中国大数据算法大赛-用户购买时间预测

队伍名称：WhyK

演讲者：邱昱（yuna_qiu）

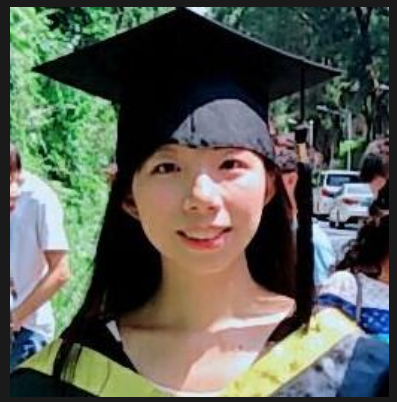
2018.07.19

JDATA

目录

JDATA

团队介绍



邱昱
WhyK(yuna_qiu)
研一学生

黄志炜
WhyK(zhazhawong)
开发工程师

赵银湖
WhyK(lake)
研一学生

容汉铿
WhyK(kengkeng)
研一学生

目录

JDATA

赛题描述&问题解读

➢ 赛题描述

提供T时间点前3个月在目标品类有购买的用户集合，预测T时间点后一个月内这些用户是否购买及首次购买时间。

- A榜T：2017-05-01
- B榜T：2017-09-01

➢ 评价指标

$$S_1 = \frac{\sum_{i=1}^{N} w_i o_i}{\sum_{i=1}^{N} w_i}$$

$$w_i = \frac{1}{1 + ln(i)}$$

$$S_2 = \frac{\sum_{u \in U_r} f(u)}{|U_r|}$$

$$f(u) = \begin{cases} 0, u \notin U_r \\ \frac{10}{10 + d_u{}^2}, u \in U_r \end{cases}$$

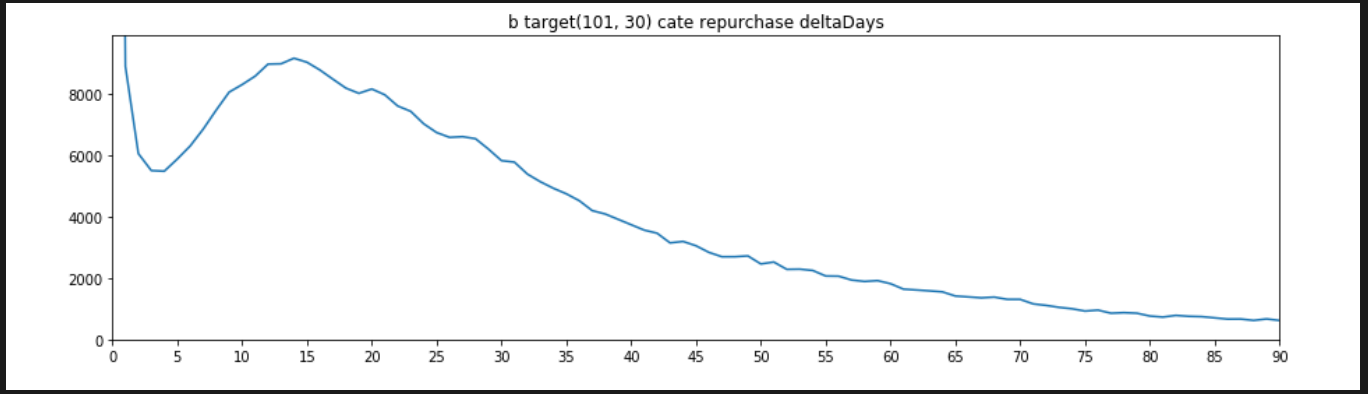
$$S = 0.4 \times S1 + 0.6 \times S2$$

➢ 问题解读

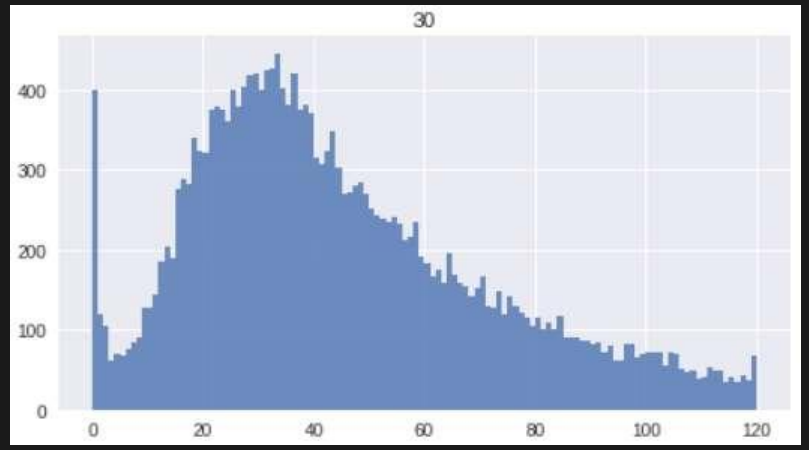
1. 在考察时间段内，目标用户是否会发生下单行为（二分类问题）
2. 在1中发生购买的用户，对目标品类商品的首次下单时间（回归问题）

数据探索--用户回购时间

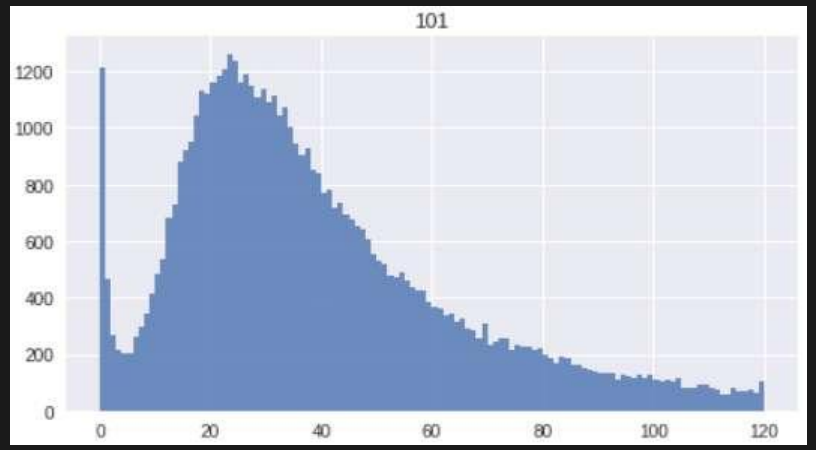
用户回购间隔天数



30类目用户平均回购时间间隔



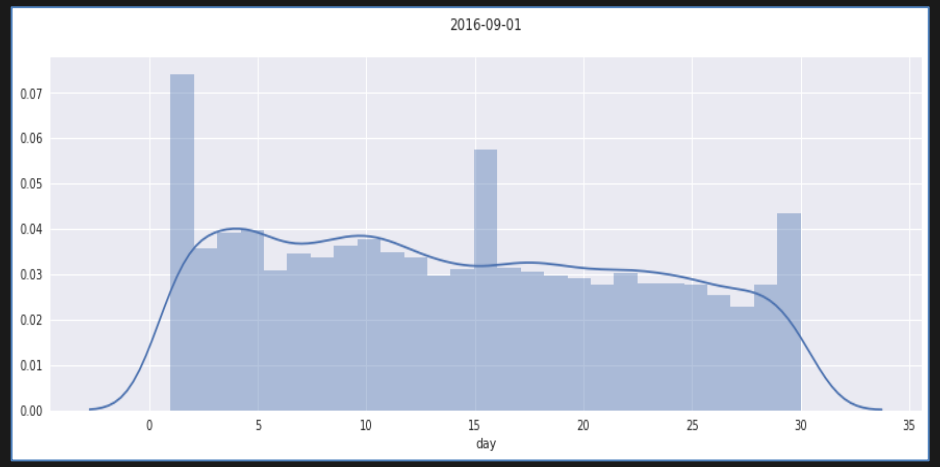
101类目用户平均回购时间间隔



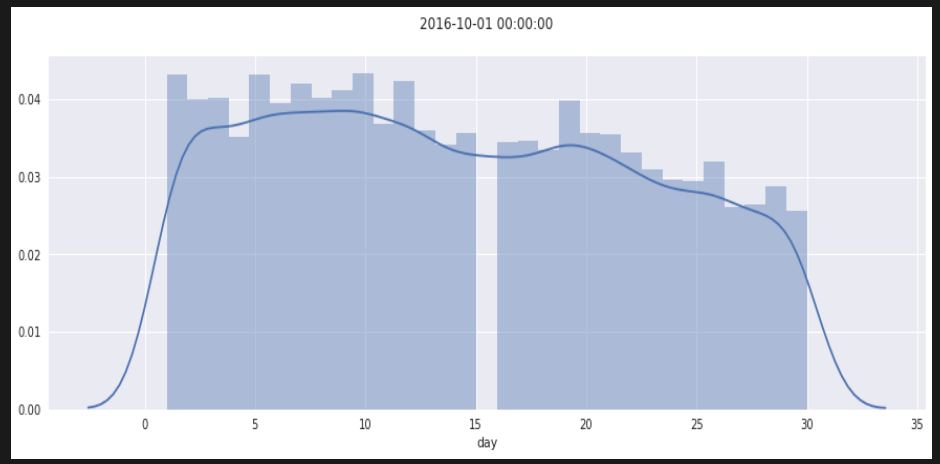
JDATA

数据探索--部分月份首次购买日期统计

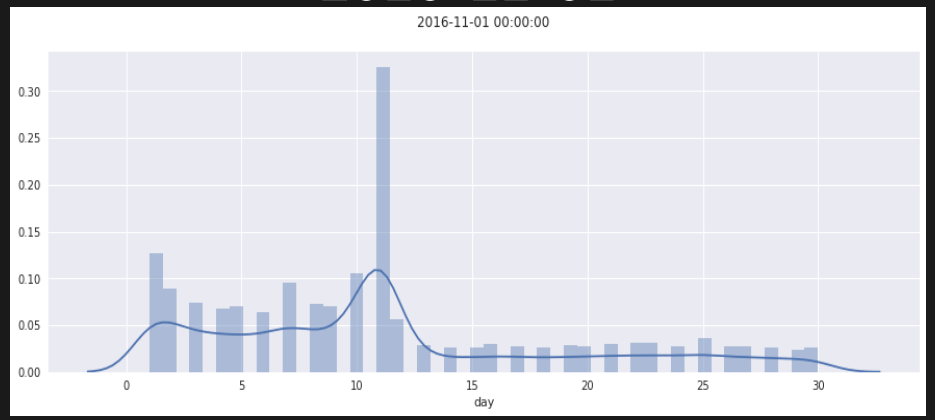
2016-09-01



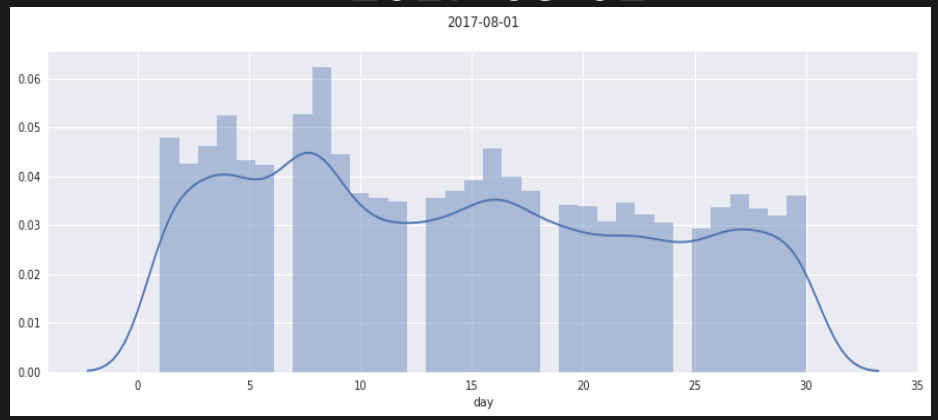
2016-10-01



2016-11-01



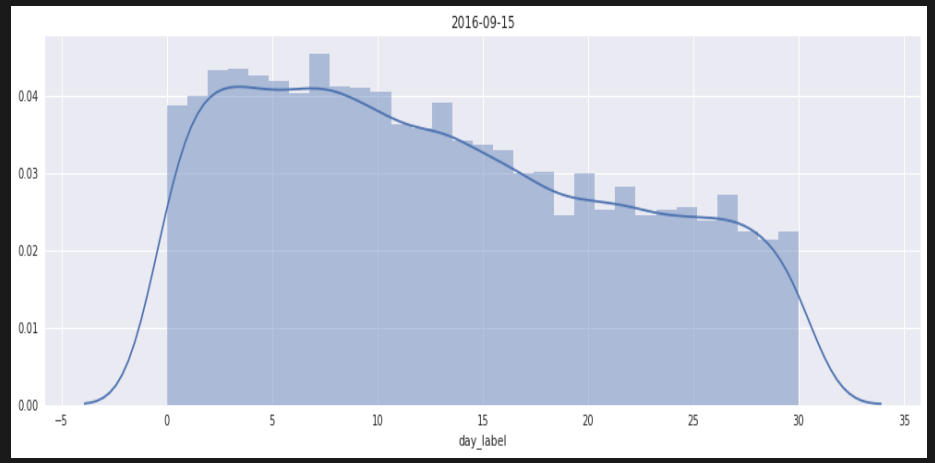
2017-08-01



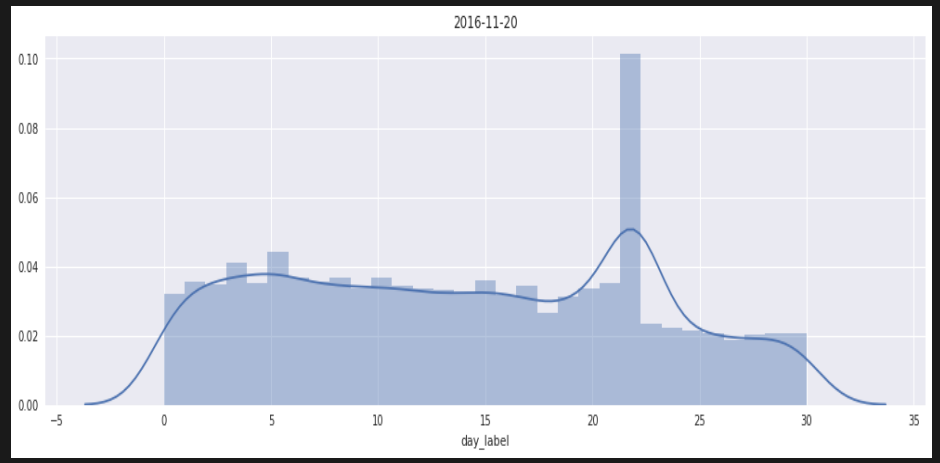
JDATA

数据探索--部分日期未来30天首次购买日统计

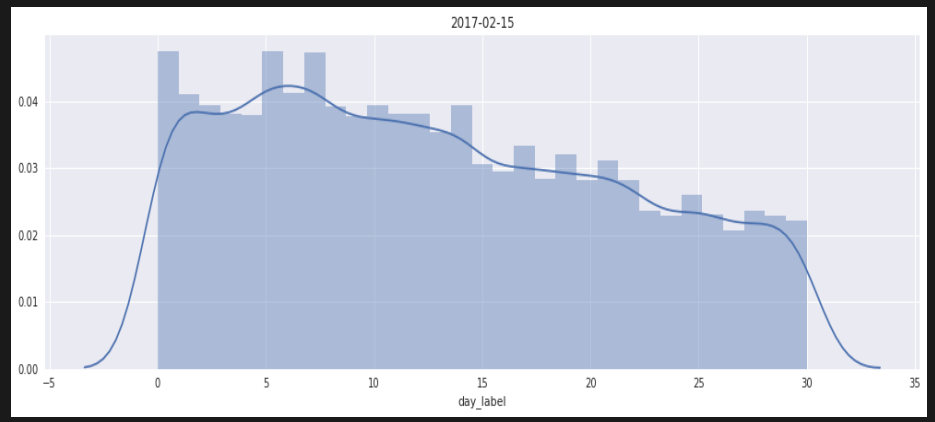
2016-09-15



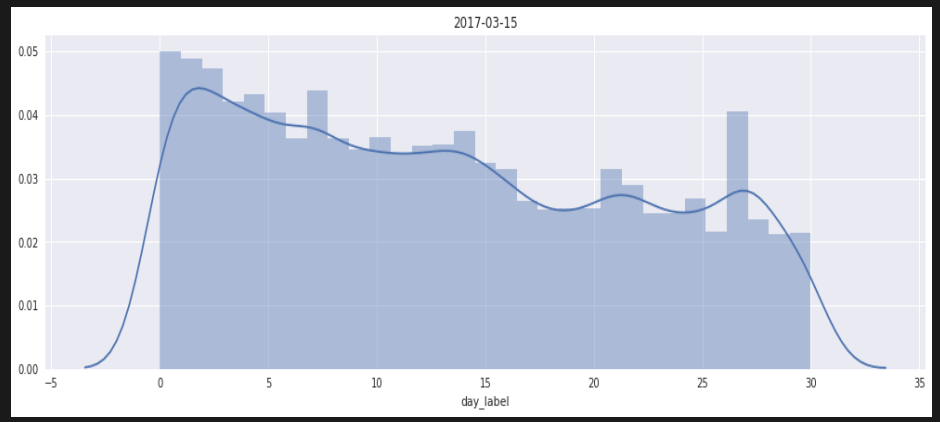
2016-11-20



2017-02-15



2017-03-15



JDATA

数据集构造--用户模型

- 方案一：滑窗倍增

线下验证集：

线上测试集：

- 方案二：避免数据穿越

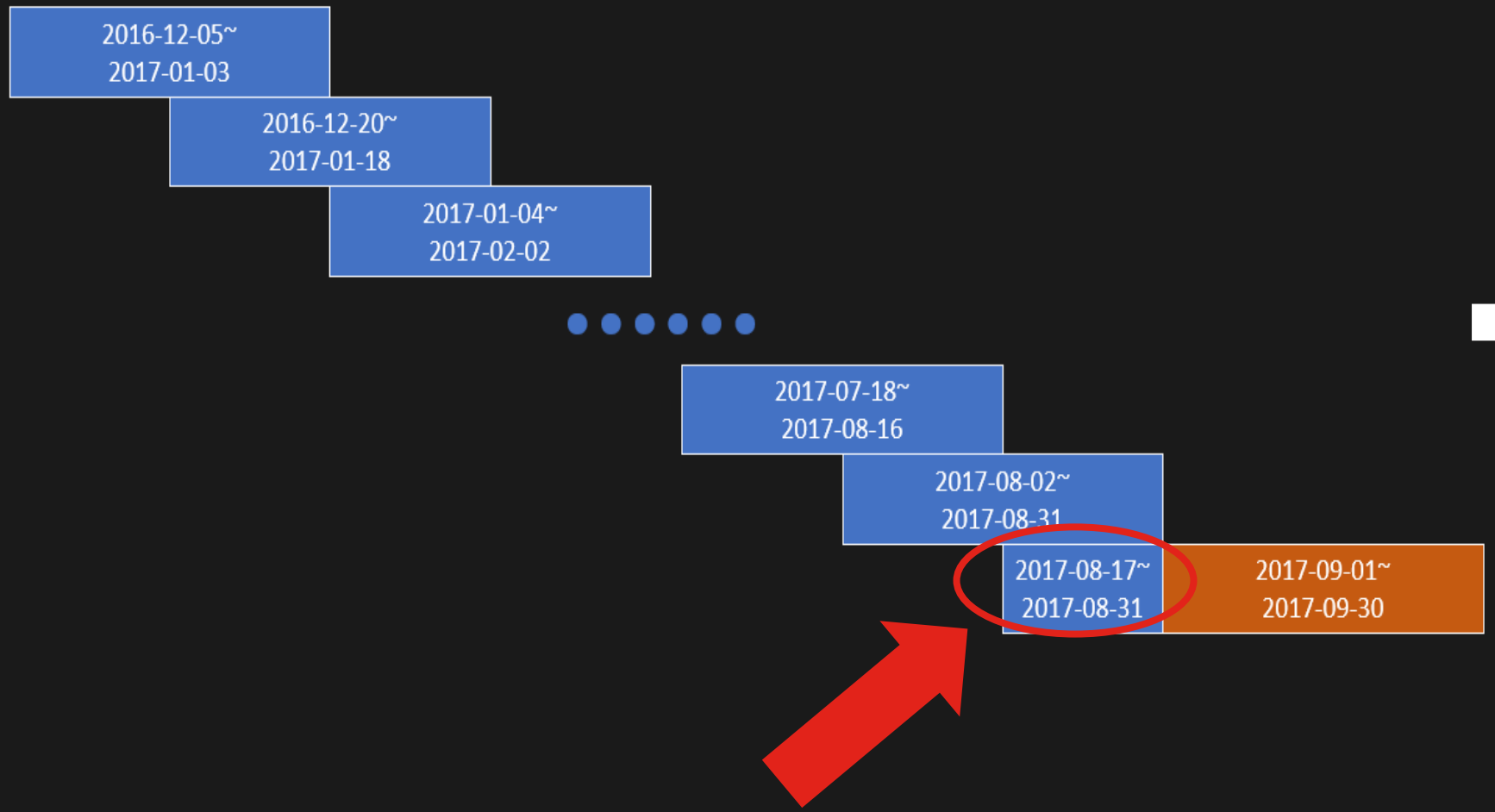
训练集：

测试集：

JDATA

数据集构造--日期模型

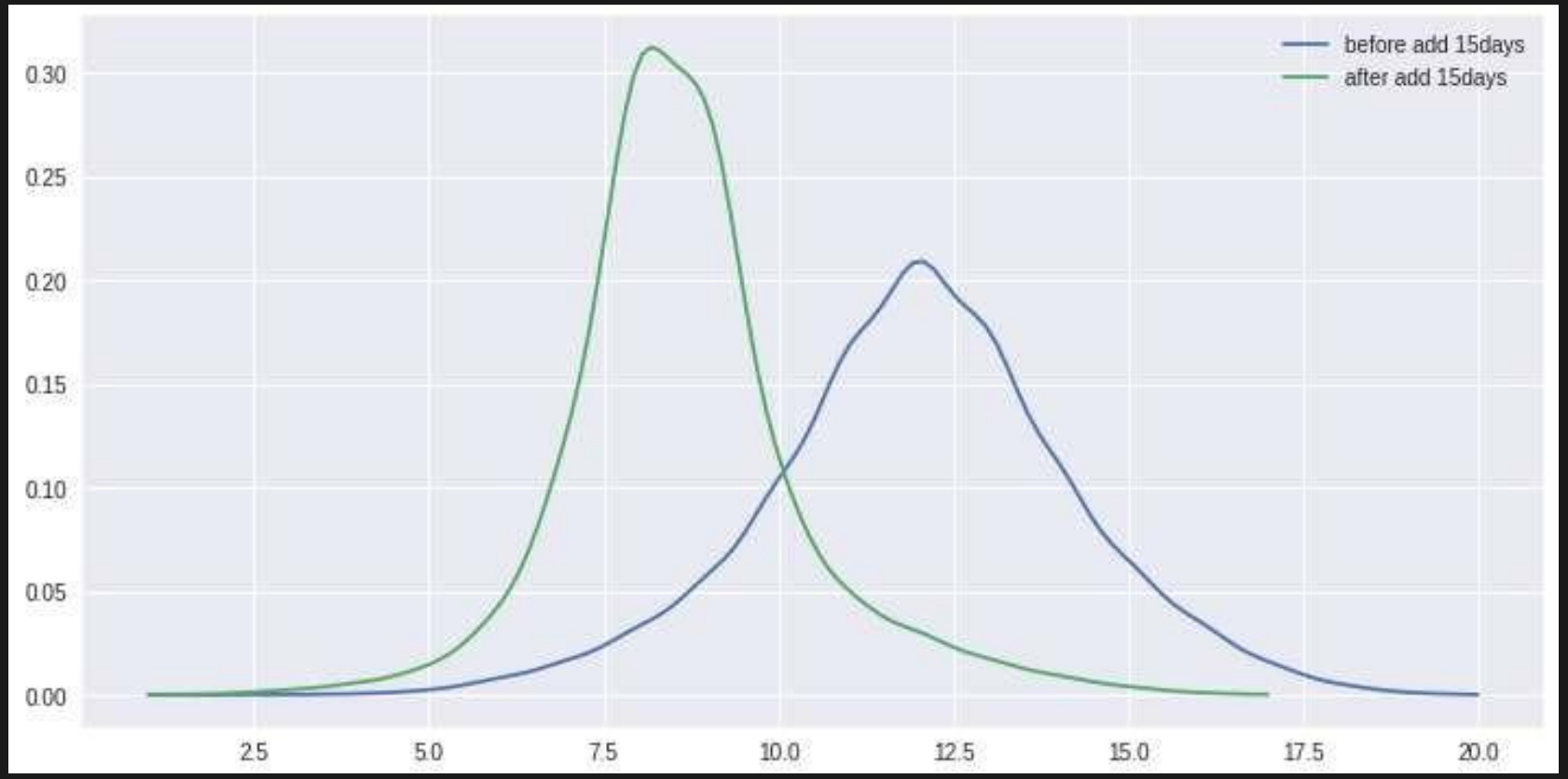
15天滑窗间隔倍增



+ 日期序列特征

JDATA

数据集构造—日期模型添加15天数据集前后对比



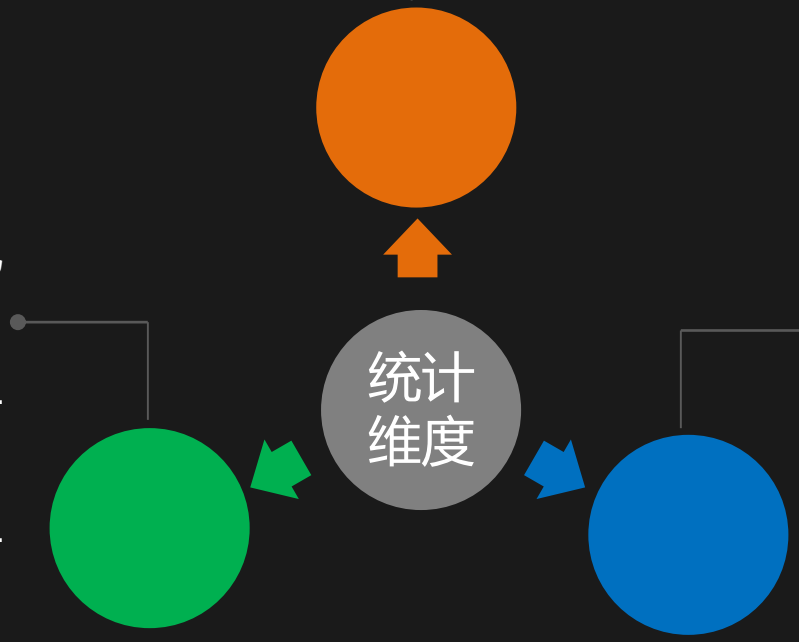
特征工程--统计方式

历史统计特征

- 用户目标类目历史平均每个月购买的天数
- 用户目标类目历史平均每个月购买数量

滑窗统计特征(15天/1个月/3个月)

- 统计区间101、30类别评价数
- 统计区间101、30类别月份、周几偏好性
- 统计区间101、30类别各类别订单数量
- 统计区间101、30类别各类别订单商品数量



最近一次行为特征

- 用户各类别、所有类别的最近一次浏览时间
- 用户各类别、所有类别的最近一次下单时间
- 用户101、30类别最近一次订单的商品参数一均值

特征工程--特征内容

用户基础特征

- 用户性别
- 用户年龄
- 用户等级码

用户行为特征

- 用户对目标品类的关注的次数
- 用户对所有品类的活跃天数
- 用户对所有品类的活动时间间隔的均值、方差

用户订单特征

- 用户对目标品类的订单数量
- 用户对目标品类订单的天数
- 用户对目标品类第一次订单的para1

用户评论特征

- 用户评论数量
- 用户最近一次评论距离T的天数

商品属性特征

- 用户半个月/一个月/三个月/六个月内购买过目标类目/相关类目/101类目/30类目商品价格的最大值/最小值/均值/总值

行为与订单交叉特征

- 用户对目标品类有行为但是没有订单的数量

特征工程—比较重要的特征

用户模型：

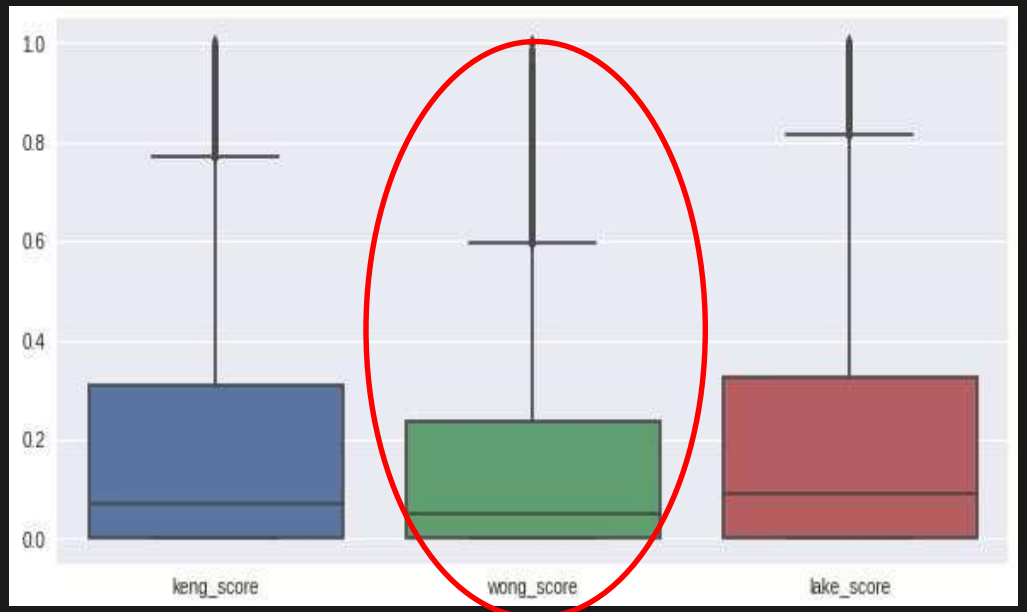
- 用户购买次数或频率
- 用户等级
- 商品参数一
- 商品价格
- 用户购买时间间隔

日期模型：

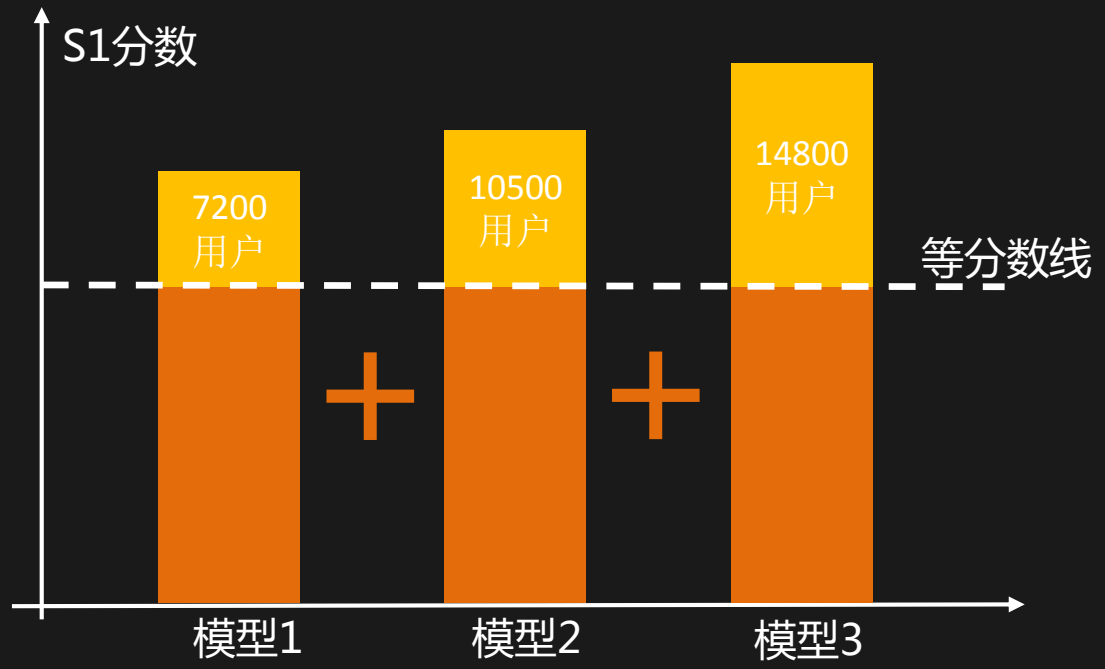
- 日期顺序序列
- 用户活跃天数
- 用户购买时间间隔
- 最近一次的各种行为时间
- 用户的订单价格
- 购买商品的参数一

用户模型融合

- 618用户：选用5、7、8月用户，从而剔除618凑热闹用户
- 各模型比例：根据A榜分数确定重要性比例
- （难点）评分分布不一致：用户评分归一化后，再根据数量比例进行权值换算



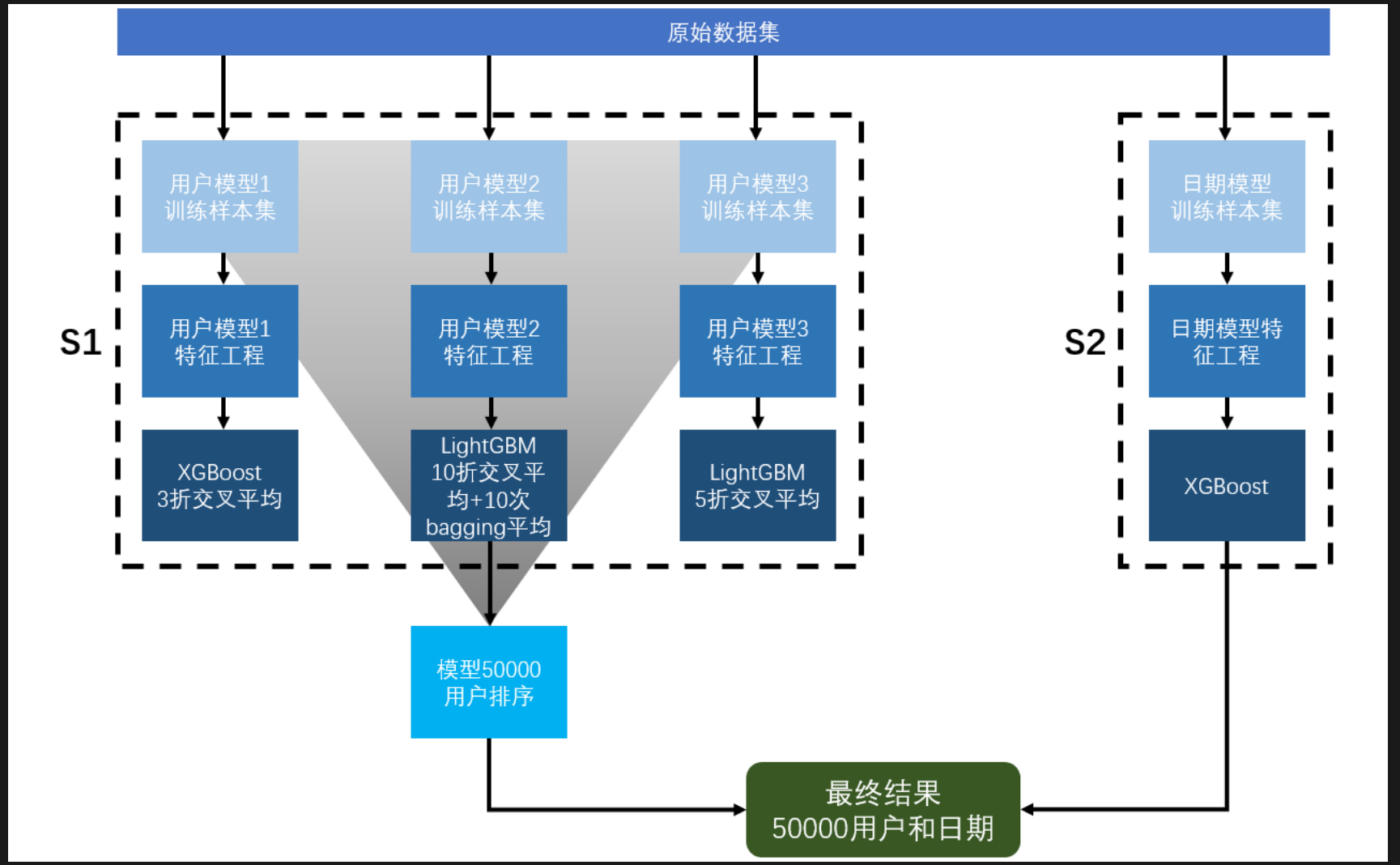
归一化后模型评分分布



模型权值换算方案

JDATA

模型融合


原始数据集
S1
用户模型1
训练样本集
用户模型2
训练样本集
用户模型3
训练样本集
用户模型1
特征工程
用户模型2
特征工程
用户模型3
特征工程
XGBoost
3折交叉平均
LightGBM
10折交叉平均+10次bagging平均
LightGBM
5折交叉平均
S2
日期模型
训练样本集
日期模型特征工程
XGBoost
模型50000用户排序
最终结果
50000用户和日期

JDATA

目录

JDATA

比赛经验总结

- 特征工程决定模型上限
- 多审题，多分析数据，多思考，可以少走一些弯路。
- 比赛应该尽早寻找队友，多交流，尽早找到更优的方案
- 做好工作记录。可以通过云协作，每个人记录自己做了什么工作，提取了什么特征，线上线下效果如何等等，方便队友了解


感谢您的时间。
THANKS.
JDATA