# 中国大数据算法大赛-用户购买时间预测

队伍名称: 珞珈山第一菜鸡

演讲者:王贺

2018.07.19

# 目录

- 1 团队介绍
- 2 算法核心设计思想
- 3 比赛经验总结

# 1.团体介绍











六号: (队长) ijcai前30,华为前 20, 原名刘好

亚克西: (成员) ijcai前30,原名王 (成员)ijcai前30, 超

鱼遇雨欲语与余: 原名王贺

小幸运: (成员) ijcai前20,拍拍贷 ijcai前30,原名赵 腾讯比赛11名, 前5, 原名张浪浪 成伟

zhao: (成员)

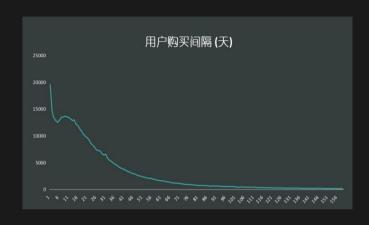
# 目录

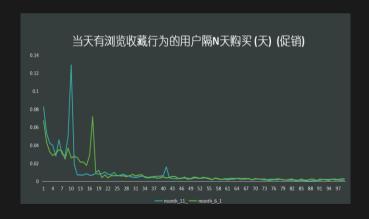
- 1 团队介绍
- 2 算法核心设计思想
- 3 比赛经验总结

# 数据分析

我们主要分析的是用 户购买间隔的来做的 数据分析,如右图

- ▶ 用户购买间隔
- 有收藏行为的用 户购买间隔天数
- 有浏览行为的用户购买间隔天数(分为促销和非促销)





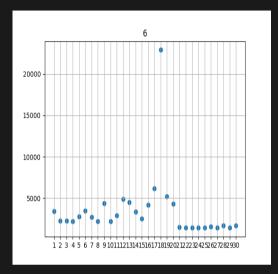


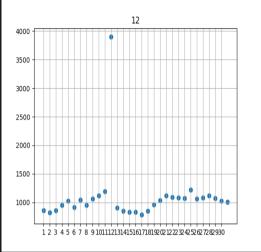


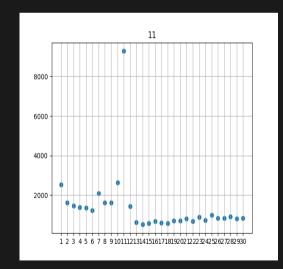


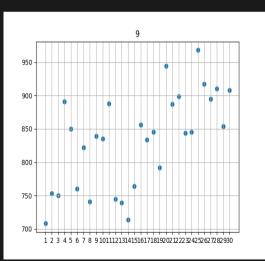
### 数据处理

- ▶ 对订单表和行为表去重,减少了噪声
- 缺失值填充:离散特征众数填充,连续 特征均值填充
- ▶ 删除缺失值较多的特征,例如sex特征, 近一个千分点的提升
- ▶ 看右边的图表,可以看出618,双十一 和双十二的流量出现异常,因此我们做 s2时去掉了一些节日



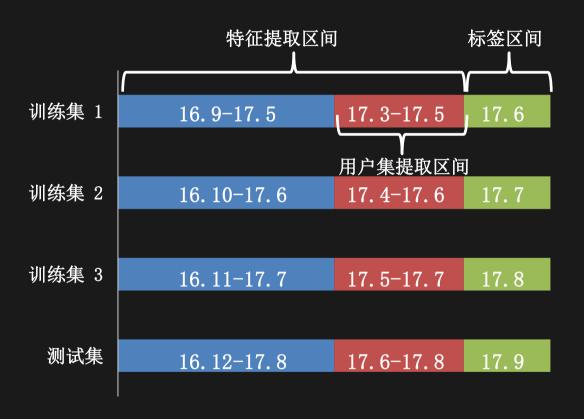




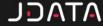




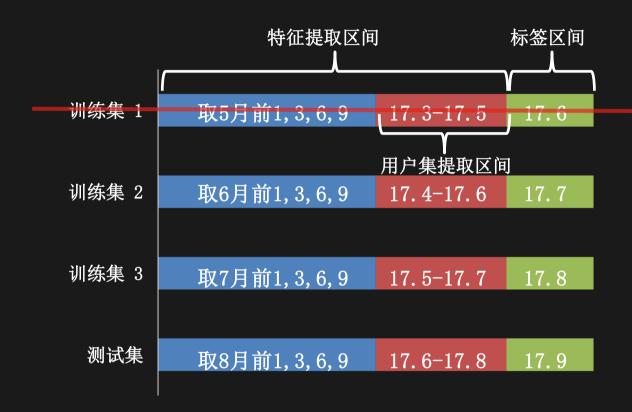
## S1训练样本构建



- 4组样本分布不同,标记区分样本组别, A榜带来一个百分点的提升
- ▶ 线下: 训练集1、2 验证集: 3
- ▶ 线上: 训练集1、2、3 测试集
- ▶ 标签日前9个月提取特征
- ▶ 标签日前3个月构造用户集合
- ▶ 与线上评测保持分布一致,2<sup>~</sup>3个千分点的提升



## S2训练样本构建



- 考虑618影响过大,最后S2只用了训练集2和;
- ▶ 线下: 训练集2 验证集: 3
- ▶ 线上: 训练集2、3 测试集
- ▶ 标签日前1,3,6,9个月提取特征
- ▶ 标签目前3个月构造用户集合



### S1主要特征

#### 用户订单表特征 用户行为表特征 用户评论表特征 重要的时间特征

> 用户订单数

. . . . . .

- ▶ 用户购买了几个月
- ▶ 用户连续购买了几个月
- ▶ 用户购买目标品类商品 的价格统计特征群
- 用户购买目标品类商品的属性统计特征群

- > 用户浏览的天数
- ▶ 用户浏览了几个月
- 用户连续浏览了几个月
- ▶ 用户浏览目标品类商品 > 的价格统计特征群
- ▶ 用户浏览目标品类商品 ▶ 的属性统计特征群

. . . . . .

- 用户评论的最早、最晚、 平均时间
- 用户最后评论与最后购买的时间差
- 用户最后评论与最后浏览的时间差
- 用户最后评论距离标签日的时间差

.....

- 用户购买的最早、最晚、 平均时间
- 用户浏览的最早、最晚、 平均时间
- ▶ 用户评论的最早、最晚、 平均时间
- ▶ 用户购买的时间间隔的 统计特征群
- ▶ 用户浏览的时间间隔的 统计特征群
- ▶ 用户最后购买距离标签日的时间间隔
- ▶ 用户最后浏览距离标签日的时间间隔

.....



### S2主要特征

#### 基本特征

#### 统计特征

#### 时间特征

#### 比率特征

- 用户年龄
- 用户等级
- 商品属性
- 商品价格 . . . . . .

- 基于商品统计特征 (商品属性、价格、种类等)
- 基于行为的统计特征 (用户行为类型,行为次数 等)
- 基于订单的统计特征 (购买数量,购买次数等)
- 基于评论的统计特征 (评论次数、用户等级等)
- > 交叉统计特征 (基于用户id和用户等级的商 品统计特征等)

- 时间差特征 (比如用户每次订单的间隔 时间特征,用户行为的间 隔时间等)
- 时间统计特征 (用户首次购买的时间,最 后一次购买时间,每个月 首次购买时间, 平均第几 周购买,购买的天数等)

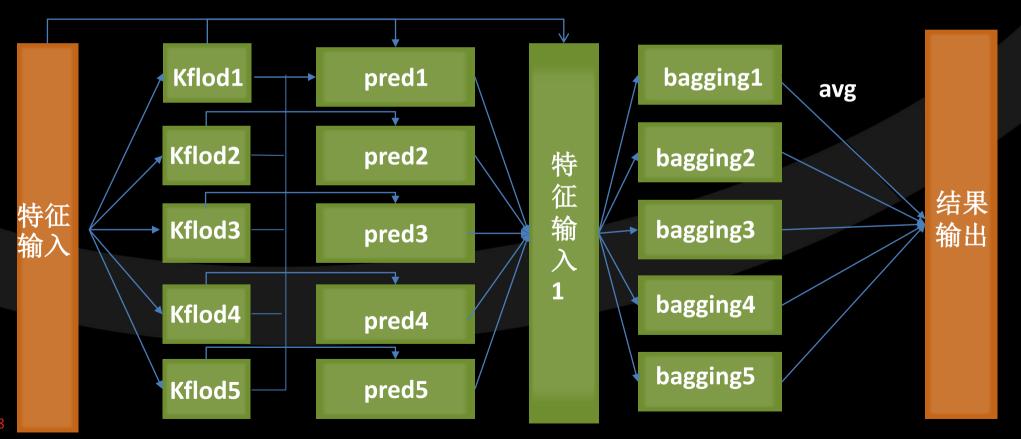
- 用户购买非目标的数量/购 买目标的数量
- 用户行为后购买的次数/用 户行为的总次数
- 不同年龄购买目标数量/该 年龄的购买总数量

- ▶ S2是预测时间,肯定和时间特征联系紧密, 因此做了很多时间特征
- 其他特征就是根据实际业务逻辑提取



# 模型模块

- 模型使用lightgbm和xgboost
- Cross validation,增强模型鲁棒性
- Stacking和bagging,保证模型的精度



# 目录

- 1 团队介绍
- 2 算法核心设计思想
- 3 比赛经验总结



#### 比赛经验总结

- > 滑窗采样的同时需要注意每个窗口的正负样本分布
- ► 每天评测次数有限,需要保证线下的验证结果可靠,因此构造数据 时线上线下需要保证分<u>布一致</u>
- ▶ 比赛中S2部分没有尝试使用线上评测函数作为目标函数来训练模型,有点遗憾,据说提升很大。
- ▶ 对于618和双十一的噪声数据没有处理好,这部分和前排差距明显。



### 致谢

- ▶ 感谢京东主办这次比赛,让我们能够接触到真实的业务数据,在 比赛不断探索的过程中得到了锻炼和展示。
- ▶ 感谢比赛以来帮助过我们的朋友,以及给我们解决问题的相关工作人员。



