * 用户画像 - 全用户档案
     + 年龄，性别和活跃度等。
3. 模型

用户意图模型：用户来访，点击，收藏，加购，**购买**，价格偏好，商品偏好等意图模型。

版本一：规则参数模型（近似lr）预计完成时间2021年2月底  
根据经验与线下的数据探索，建立lr模型。根据特征权重，进行实时计算。

版本二：FTRL（alink） + mckp

1. 评估

购买意图：auc  
针对与离站人群，发放权益后，可以通过ab test观察30分钟内重登，下单人数，红包使用率等指标。还需要针对红包有效时长，统计红包转化率。

1. 任务计划