

中国大数据算法大赛-用户购买时间预测

队伍名称：D国反击战

演讲者：段文强

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

团队介绍（绝Data武士）



张杰

趋势科技数据科学家
天池数据科学家
天池乘用车销量
美年健康AI亚军

Datacastle 金融赛二等奖
Corporación Favorita Grocery Sales
Forecasting top 1%



卢杰

4399算法工程师
2017CCF大数据与
计算智能大赛冠军
Kaggle:



段文强
江西财经硕士

神秘？



杜旭浩

西澳大学机械博士
诺贝尔奖实验室
马歇尔疾病中心研究员
融360数据风控大赛冠军
首届滴滴大赛top 1%



刘洋

东南大学博士
天池数据科学家
天池天文赛，
安全赛冠军
IEEE UAI 冠军

JDATA

目录

- 1 团队介绍
- 2 算法核心设计思想
- 3 比赛经验总结

赛题分析

- 任务重述
- 数据描述
- 解题思路
- 样本集构造
- 数据集划分

任务重述

任务描述 根据赛题方提供的数据，预测未来1个月内最有可能购买目标品类的用户，并预测他们在考察时间段内的首次购买日期。

评价指标

$$S1 = \frac{\sum_{i=1}^N w_i o_i}{\sum_{i=1}^N w_i} \quad w_i = \frac{1}{1+\ln(i)}$$

$$S2 = \frac{\sum_{u \in U_r} f(u)}{|U_r|} \quad f(u) = \begin{cases} 0, u \notin U_r \\ \frac{10}{10+d_u^2}, u \in U_r \end{cases}$$

$$S = 0.4 * S_1 + 0.6 * S_2$$

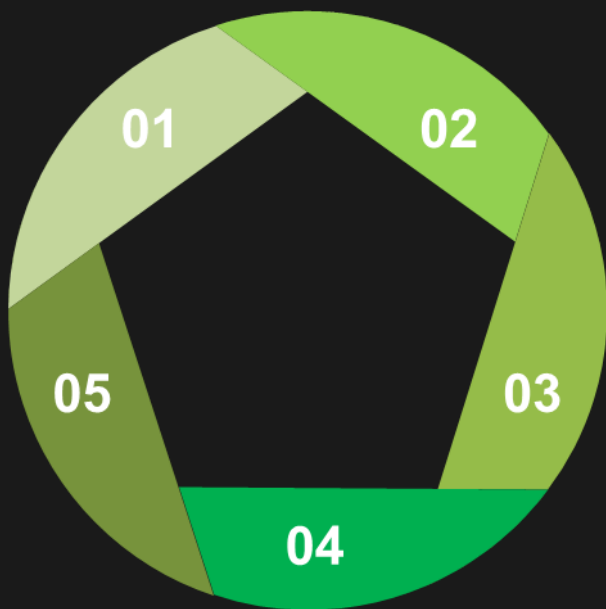
数据描述

01.sku基本信息

商品标识
价格
品类
参数一、二、三

05.评论分数数据表

用户标识
评论时间
下单标识
评分级别



04.用户订单表

用户标识
商品标识
下单标识
下单日期
下单区域
下单件数

02.用户基本信息表

用户标识
年龄
性别
用户等级码

03.用户行为表

用户标识
商品标识
行为日期
行为次数
行为类型

解题思路

单任务 预测未来一个月的每一天用户是否购买目标品类商品，二分类问题，数据量 $99446 \times 30 = 2983380$

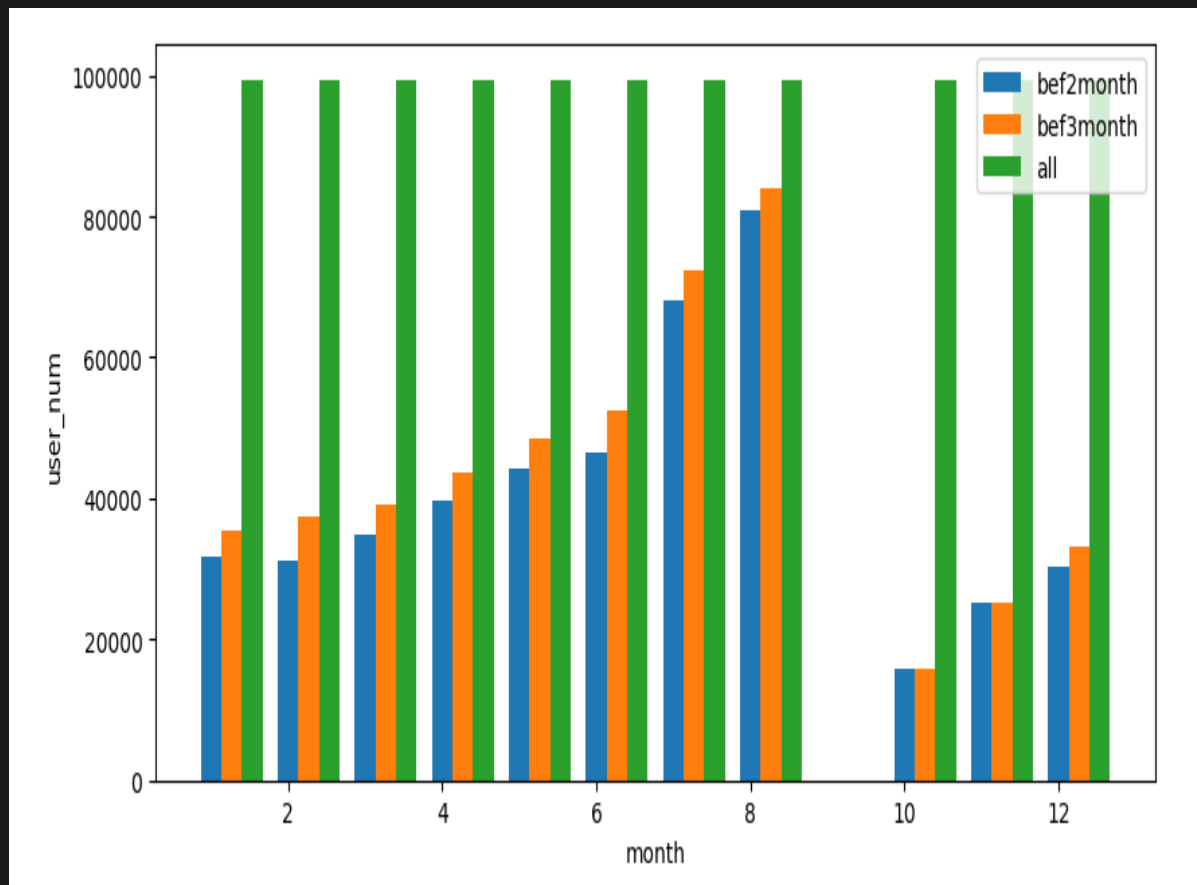
多任务 预测未来一个月用户是否购买目标品类商品及用户未来一个月目标品类的首次购买时间，数据量 $99446 + 99446 = 198892$

样本集构造

全集用户 每月样本都用全集用户构造

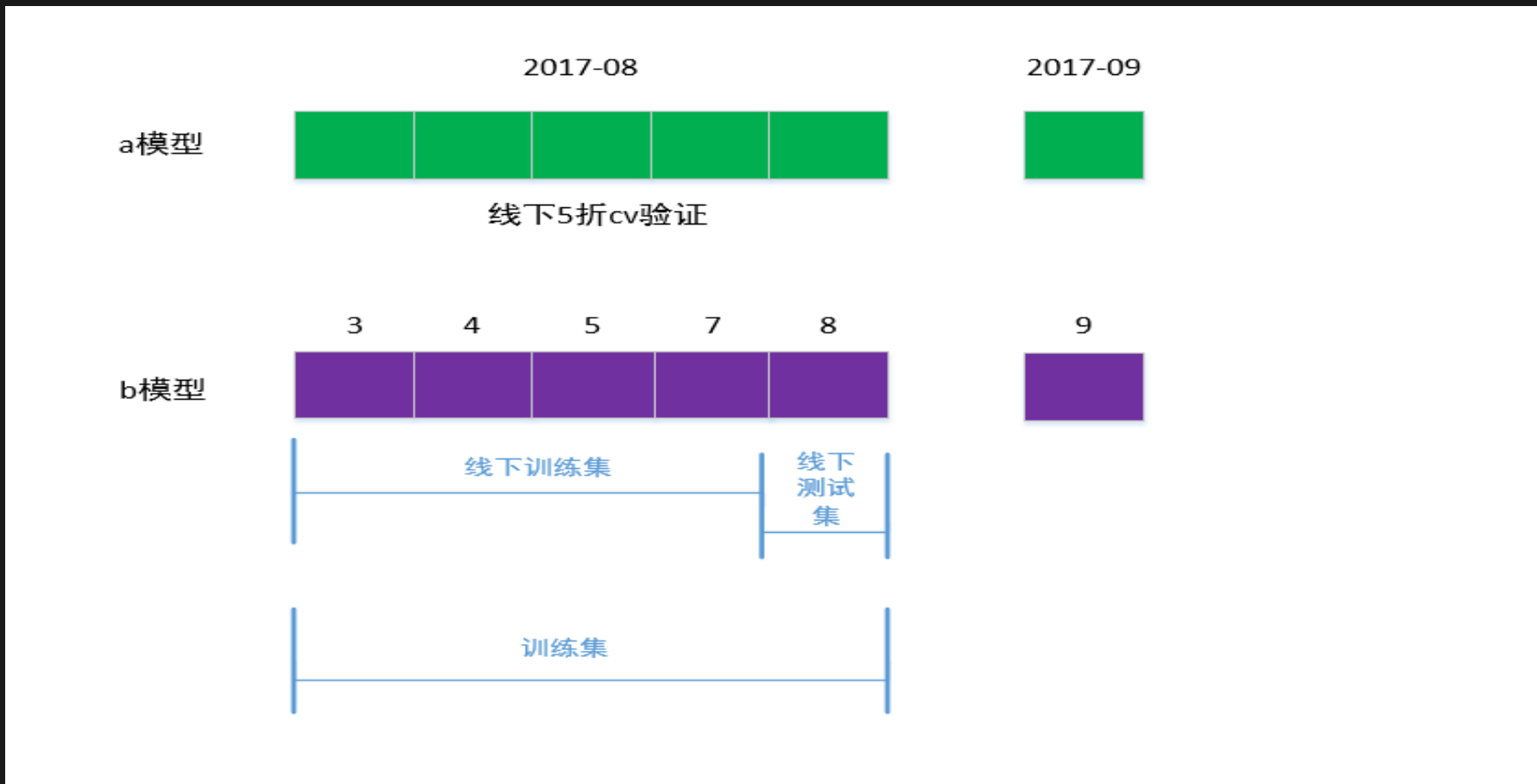
部分用户一 每月的样本都用前三个月购买过目标品类的用户构造

部分用户二 每月的样本都用前两个月购买过目标品类的用户构造



数据集划分

S1数据集划分



数据集划分

S2数据集划分

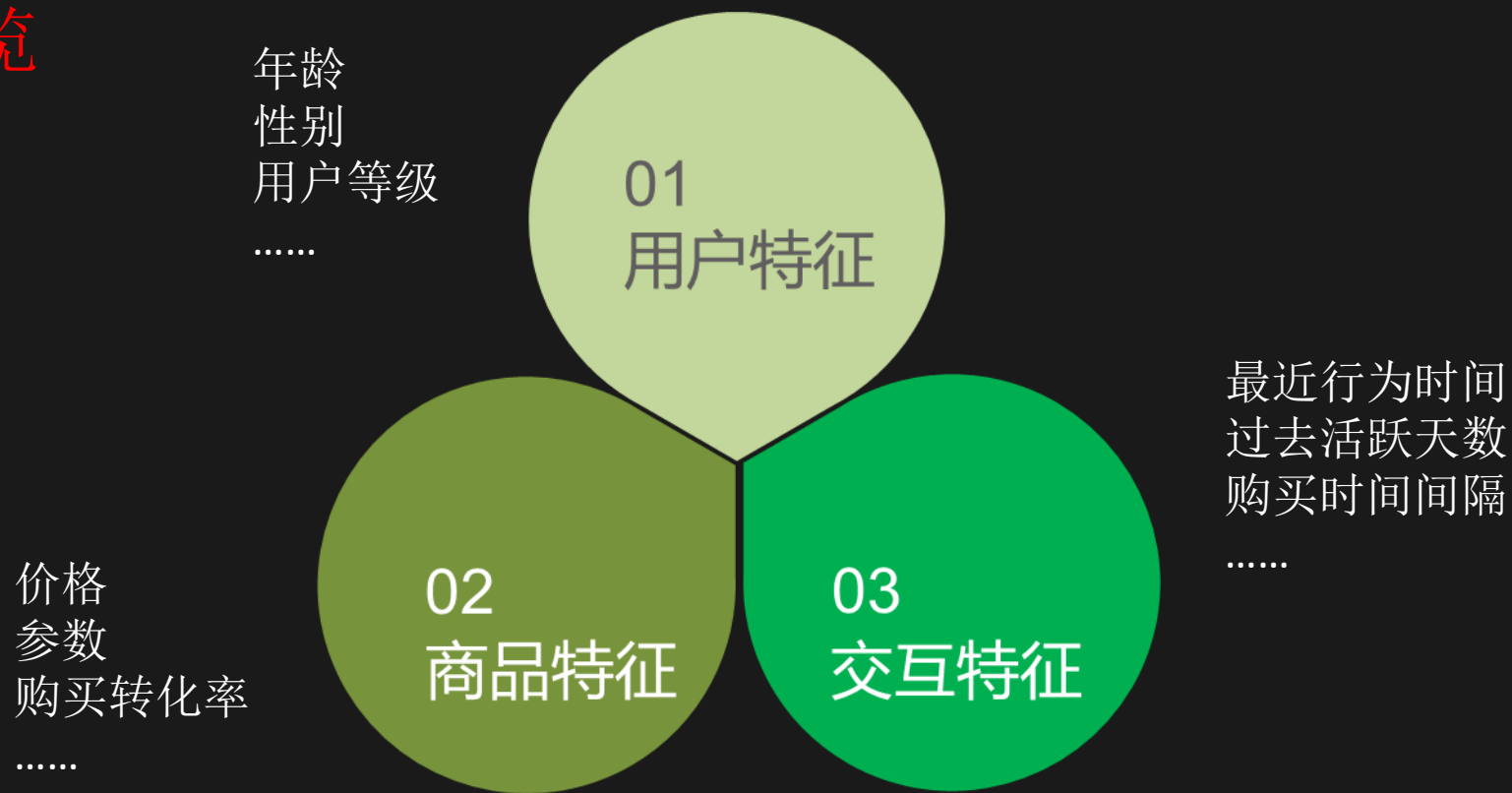


核心思路

- 特征工程
- 算法选择
- 特征选择
- 模型构建

特征工程

特征概览



特征工程

特征构造

初级特征：用户年龄、性别等；

二级特征：用户平均购买时间间隔、平均行为时间间隔等；

三级特征：用户sku的平均购买间隔再平均、sku平均行为间隔再平均等。

固定步长滑窗特征：以21天和30天分别为步长来滑窗提取用户行为特征和用户订单特征；

随机步长滑窗特征：以30天、45天、63天、78天、90天、121天.....不同时间段来提取交互特征。

算法选择及特征选择

XGBOOST

- 高效的C++实现
- 多线程并行计算
- 优秀的性能

LIGHTGBM

- 更快的训练速度
- 更少的内存消耗
- 善于处理大规模数据

特征选择

根据相关性筛选：部分特征相关性较高，设定的阈值是0.95。

根据队友杜旭浩开源的特征选择代码进行筛选，包括：

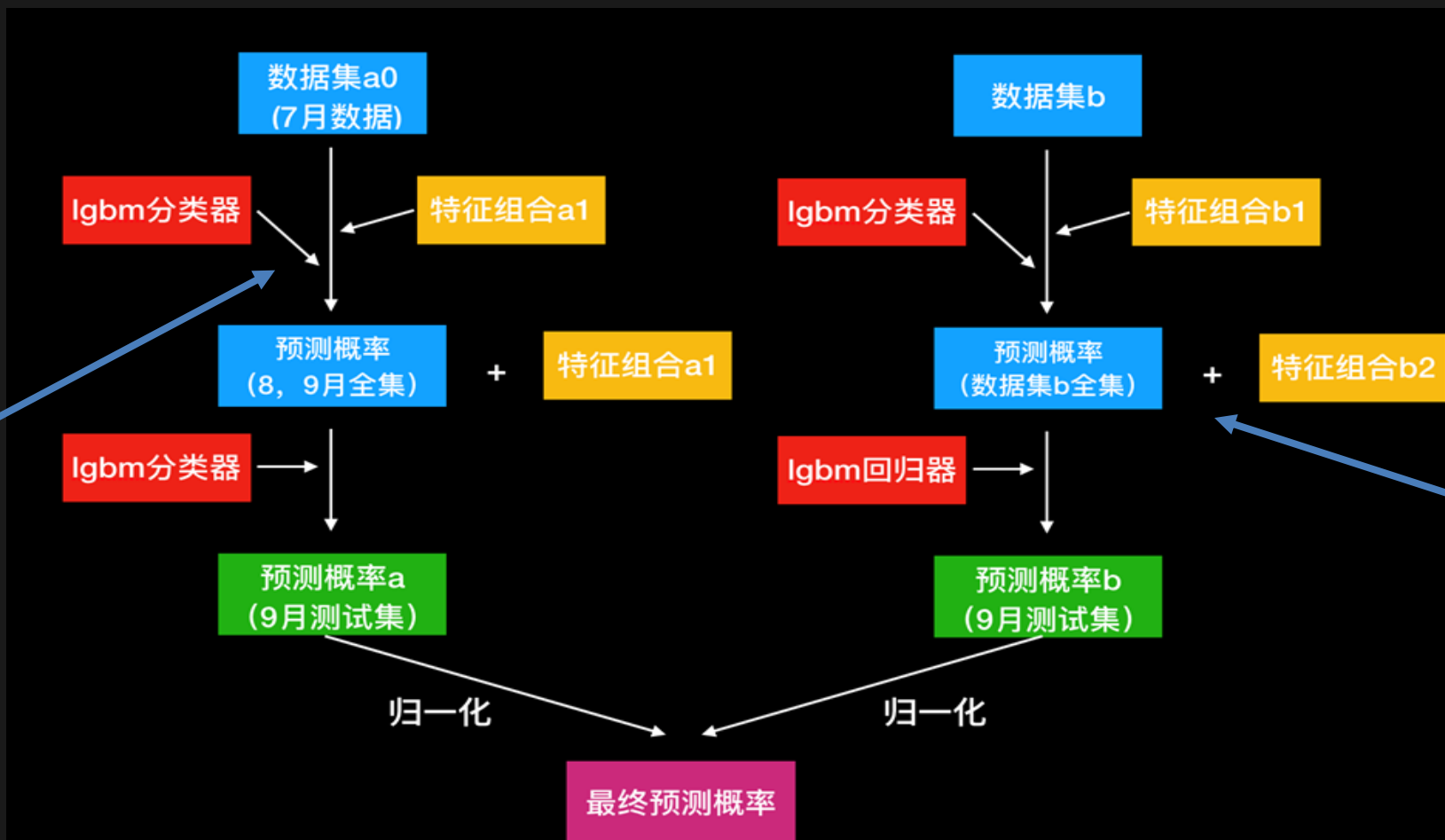
1. 相关性筛选
2. 重要性筛选
3. 贪心筛选

<https://github.com/duxuhao/Feature-Selection>

模型构建

S1模型框架

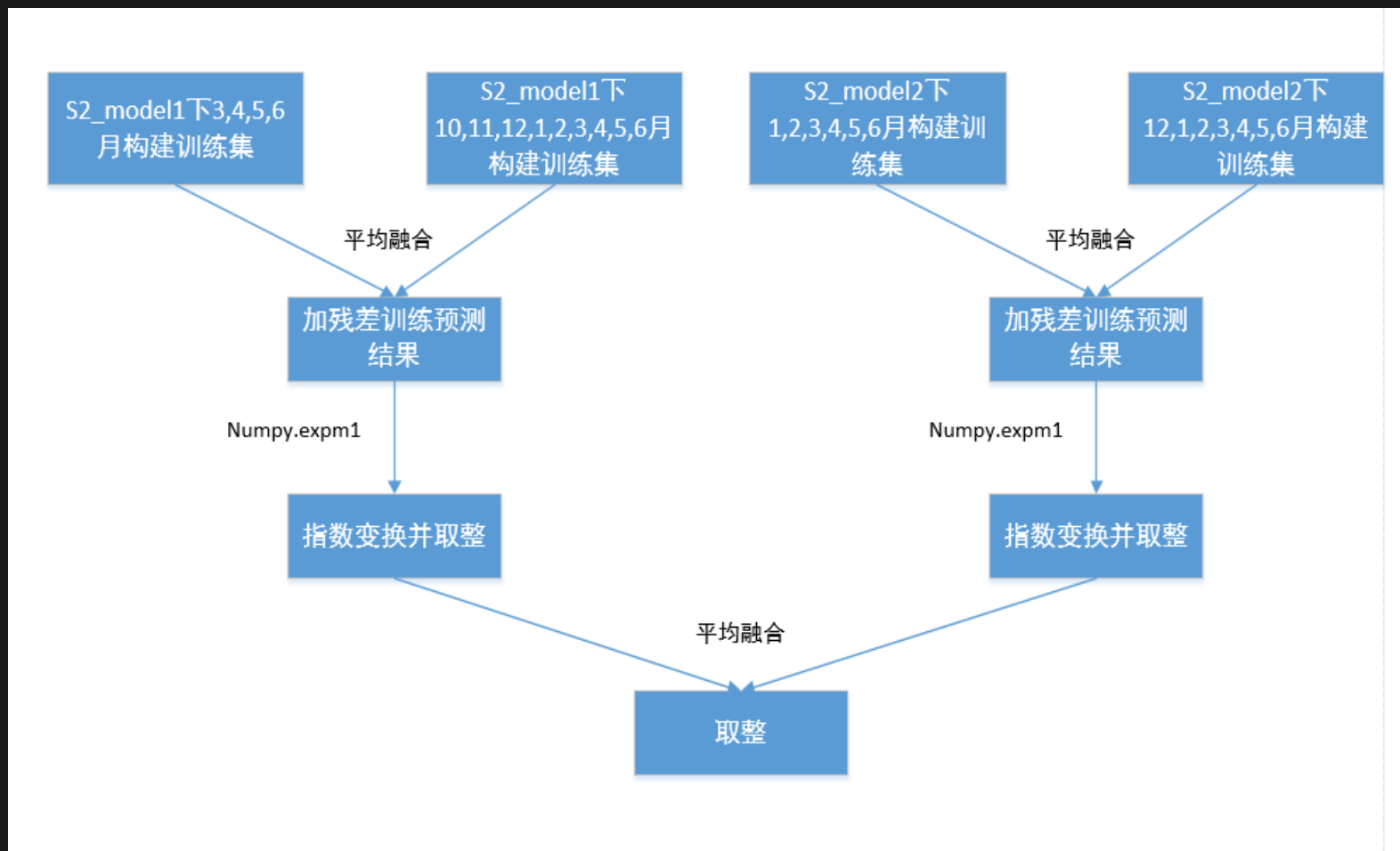
迁移过去一个月的信息进行二次预测



针对瓶颈进行二次预测

模型融合

S2模型框架



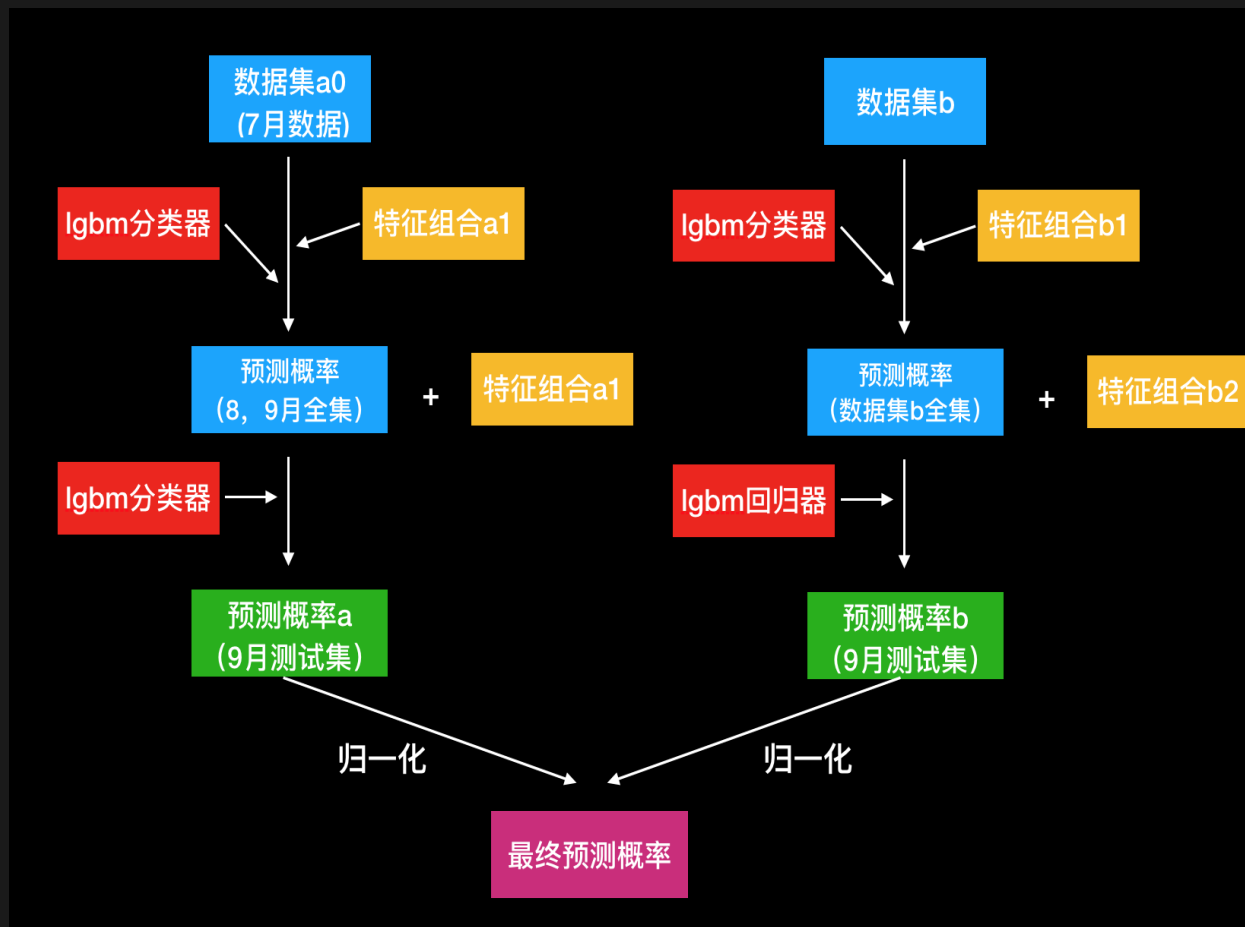
细节解析

- 模型构建
- 特征选择
- 残差训练

S1模型构建

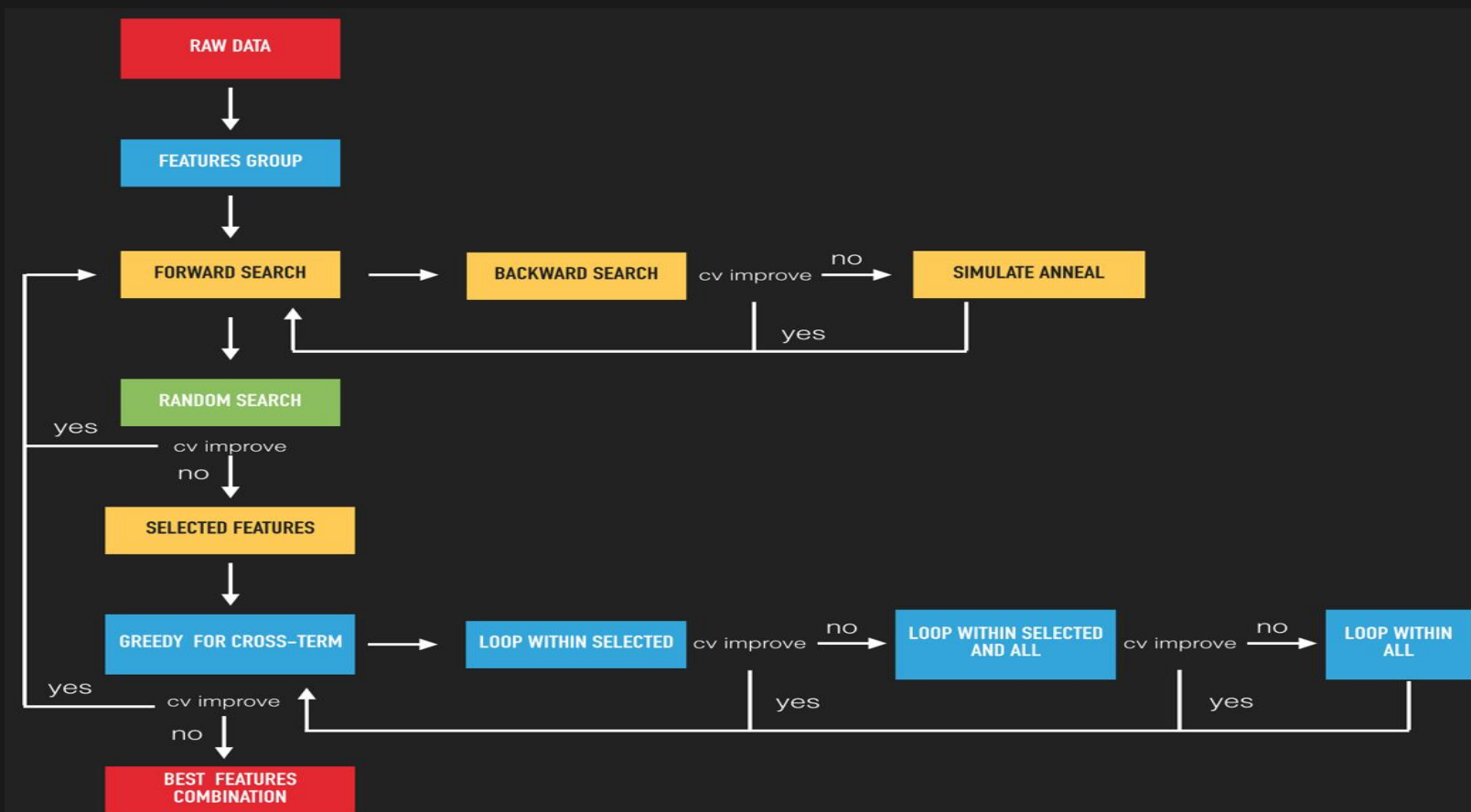
a模型：借用了植物大神迁移学习的思想，7月为样本的数据集训练一个模型并预测8月9月的作为新的特征，再进行训练，这个目的是在保证数据无泄漏的情况下（只用6，7月有交互的用户构建数据集），更多地去使用到过去的的数据信息。

b模型：主要是因为在经过线下瓶颈分析后发现相当部分用户之前所有月份都进行购买后，最后一个月突然没有购买，这部分人可以额外的用一些选择过的特征来进行分辨，因此在第一次预测的基础上进行了二次的预测。

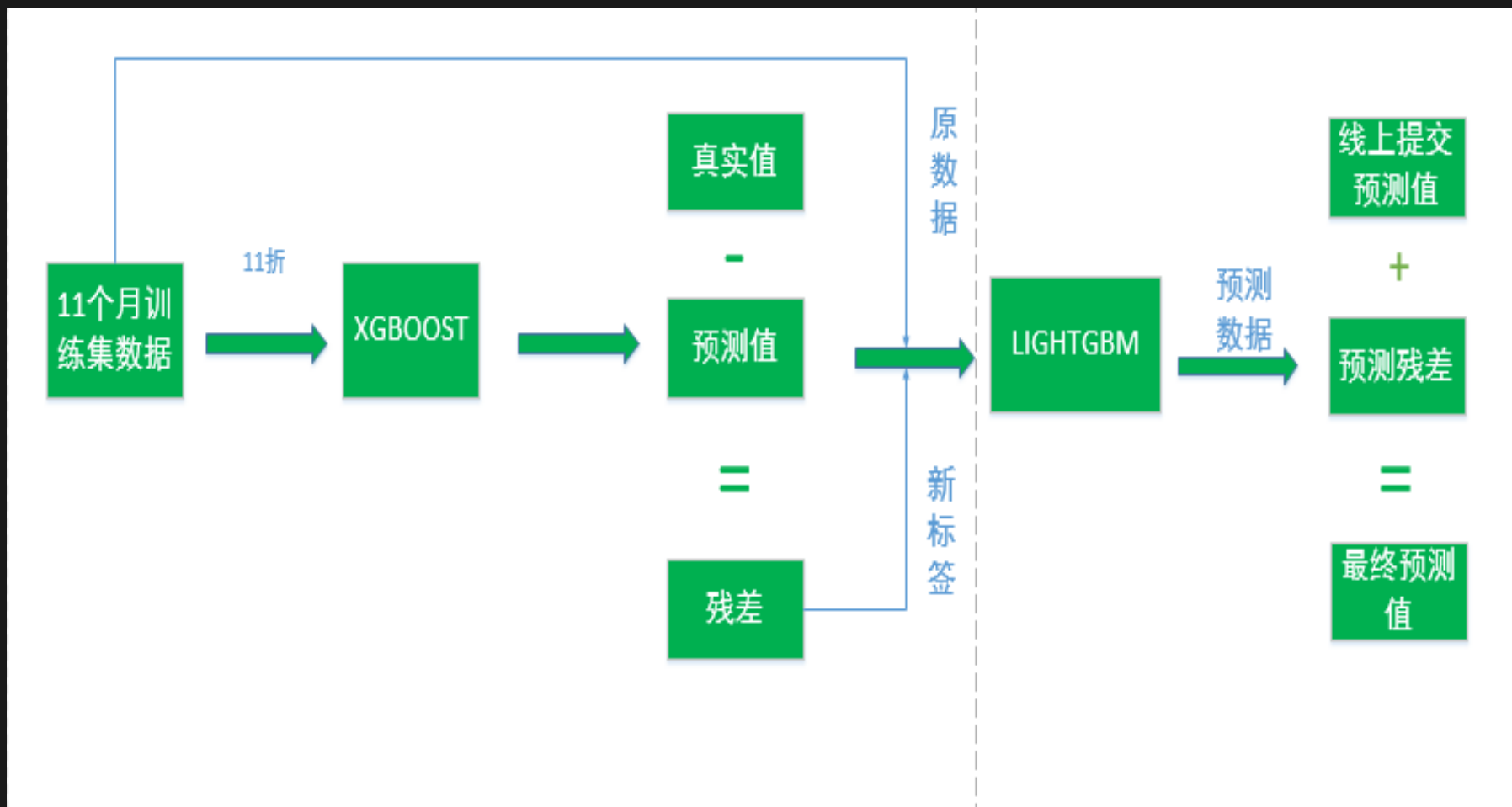


特征选择

经过选择后的特征具有运行时间短，需要内存小的，而且在准确率上也有提升的优点。选择的特征主要是借助我们的队友杜旭浩在网上开源的特征选择库进行的，其部分的框架如下图。



S2残差训练



目录

- 1 团队介绍
- 2 算法核心设计思想
- 3 比赛经验总结

算法优势



两段式学习

一个学习目标，一个学习残差，不同价值的信息都学到。



迁移学习

最大化挖掘数据集所能提供的所有信息



特征选择

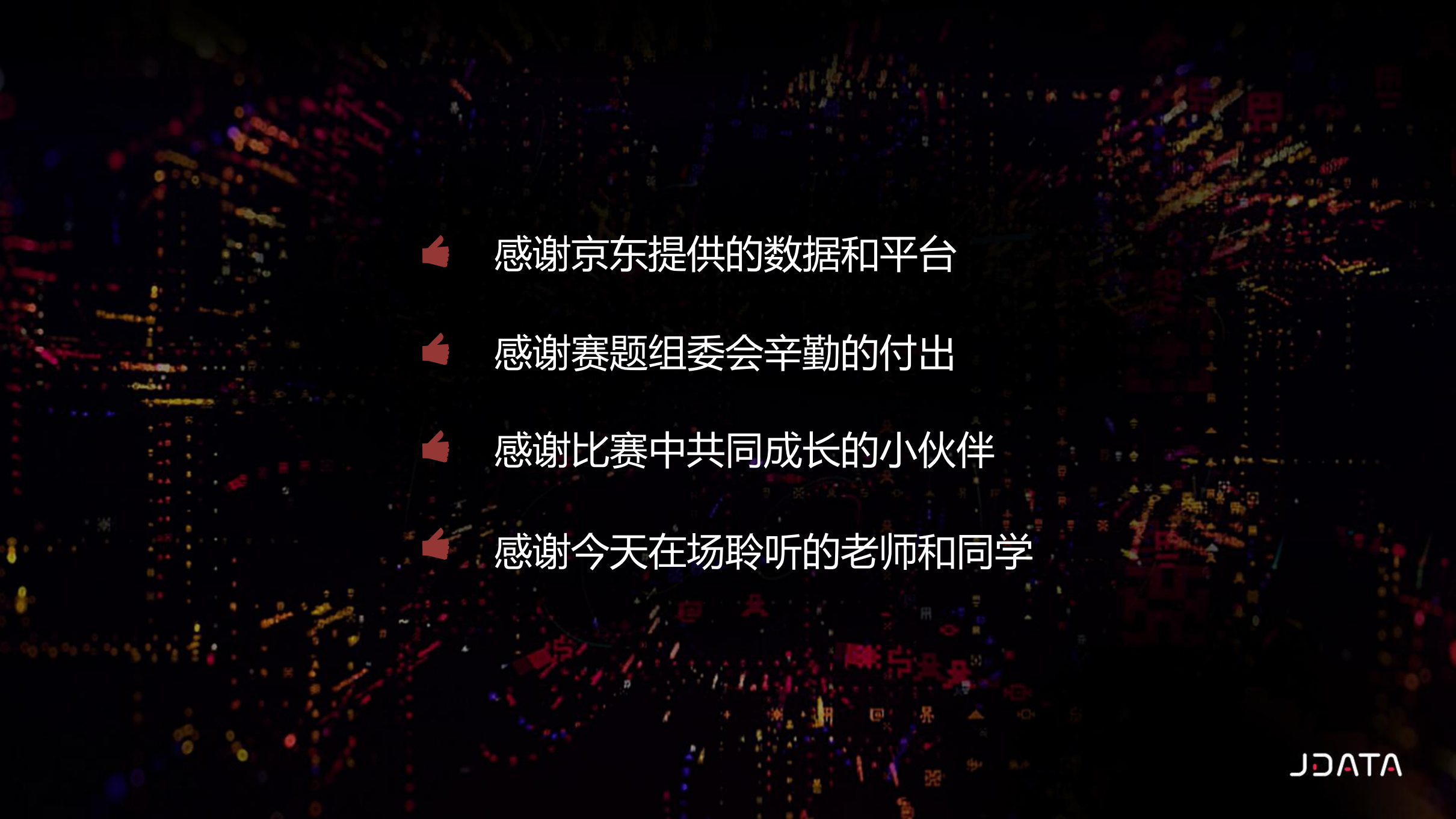
能尽可能减少训练和运行时间，并能对模型解释度有一定提升。

经验总结

数据探索 做简单的统计，时序数据按时间分割寻找规律，或对历史记录进行探索，作图分析，通过各种方式加深对数据理解。

问题理解 赛题相关问题往往是有相关资料可查询的，可以查找相似的题目及解决方案，或者查询论文。结合现有解决方案和赛题特点，提出自己的解决方案。

选手交流 不同的人对赛题理解是不同的，多和选手沟通交流，往往会碰撞出不一样的解题思路，三个臭皮匠顶个诸葛亮，学习别人的优秀方案，也思考自己解决方案，多思考才更有助于加深对赛题的理解。

- 
- 感谢京东提供的数据和平台
 - 感谢赛题组委会辛勤的付出
 - 感谢比赛中共同成长的小伙伴
 - 感谢今天在场聆听的老师 and 同学



That's all
Thanks