中国大数据算法大赛-用户购买时间预测

队伍名称：何以解忧

演讲者：周宏

2018.07.19

JDATA

目录

1 团队介绍

2 问题描述

3 数据探索

4 算法核心设计思想

5 比赛经验总结

JDATA

团队介绍

团队介绍

周宏

江西理工大学/大三，数据挖掘爱好者

曾露

招商银行/数据分析师，数据挖掘从业者

罗江伟

亿橙信息技术有限公司/java开发，数据挖掘爱好者

李智

腾讯/应用研究，数据挖掘从业者

李佳忆

腾讯/产品运营，数据挖掘爱好者

- 团队介绍：我们团队成员来自不同的地方，每个成员都有着丰富的比赛经验，在国内外各大比赛平台，天池，df，kaggle等都有过获奖经历

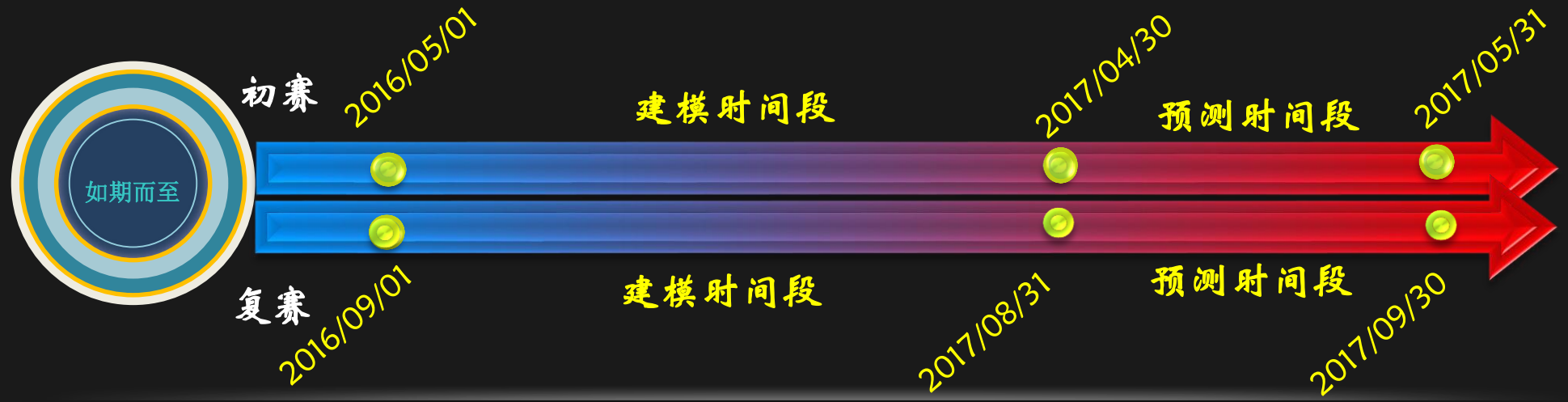
问题描述

数据来源：京东商城

数据内容：用户在一段时间内的行为数据

初赛：用户从2017/02-2017/04发生购买行为的用户中抽取

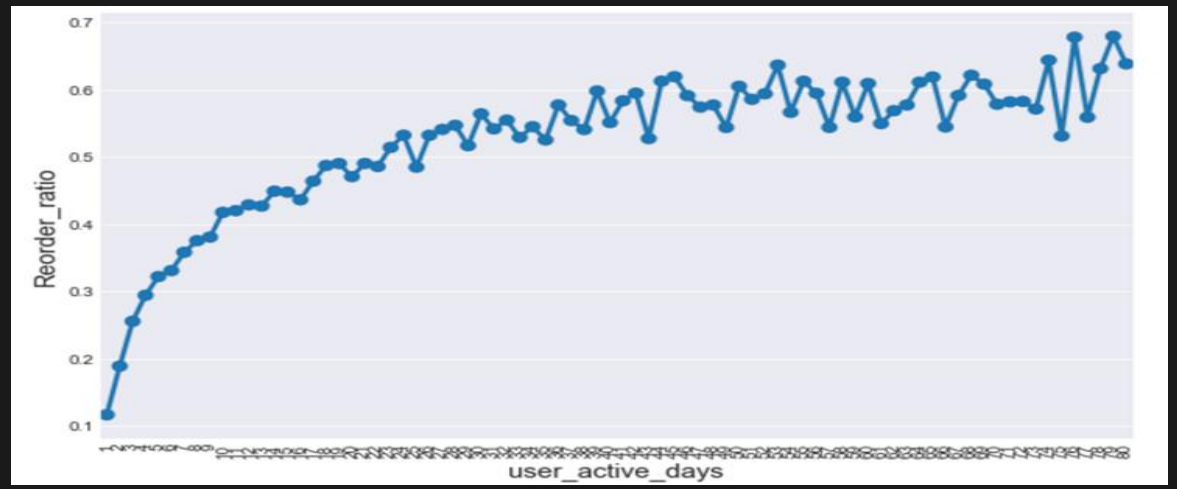
复赛：用户从2017/06-2017/08发生购买行为的用户中抽取

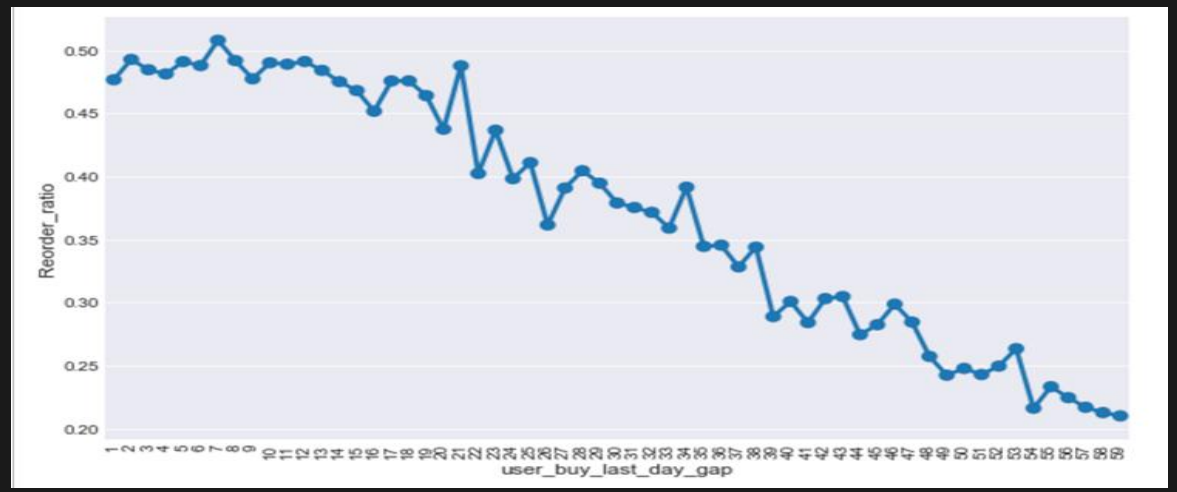


根据用户历史行为数据，预测用户是否会在规定时间内发生购买行为，并且预测购买日期

JDATA

数据探索

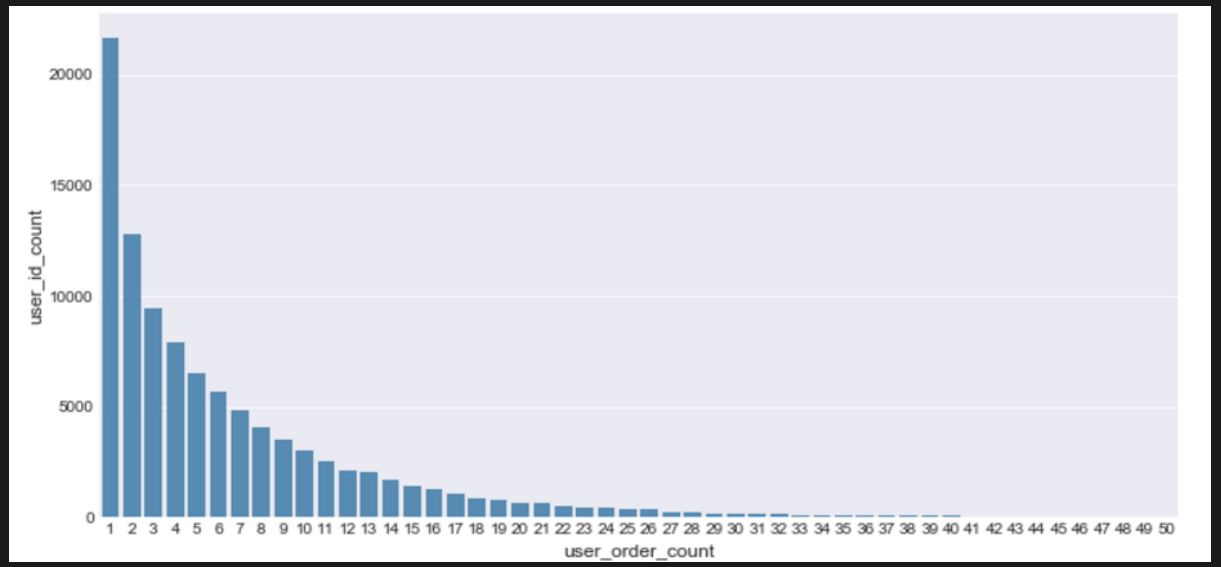


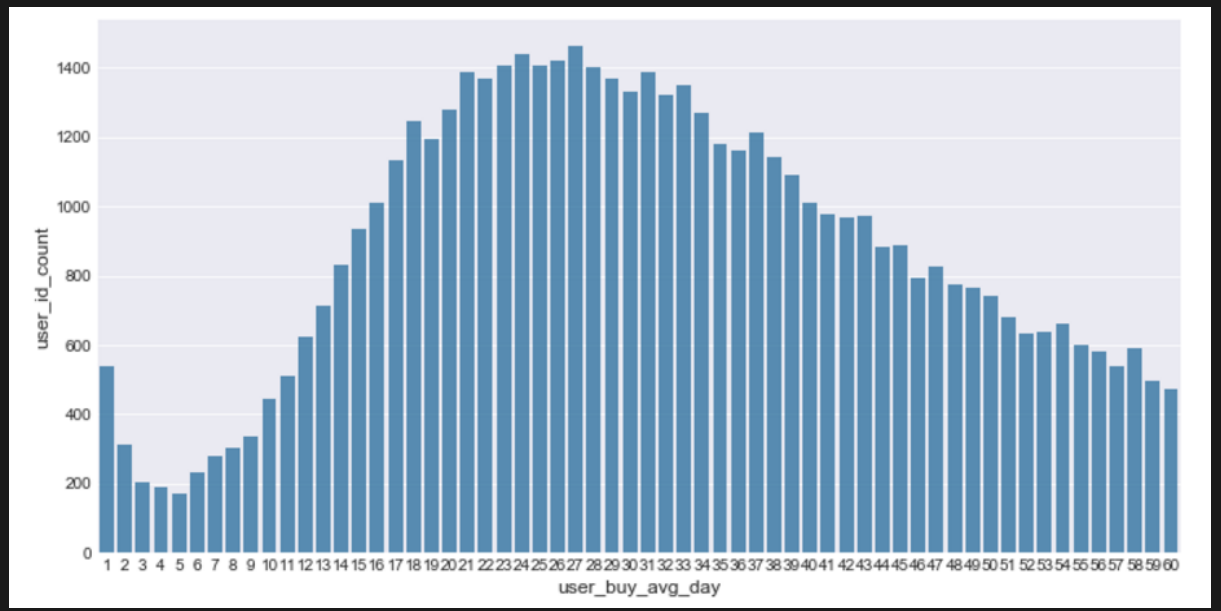


通过数据探索，我们可以看到
用户活跃时间越长，复购概率越高。
用户最后一次购买距离label日的间隔越短，购买概率越高。因此，时间相关的特征非常重要，并且，具有强烈的时效性。

JDATA

数据探索

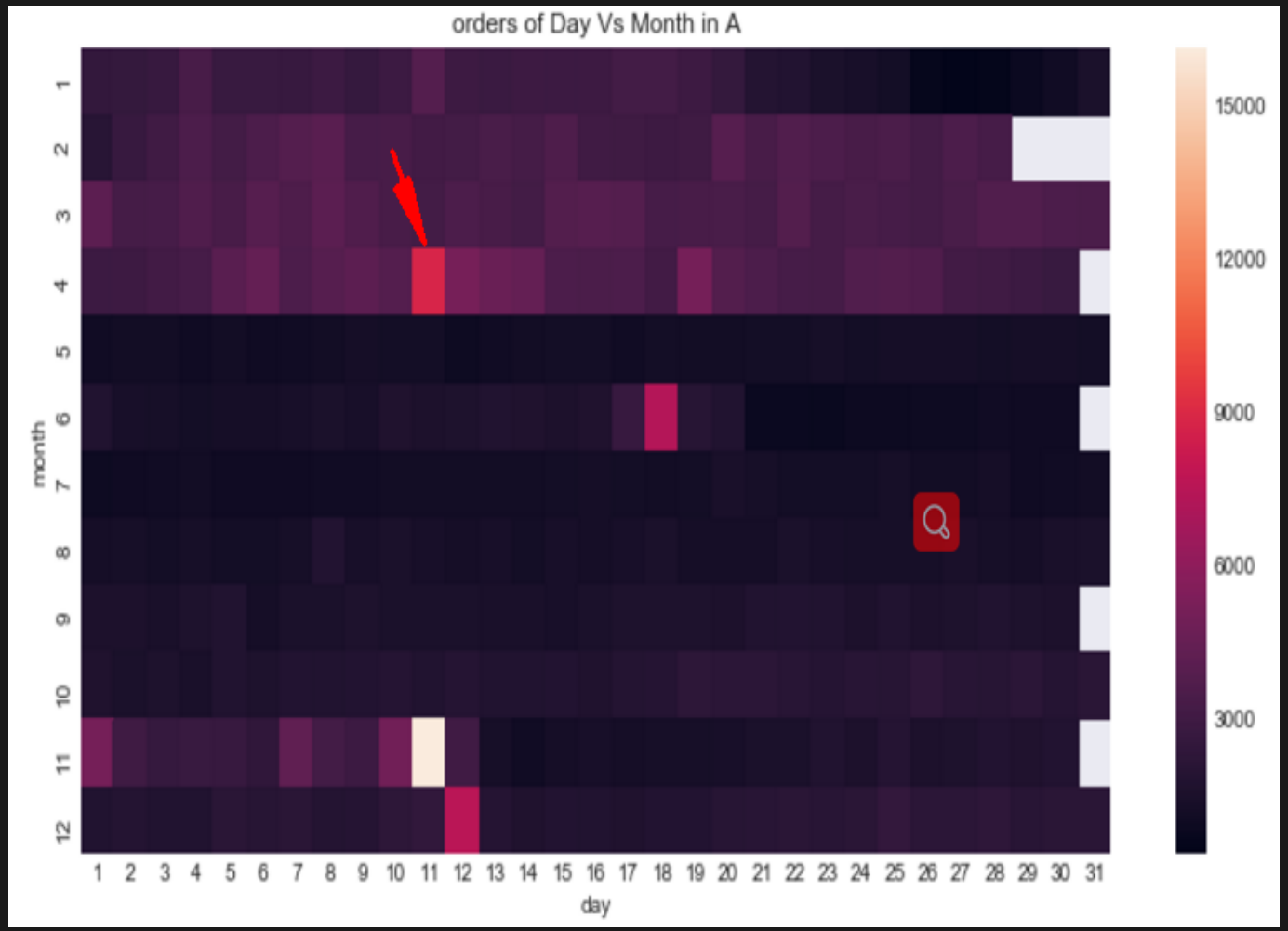




通过数据探索，我们可以看到绝大多数用户只购买了几次(1-5),呈现明显的长尾分布,并且绝大多数用户的平均购买时间间隔在20-35天左右，所以滑窗统计用户特征的方式是行不通的(最近三天，最近七天，最近15天.....)

JDATA

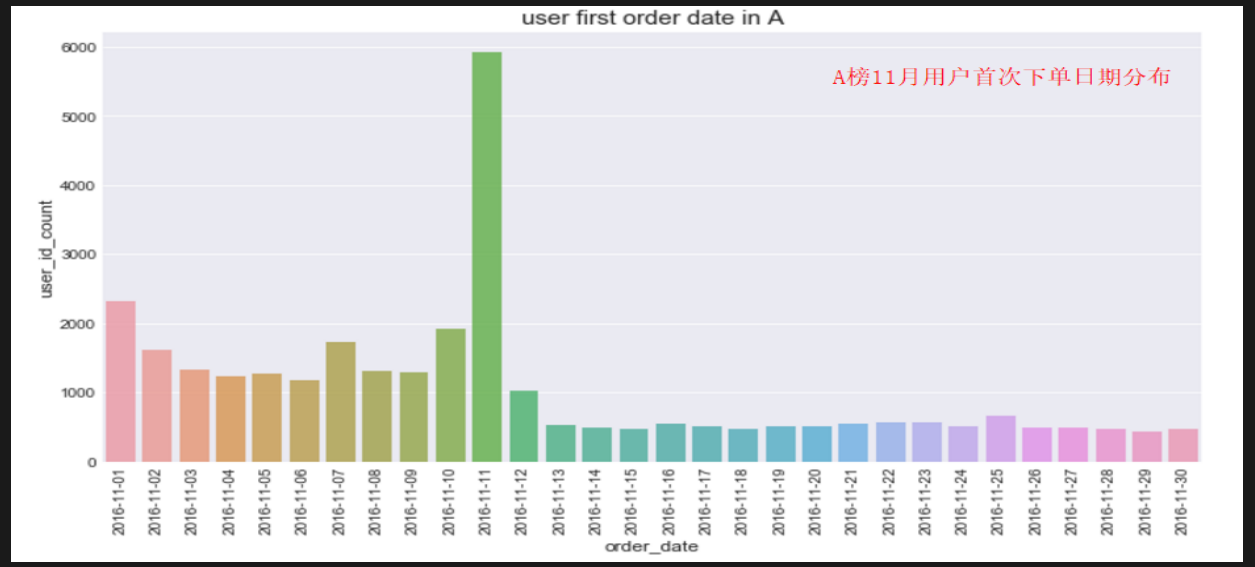
数据探索-异常数据

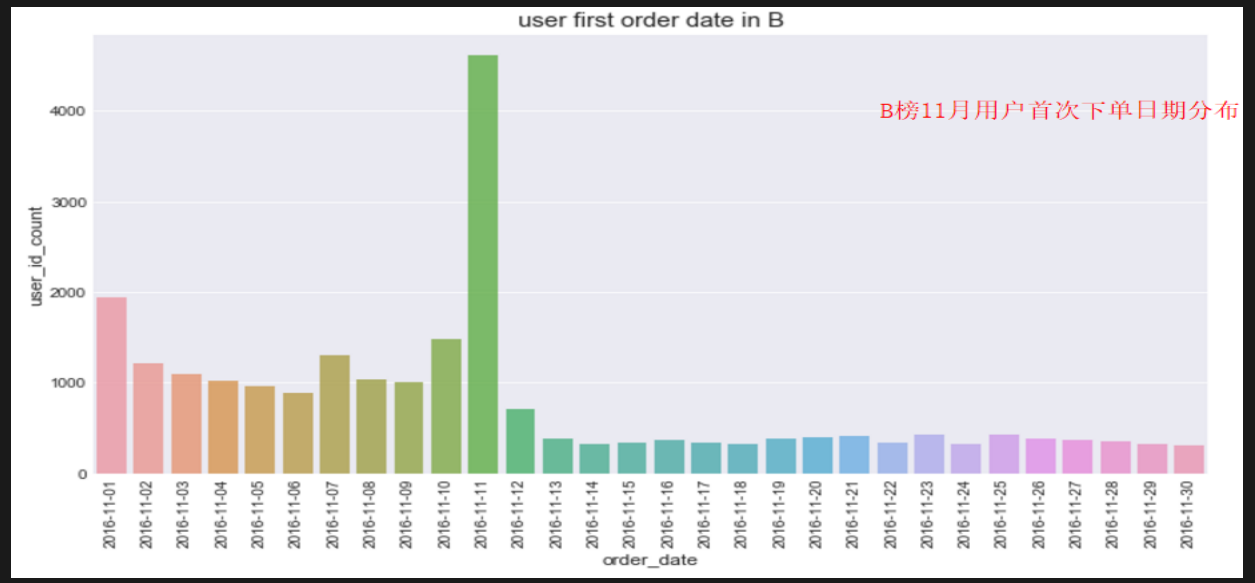


初赛的时候，我们发现了在A榜数据中，411这一天的销售量仅次于双11，因此我们把这一天当作是异常数据进行处理，但是效果却非常的差，因此，A榜数据没有受到随机用户影响。

JDATA

数据探索-异常分析

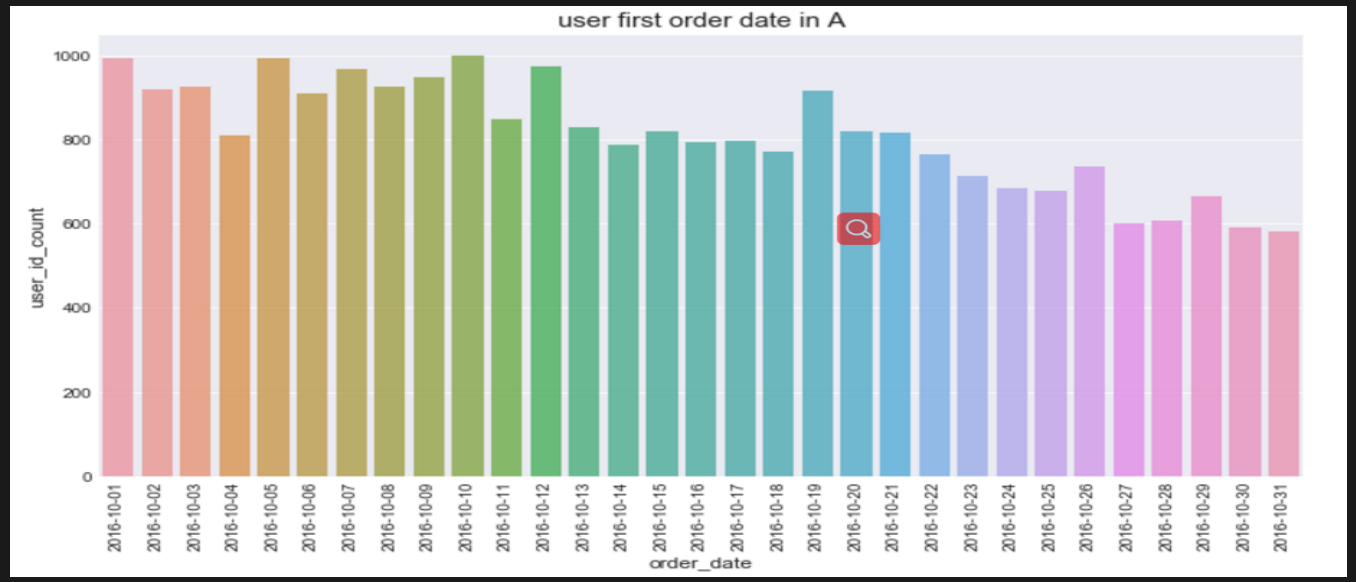


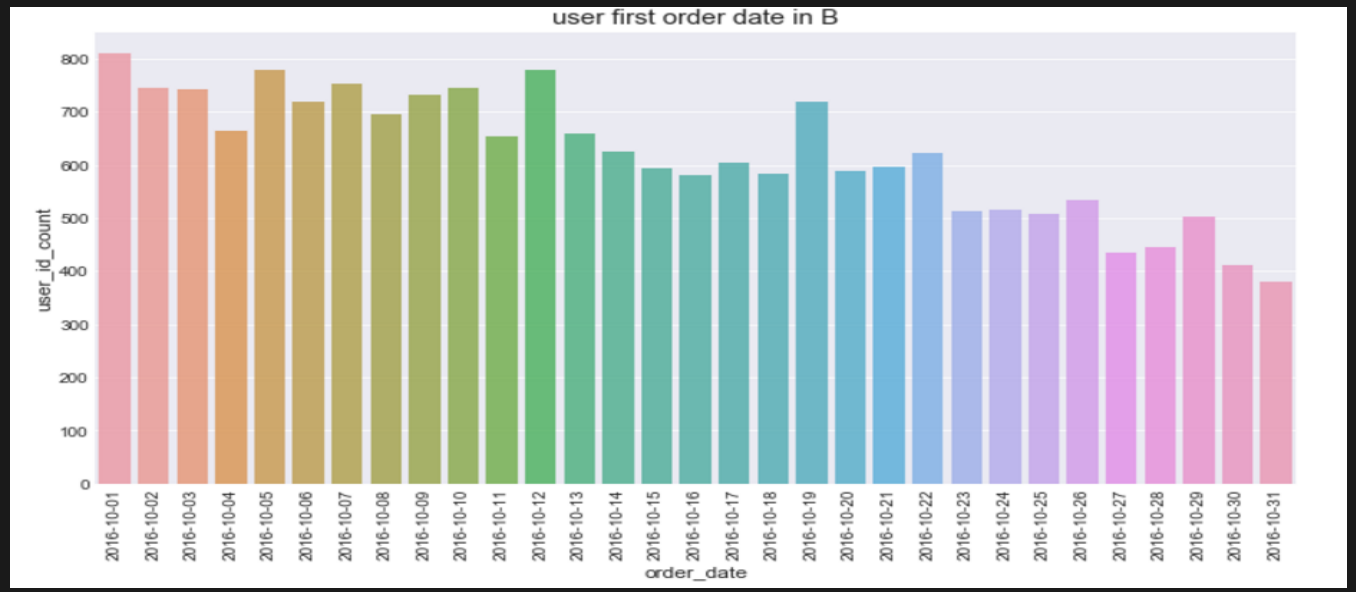


所谓随机用户指的是冲着618来的用户，只在618大促期间购买用户，因为训练数据采样的方式，A榜和B榜的11月数据都不受随机用户影响，因此可以看到，两者用户首次下单日期的分布极其一致，实际上，每个月的分布都很类似。

JDATA

数据探索-异常分析

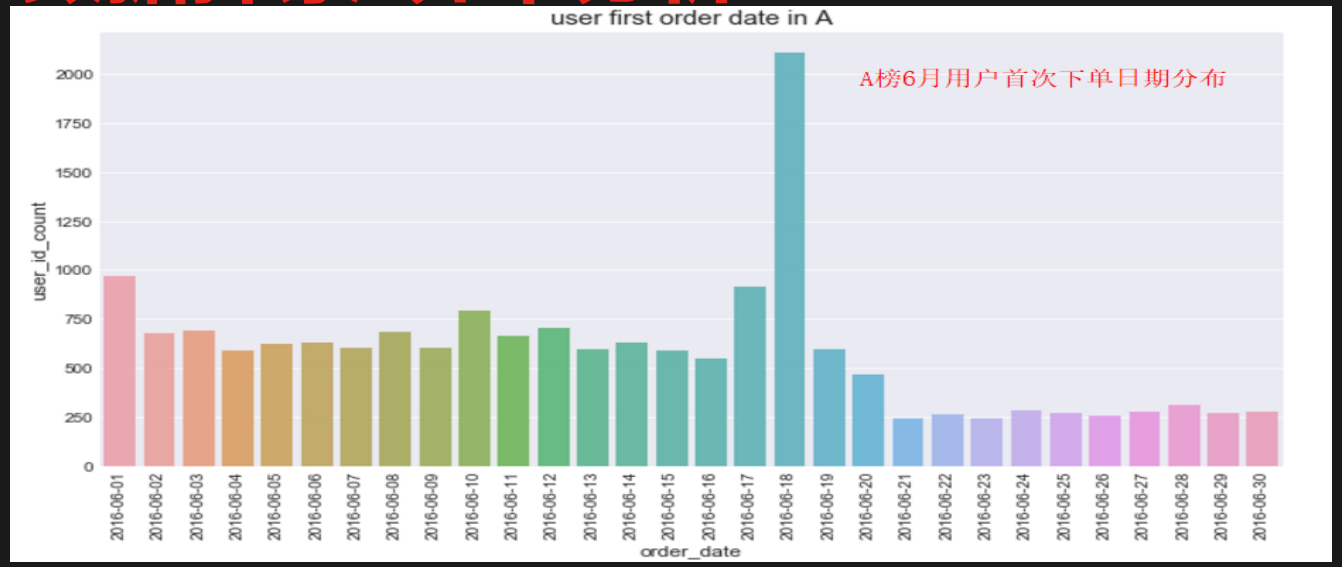


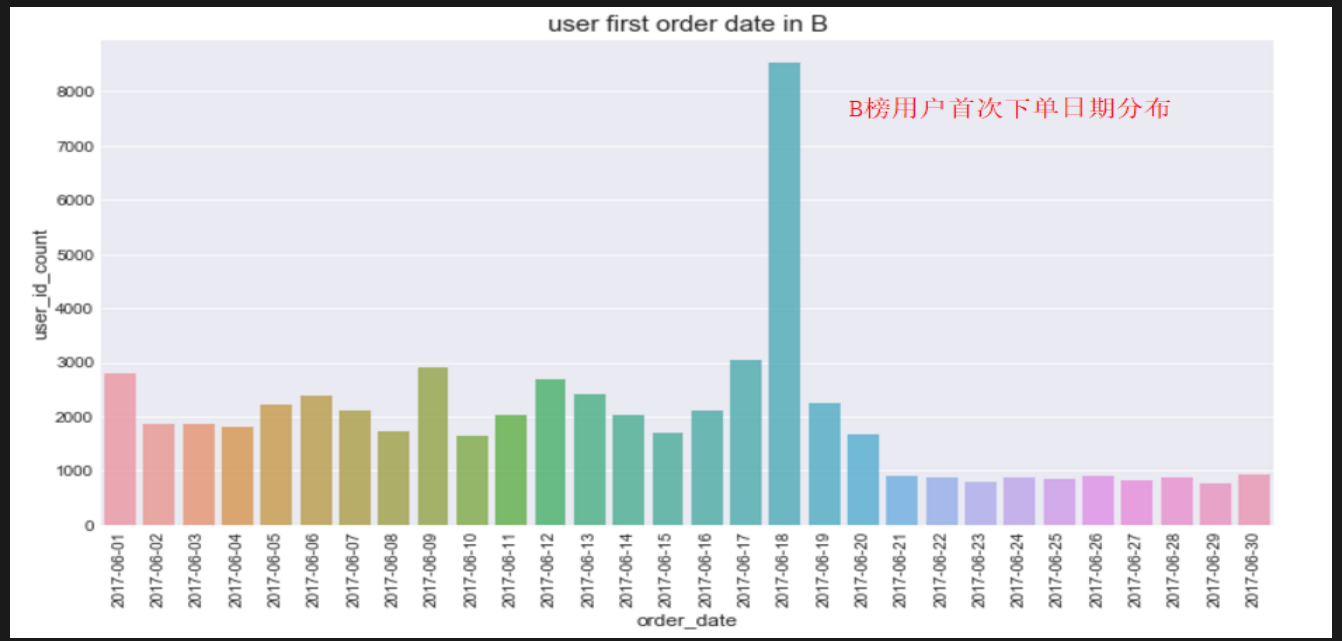


AB榜数据10月同样不受随机用户影响，两者分布基本一致

JDATA

数据探索-异常分析

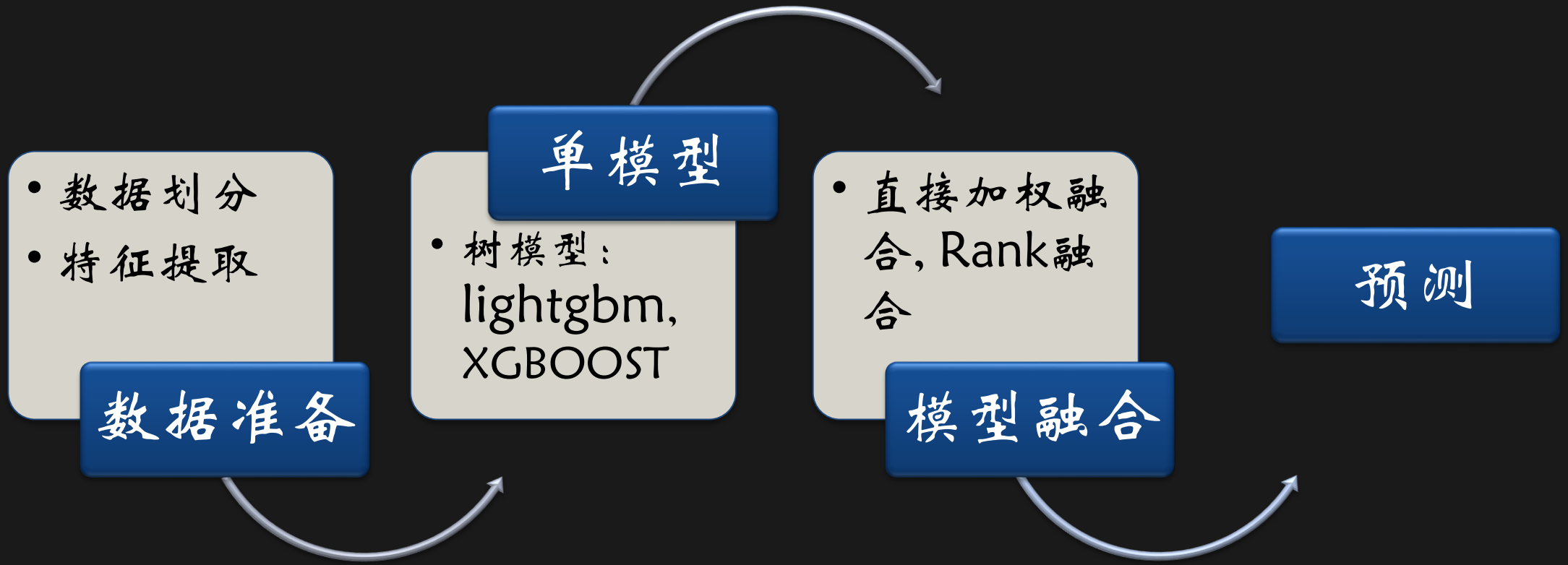




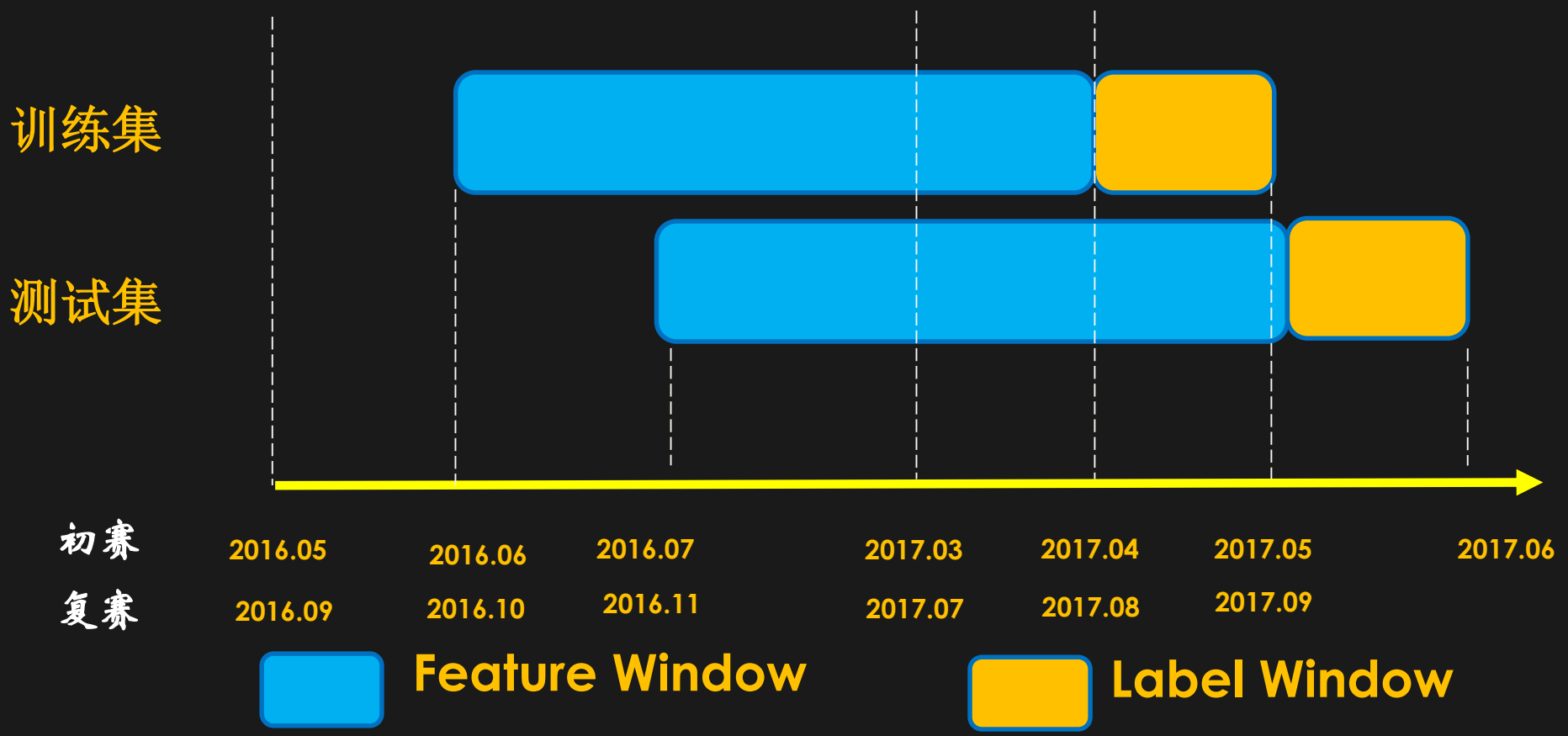
A榜6月不受随机用户影响，但是B榜6月受到随机用户影响，两者分布有一些小区别，可以看到B榜618这一天比A榜要更加突出，因此，根据上述推理，我们认为，6月的用户删多了效果应该不会好，毕竟有很多正常的用户会赶上618大促，因此我们建模的时候只是保守的删除了仅仅只在618当天购买的用户

JDATA

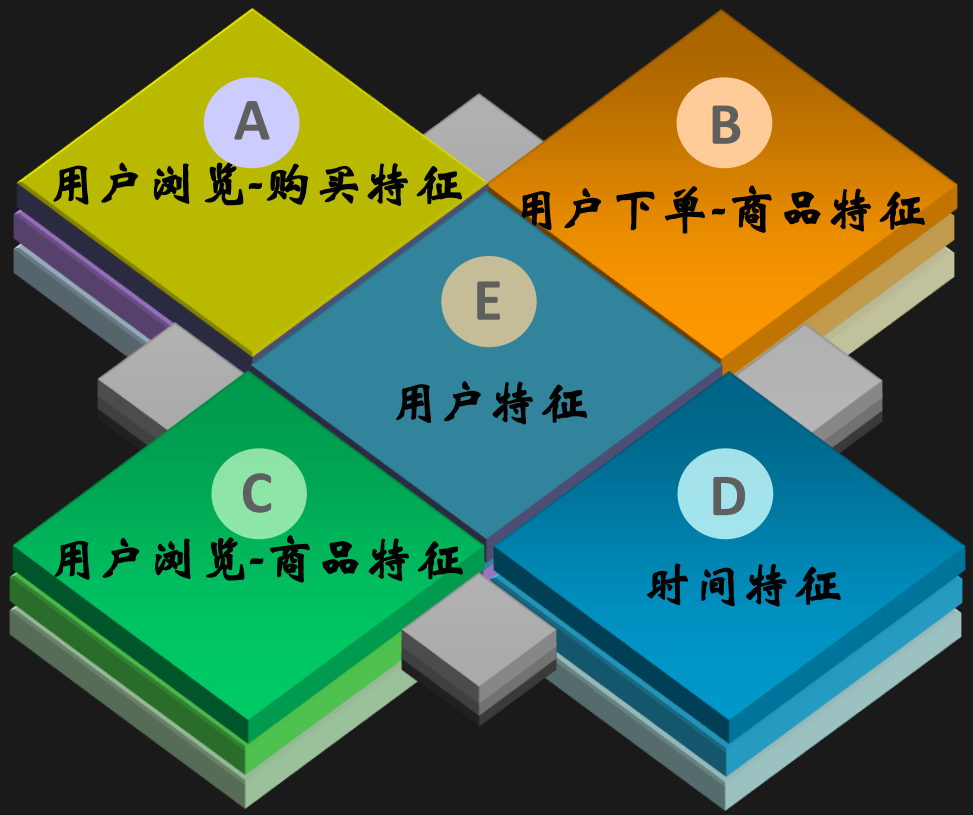
算法思想-算法框架


• 数据划分
• 特征提取
数据准备
单模型
• 树模型：lightgbm, XGBOOST
• 直接加权融合, Rank融合
模型融合
预测

算法思想-数据划分S1


训练集
测试集
初赛
复赛
2016.05
2016.06
2016.07
2017.03
2017.04
2017.05
2017.06
2016.09
2016.10
2016.11
2017.07
2017.08
2017.09
Feature Window
Label Window

算法思想-特征工程



算法思想-特征工程

- 时间类型特征非常重要
- 大多数用户只购买了少数几次，只能通过用户购买以及浏览的sku的固定属性(price, para_1,para_2,para_3,sku复购次数以及复购率等)去推断用户的个人喜好以及购买习惯。
- 用户离考察时间段越近的行为越重要，因为绝大多数用户只有少数几次购买行为，所以我们只需要关注用户最近的几次行为，如最后一次购买/浏览，倒数第二次购买/浏览

算法思想-特征工程

用户下单-商品特征

- 用户最后一次购买目标品类的price总和，均值，最大值，方差
- 用户最后一次购买目标品类的para_1,para_2,para_3总和，均值，最大值，方差
- 用户最后一次购买的sku，平均被用户复购了多少次以及复购率
- 用户倒数第二次购买的上述情况
- 用户最后一次购买的price,para_1,para_2,para_3 减用户倒数第二次对应商品属性
- 用户对目标品类商品属性的偏好，
- 用户对目标品类购买次数
- 用户在多少个区域下过单，每个区域下单次数多少
- … …

用户浏览-商品特征

- 用户浏览对商品属性的偏好
- 用户重复浏览商品的情况
- 用户最后一次浏览的商品属性
- 用户最后一次浏览的商品价格
- 用户对目标品类的总浏览次数
- … …

用户特征

- 用户的年龄，级别，性别

算法思想-特征工程

时间特征

- 用户最后一次购买/浏览目标品类距离label的时间间隔
- 用户倒数第二次购买/浏览目标品类距离label的时间间隔
- 用户最近第一次购买/浏览目标品类的时间间隔
- 用户平均购买目标品类的时间间隔，最大时间间隔，方差
- 用户最近一次连续多少个月对目标品类下单/浏览
- 用户的活跃天数
- … …

用户浏览-购买特征

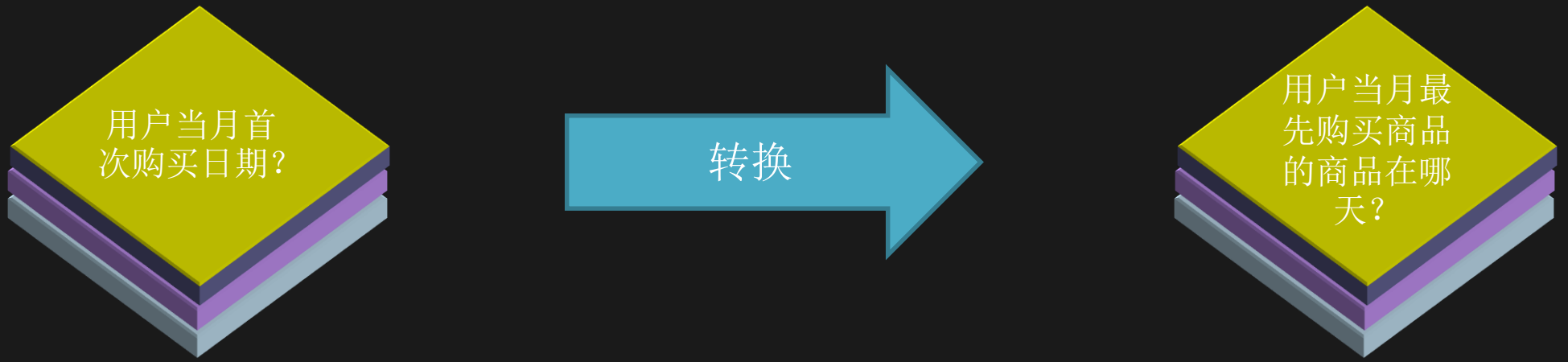
- 用户最后一次购买目标品类之后，之后是否浏览过目标品类，浏览过多少次
- 用户最后一次浏览目标品类之后，之后是否购买过目标品类，购买过多少次

算法思想-规则

- 我们分析数据发现，用户在最后几天发生浏览行为但是却没有购买的用户，有非常大的概率在未来一个月发生购买行为，因此我们设计了一条规则，用户在8月25日之后没有购买目标品类，但是浏览过目标品类的用户，按照最后的浏览日期排序
- 通过分析，还有一种用户，在很多个月都有购买行为，那么这些用户在未来一个月也有很大概率购买，因此我们设计一条规则，用户订单数>30并且用户活跃的月数在6个月以上，按照用户的习惯购买日排序(习惯在月底购买的用户优先于习惯在月初购买的用户)
- 把满足第一个规则的用户放在满足第二个规则的用户前面，去重，取代模型预测出的尾部用户

以上两个规则在S1可以提高2个千分位

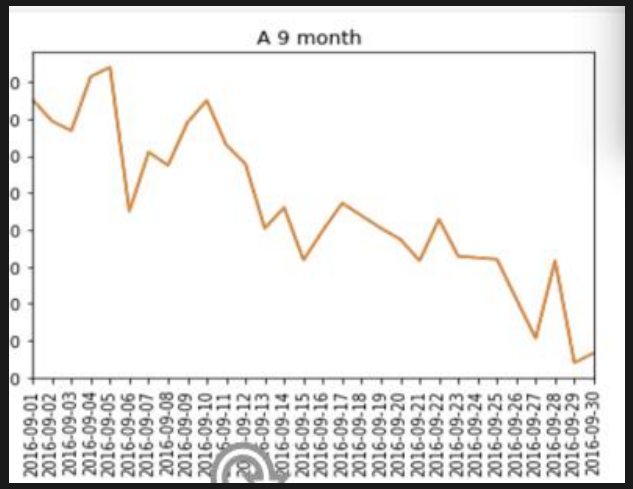
算法思想-S2思路

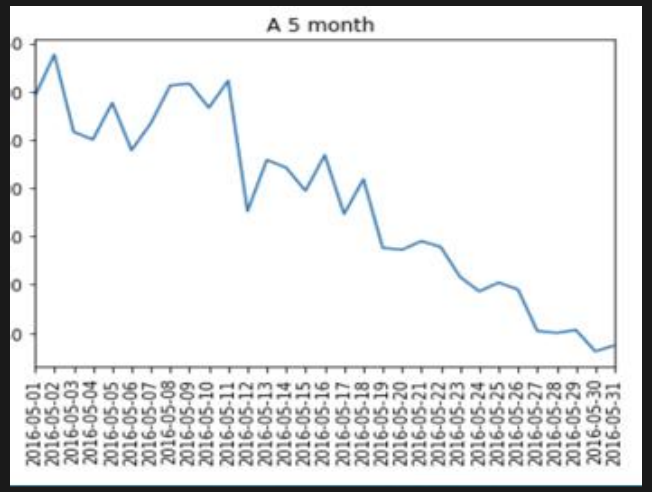


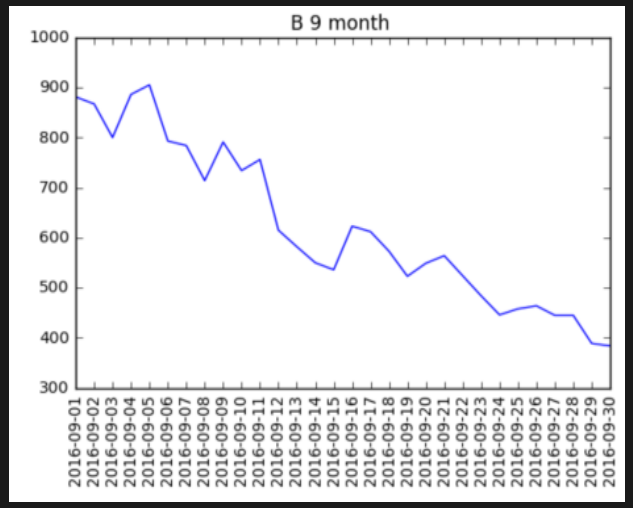
- 测试样本构造细粒度到每个之前下单过的sku,以<user_id,sku>作为样本唯一标识，该方法在线上有3-4个千分点提升
- 如 原本样本 : user_id　　转换为： user_id　sku_id ,这样便能拼接sku特征，并且将测试集扩充了十倍

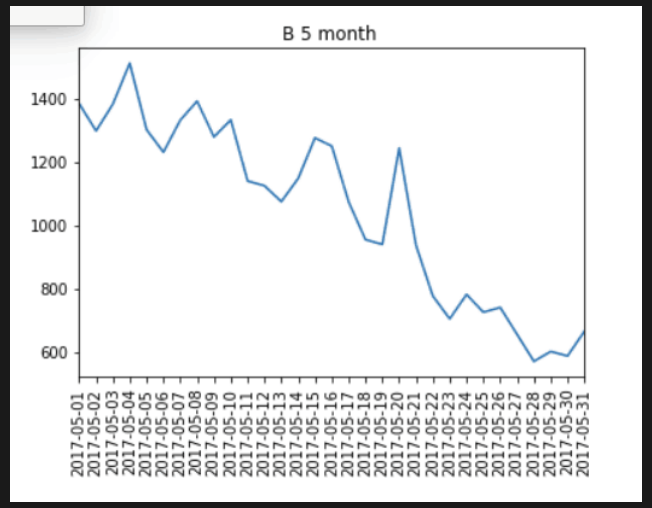
1	1	1
2	1	2
3	1	3
4	2	1

算法思想-S2趋势分析





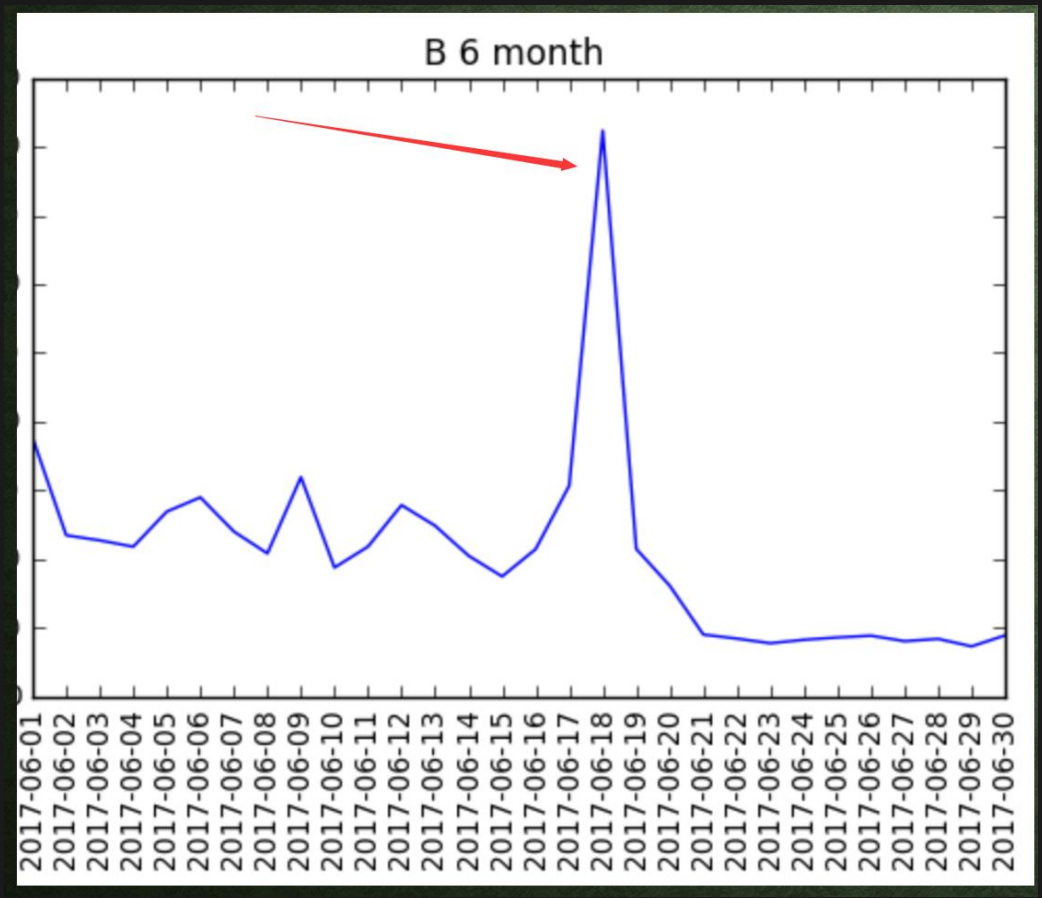




我们分别画出了a榜数据5月份和b榜数据5月份对比，a榜数据9月份和b榜数据9月份进行对比，观察发现，基本上所有月份的每天下单数数目整体趋势都是相同的，所以我们有理由认为在非促销月份的的每天下单数同样服从这个趋势

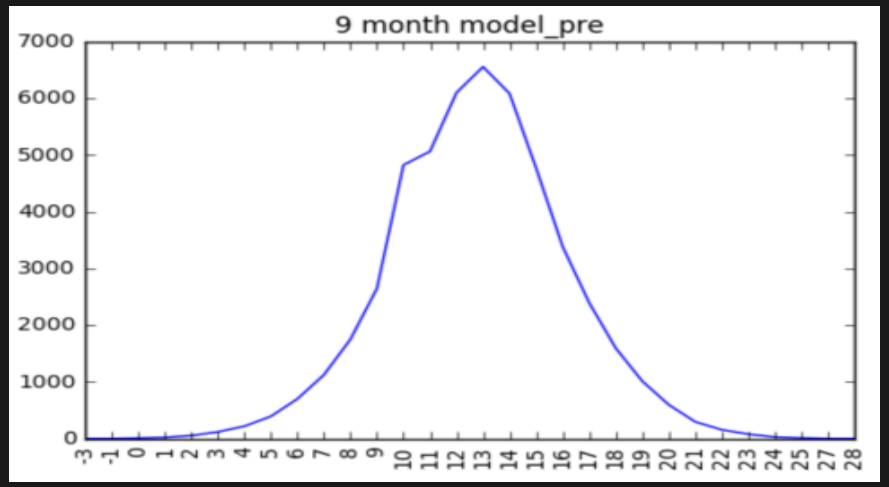
JDATA

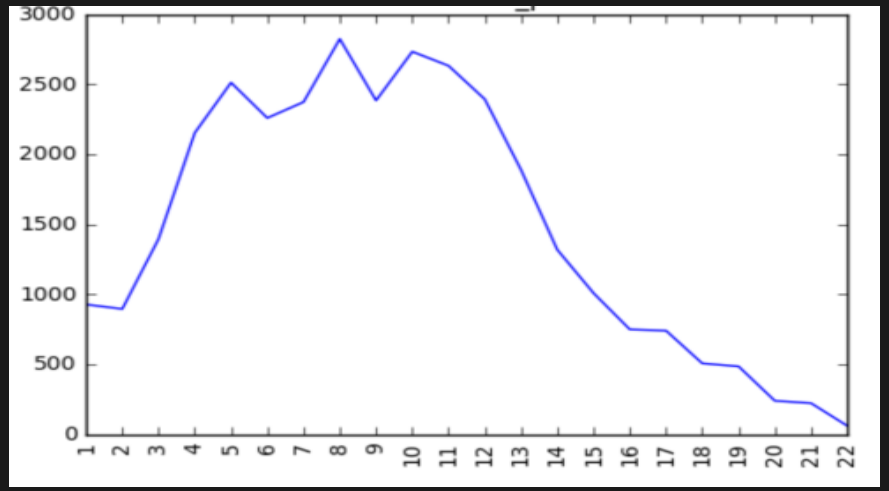
算法思想-B榜618趋势分析



我们分析6月份的下单分布发现，受到京东618的影响，六月份下单天数分布和其月份十分不同

算法思想-S2模型预测趋势分析

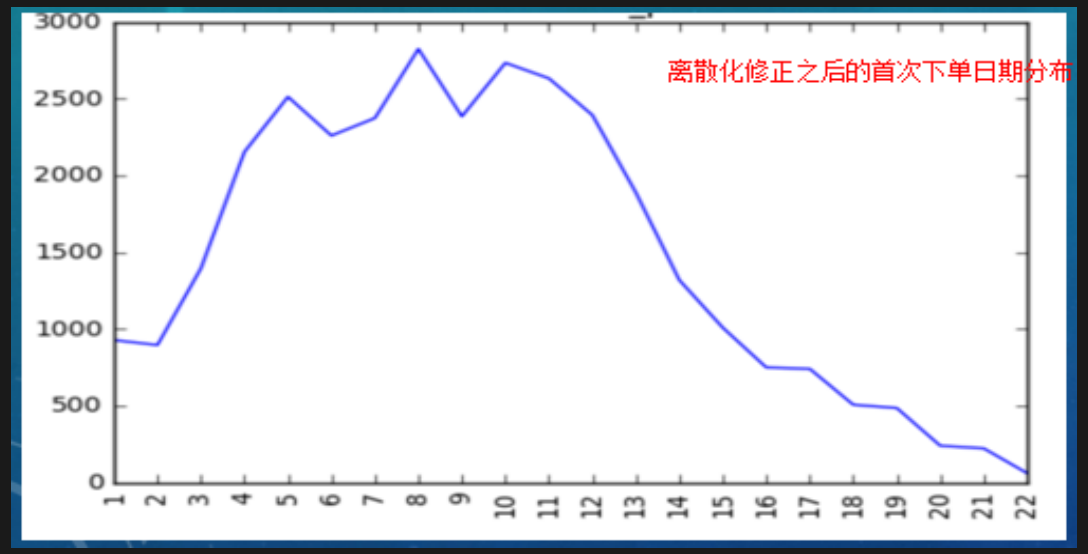


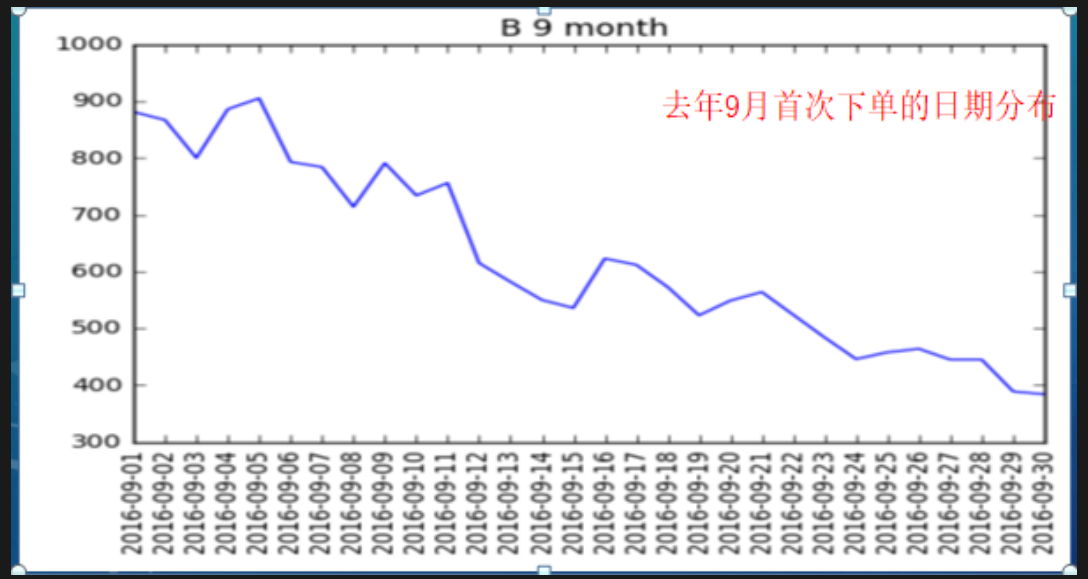


上图是模型预测出的9月用户首次购买日期的分布，由图可看出，2016年9月实际的分布和模型预测的分布差距较大，因此，我们对模型预测结果，做了人工离散化的修正，下图为修正后的分布，可认为模型预测的结果都集中在了10-13这个区间内，因此我们根据往月分布，人为的对10-13预测范围根据小数点，人为进行一个划分，在该方法在线上获得了百分位的提升

JDATA

算法思想-S2离散化操作






```python
def ggl(x):
    if 12.8<=x<13.25:
        return 10
    elif 12.4<=x<12.8:
        return 9
    elif 11.9<=x<12.4:
        return 8
    elif 11.4<=x<11.9:
        return 7
    elif 10.8<=x<11.4:
        return 6
    elif 10<=x<10.8:
        return 5
    elif 9<=x<10:
        return 4
    elif 8<=x<9:
        return 3
    elif 7<=x<8:
        return 2
    elif x<7:
        return 1
    elif 13.25<=x<13.7:
        return 11
    elif 13.7<=x<14.15:
        return 12
    elif 14.15<=x<14.6:
        return 13
    elif 14.6<=x<15:
        return 14
    elif 15<=x<15.4:
        return 15
    elif 15.4<=x<15.8:
        return 16
    elif 15.8<=x<16.3:
        return 17
    elif 16.3<x<16.8:
        return 18
    elif 16.8<=x<17.5:
        return 19
    elif 17.5<=x<18:
        return 20
    elif 18<=x<19:
        return 21
    elif 19<=x<20:
        return 22
    else:
        return x
```

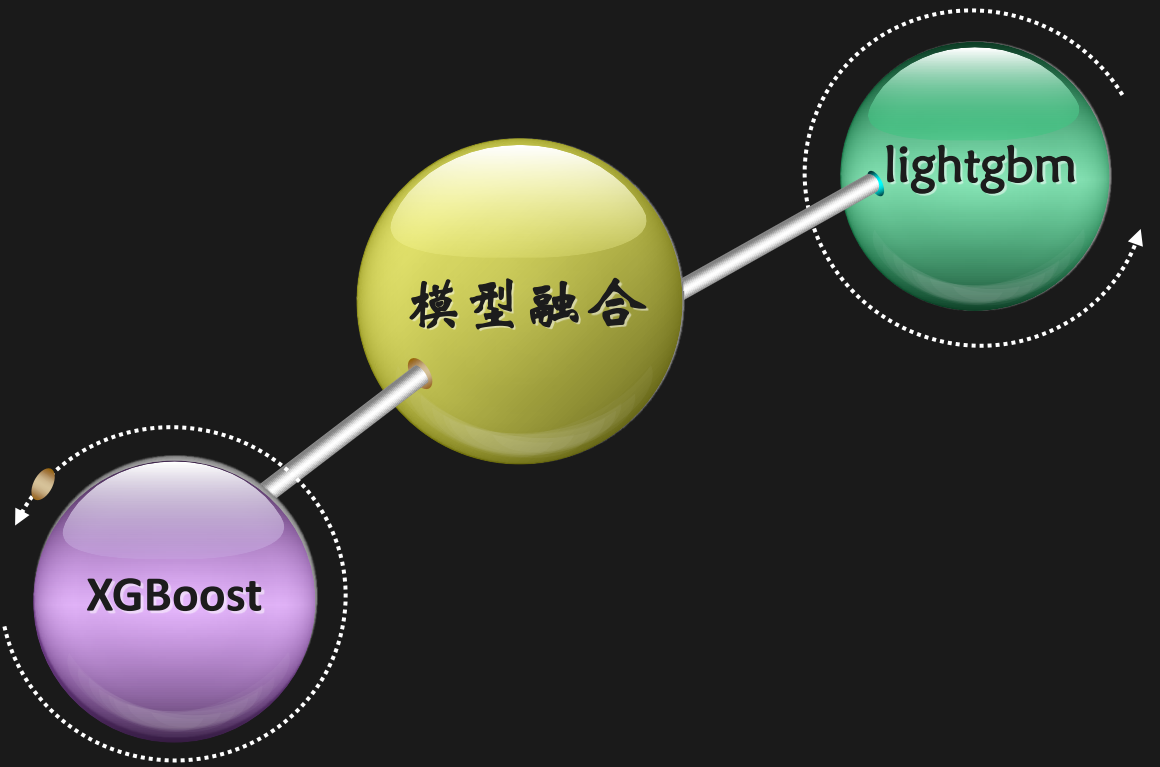
具体的离散化规则

JDATA

▸算法思想-S2模型融合

- 共训练了XGBoost, lightgbm两种模型，

s2的融合我们提取了两份特征，包括通过滑窗法构造了多份训练集分别训练模型进行加权融合。



比赛总结

在这比赛初期，我们主要做的是分析数据使用简单的统计特征并且不断的优化我们模型整体框架，比赛中期主要是强特性的挖掘，比赛后期主要是规则的探索以及模型的融合，在这里非常感谢 JDATA 给我们提供这么一个宝贵的机会去展示我们的成果

THANKS

请各位专家批评指导！

JDATA