中国大数据算法大赛-用户购买时间预测

队伍名称：最后一波咯

演讲者：汪智开

2018.07.19

JDATA

目录

1 团队介绍

2 算法核心设计思想

3 比赛经验总结

JDATA

团队介绍　最后一波咯

汪智开

华南理工大学研三学生
数据挖掘爱好者

张帆

华南理工大学研三学生
数据挖掘爱好者

梁冠强

华南理工大学研三学生
数据挖掘爱好者

邱泽增

华南理工大学研二学生
数据挖掘爱好者

曾正

华南理工大学研二学生
数据挖掘爱好者

我们团队第二次参加数据挖掘竞赛

目录

1 团队介绍

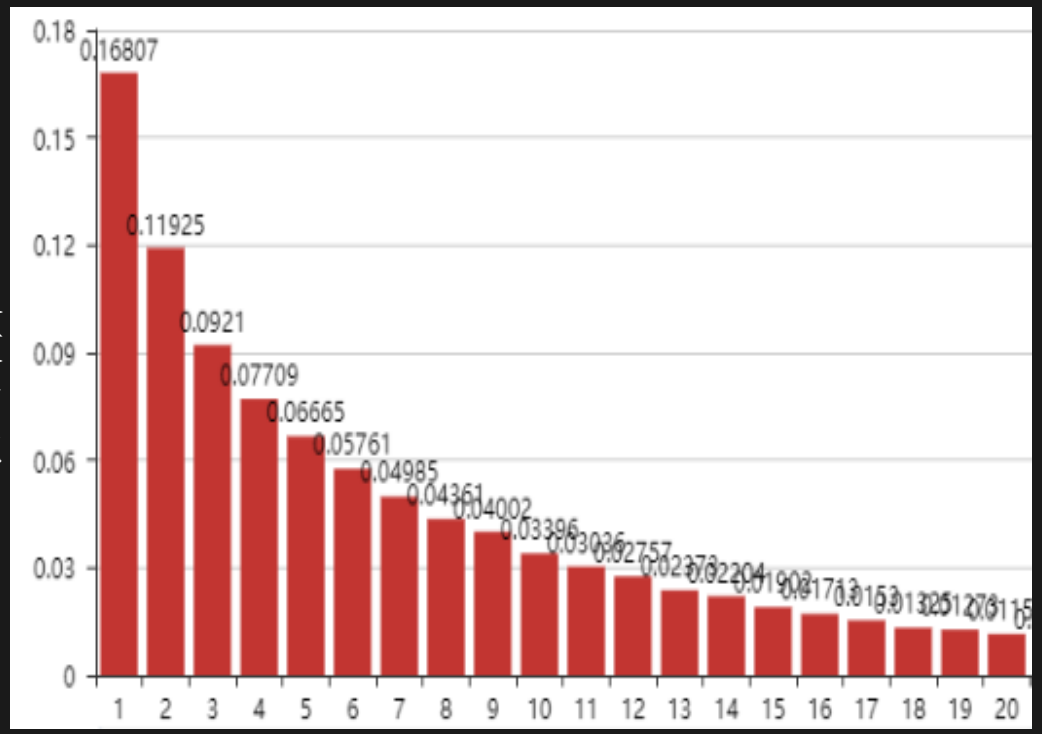
2 算法核心设计思想

3 比赛经验总结

JDATA

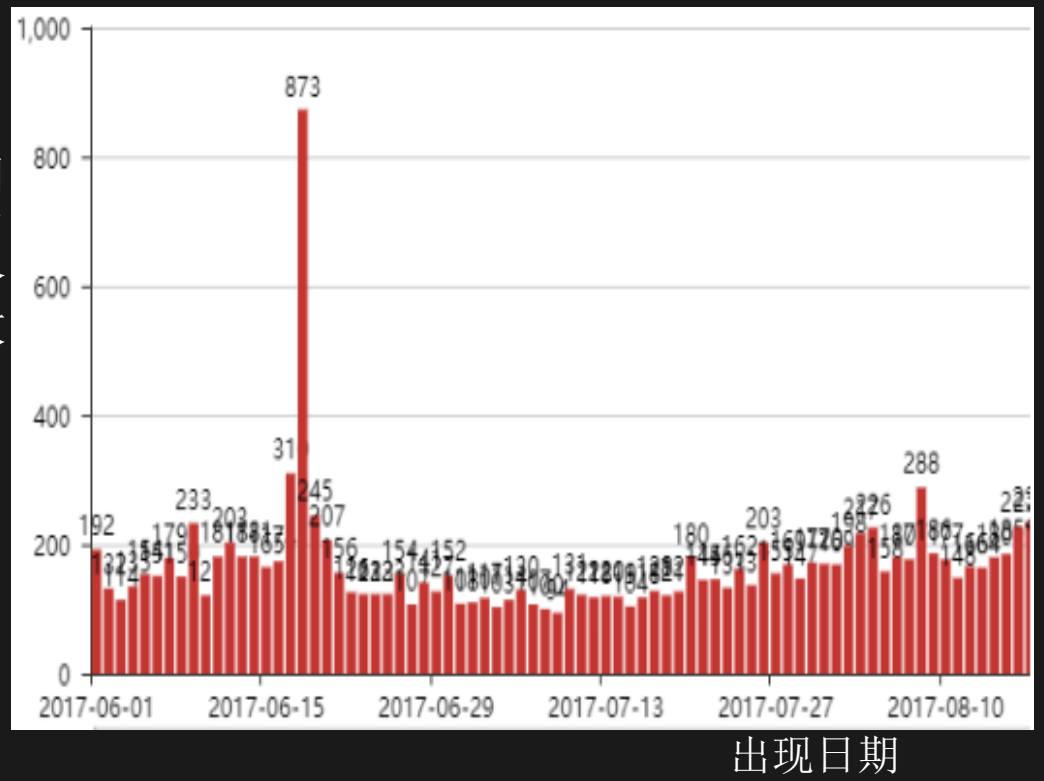
数据分析

不同订单数的用户占比



低频诉求：大多数用户购买订单数很少

只出现一次的用户

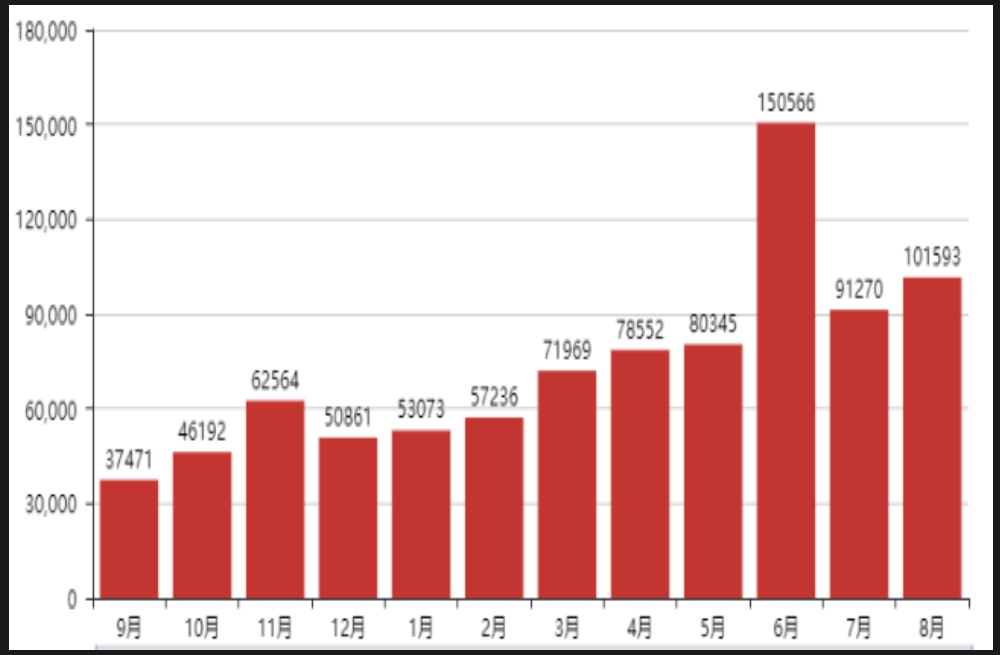


只出现一次的用户集中在预测月前3月，这些用户难以分析

JDATA

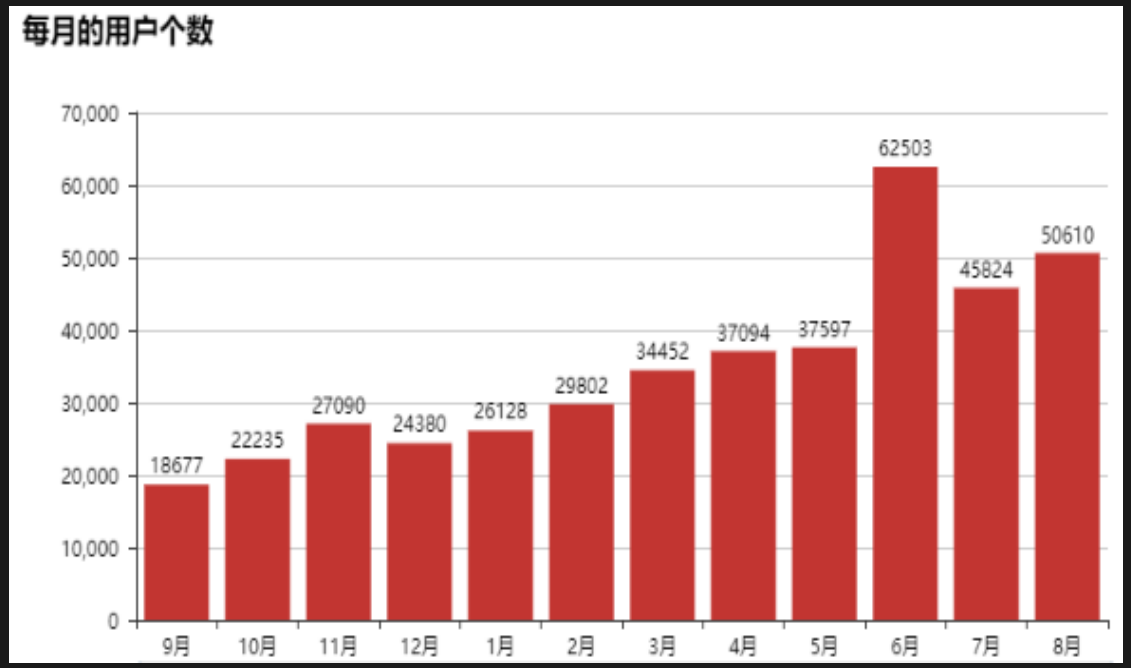
数据分析

每月订单数统计



越靠近预测月的订单数越多

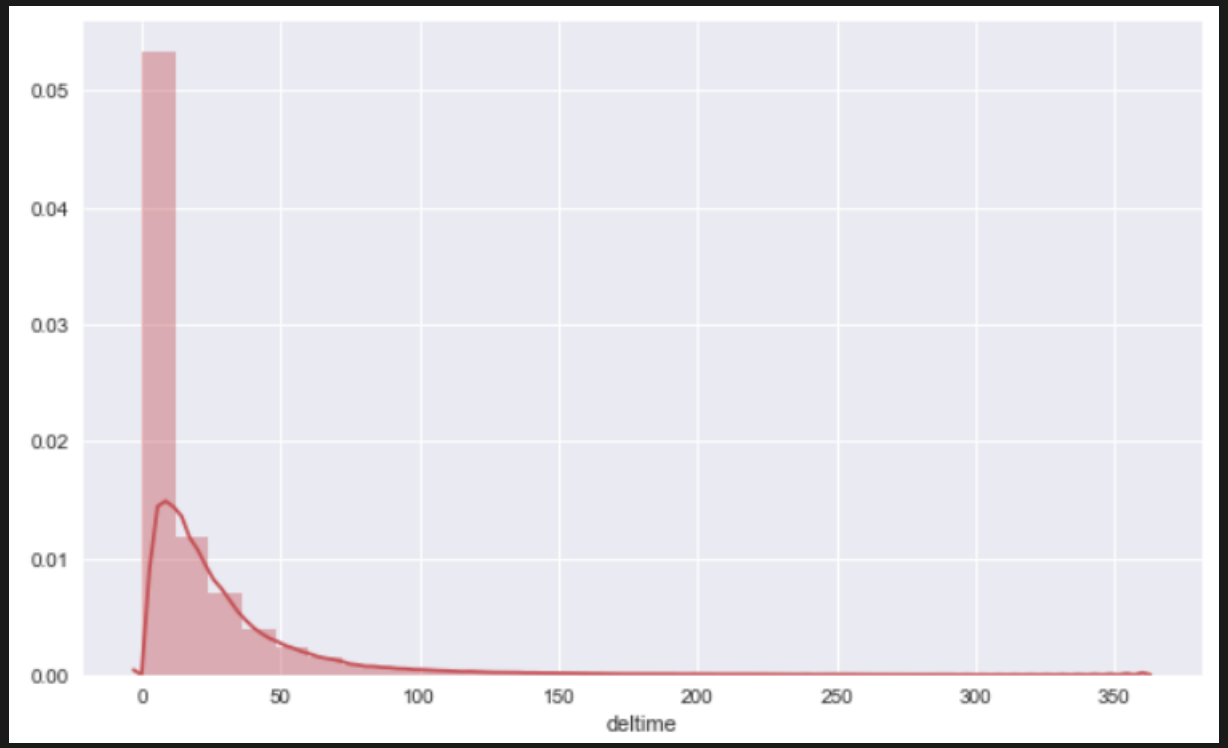
每月用户数统计



越靠近预测月的用户数越多

JDATA

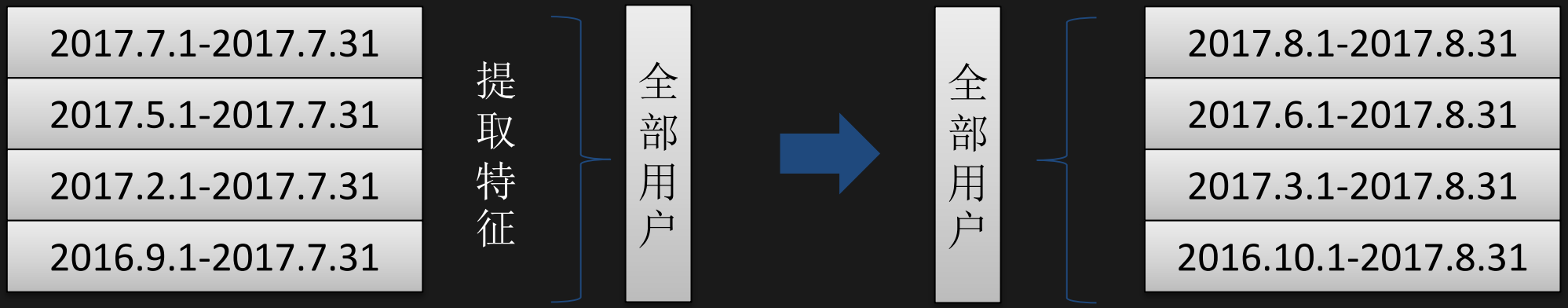
数据分析



用户购买间隔大于60天的人数极少，越靠近预测月的订单数越多，所以我们重点认为越靠近8月底购买过相关品类的人群，在9月份购买的概率越大（相对）

线上线下测试

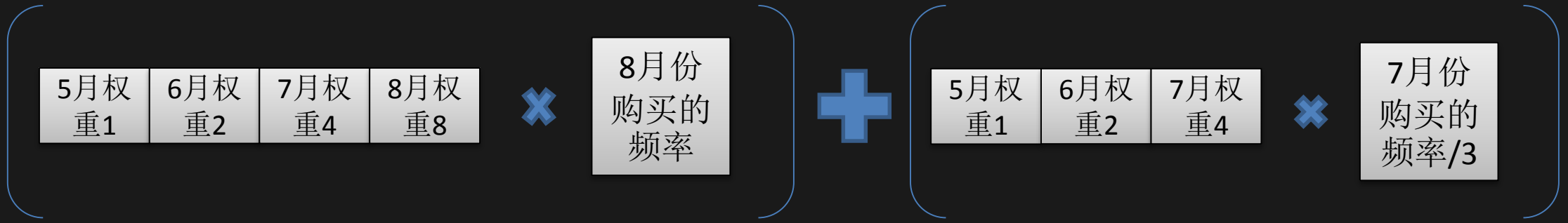
我们选择用全部用户偏移前一个月的数据作为训练集，下一个月的数据作为测试集



JDATA

模型设计

S1用户购买概率从大到小排序，采用回归模型；数据分析可看出大致的时间购买间隔，



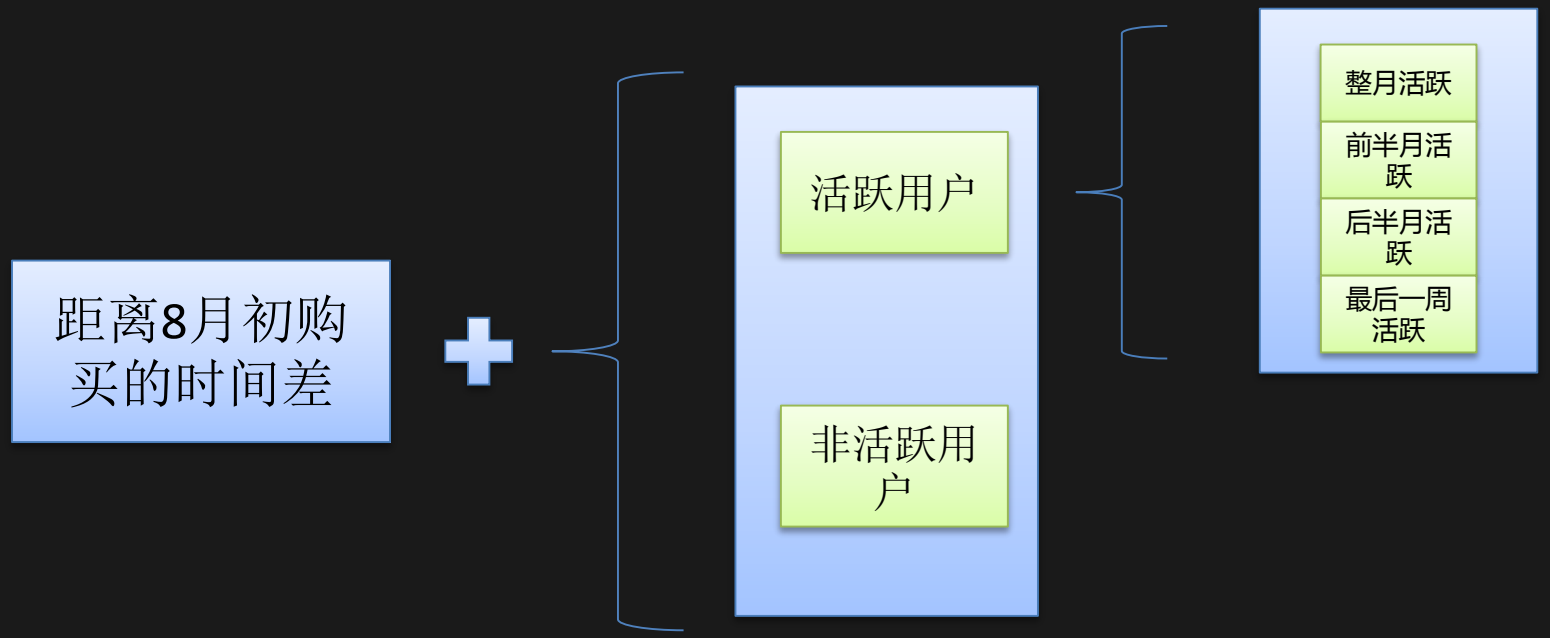
修正部分：（1）最后一星期里面购买频率大于等于3天的人，在下个月购买概率会比较大
（2）连续4个月购买的人群，在下一个月购买概率会比较大

在A榜修正这两部分S1提高了1个百分点

JDATA

模型设计

S2用户下单日期评价，采用回归模型



JDATA

特征工程

特征群（趋势特征）

趋势特征群

(1)统计用户连续7天内购买的频率，共有几天购买过101或者30品类

(2)统计用户是否有连续几个月购买

(3)统计用户连续7天内下单的件数

.....

特征工程 特征群（统计特征）

一个时间段内

订单特征群

- 用户下单次数
- 用户订单中商品id数目
- 用户订单中下单天数
- 用户订单消费价格水平
- 用户下单中地区的变化数目
- ……

行为特征群

- 用户有无发生特定的浏览关注行为
- 用户对商品浏览关注的行为程度
- 用户购买的时间差的稳定程度
- 用户有浏览关注的商品的相关特征
- ……

选取订单中品类为30和101的交易信息

特征工程

特征群（时间特征）

订单

- 用户购买订单前后时间差的平均间隔
- 用户购买30/101品类的订单平均时间间隔
- 用户购买非30/101品类订单时间间隔
- 用户倒数10次的购买日期距离9月1的时间差
- ……

商品

- 用户购买同一商品的时间的平均间隔
- 用户购买不同商品的平均时间间隔
- 各商品最后购买的日期距离预测月的时间差
- …….

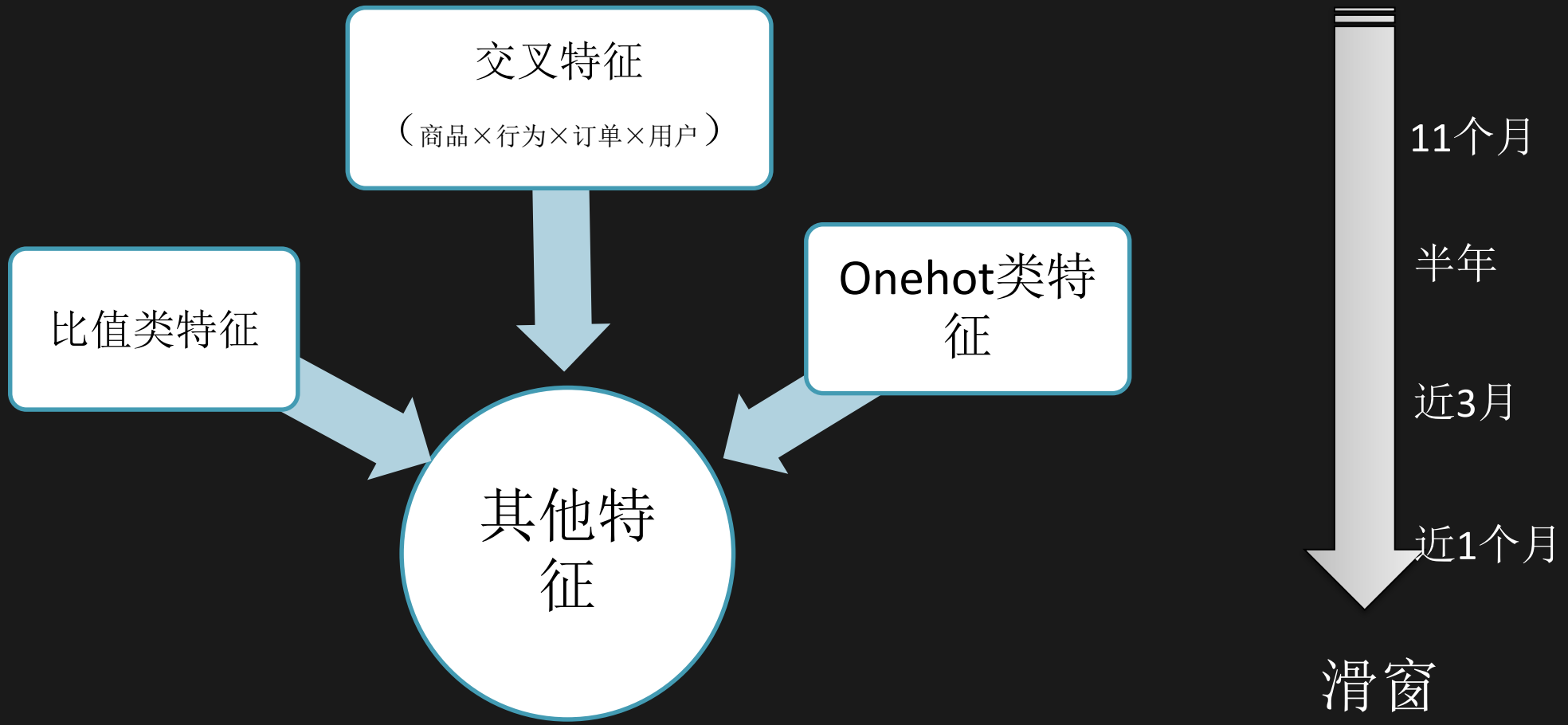
行为

- 用户开始发生行为距离购买日的时间间隔
- 用户购买前后的行为天数间隔
- 用户发生浏览/收藏行为的时间间隔

…….

特征工程

特征群（其他特征）



JDATA

特征工程

特征处理

Log（x）

处理具有长尾分布的特征

异常值处理

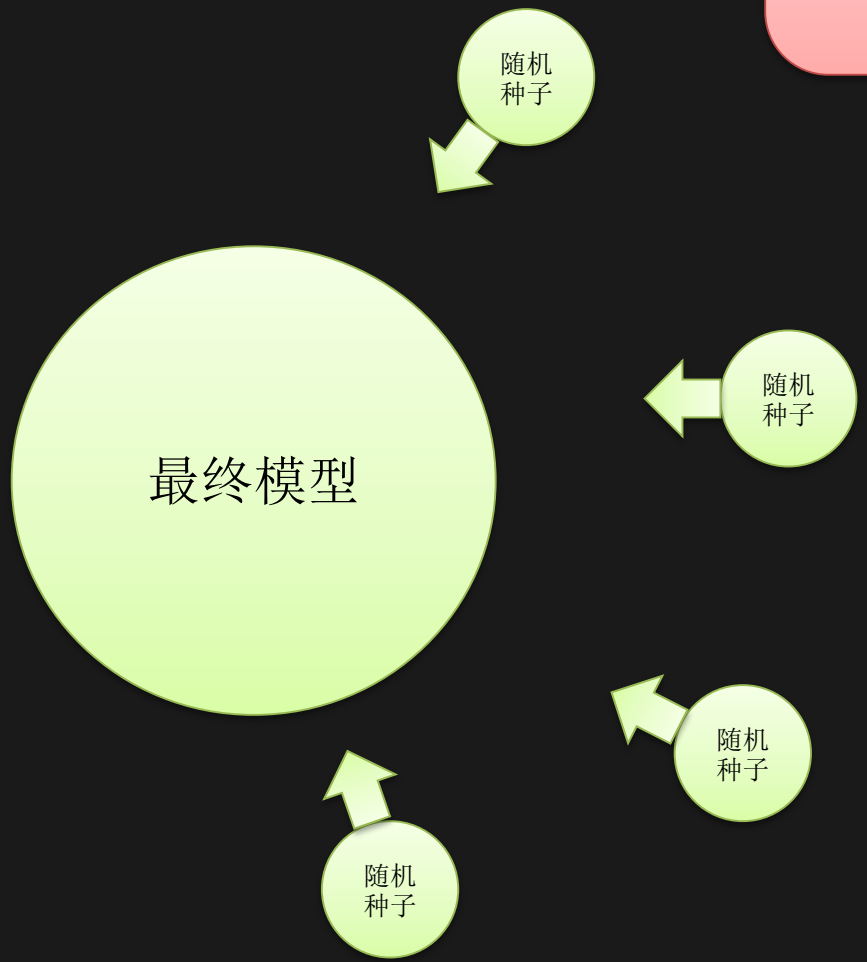
对离群点进行均值填充（比例特征、统计特征）

分箱处理

将取值相近，连续的特征进行离散化

模型融合

我们采用简单的随机种子来切分训练集方式来增强LightGBM模型的鲁棒性



JDATA

目录

1 团队介绍

2 算法核心设计思想

3 比赛经验总结

JDATA

思路优势：

这次比赛我们能有幸进入top10,有很大一方面原因在于我们仔细分析了题意，并基于题目意图和数据的可视化分析，通过合理的规则建立了相对合理的模型。

队伍优势：

由于我们队员都在一个实验室，所以平时有想法的时候也方便探讨交流。

不足之处：

虽然我们在S1上面取得了一些进展，但在S2方面，虽然我们做了很多尝试，但效果并不非常明显，这也是我们希望学习的。


Thanks
JDATA