阿里移动推荐算法大赛答辩

2015天池大数据竞赛

TIANCHI天池

队名：Sahara

2015年8月18日

提纲

- 团队介绍
- 参赛历程
- 解决方案
- 总结与思考

团队介绍—队名

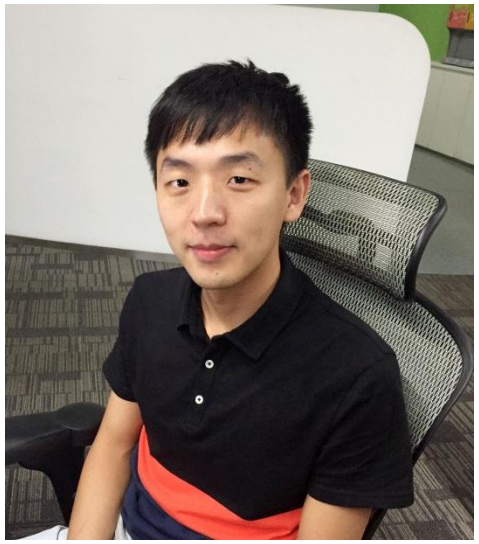
- **Sahara**
  - ✓ 撒哈拉沙漠
  - ✓ 热情、顽强、拼搏

- 关于组队
  - ✓ 混搭
  - ✓ 一面之缘、一句闲聊
  - ✓ 同一个梦想

团队介绍—队员

- 宁克锋（正澄）

清华大学



- 林浩（小眼睛）

华南理工大学

- 江少华（人大吴奇隆）

人民大学

团队介绍—分工

技术探索

基础数据分析

特征工程

方案设计

模型融合

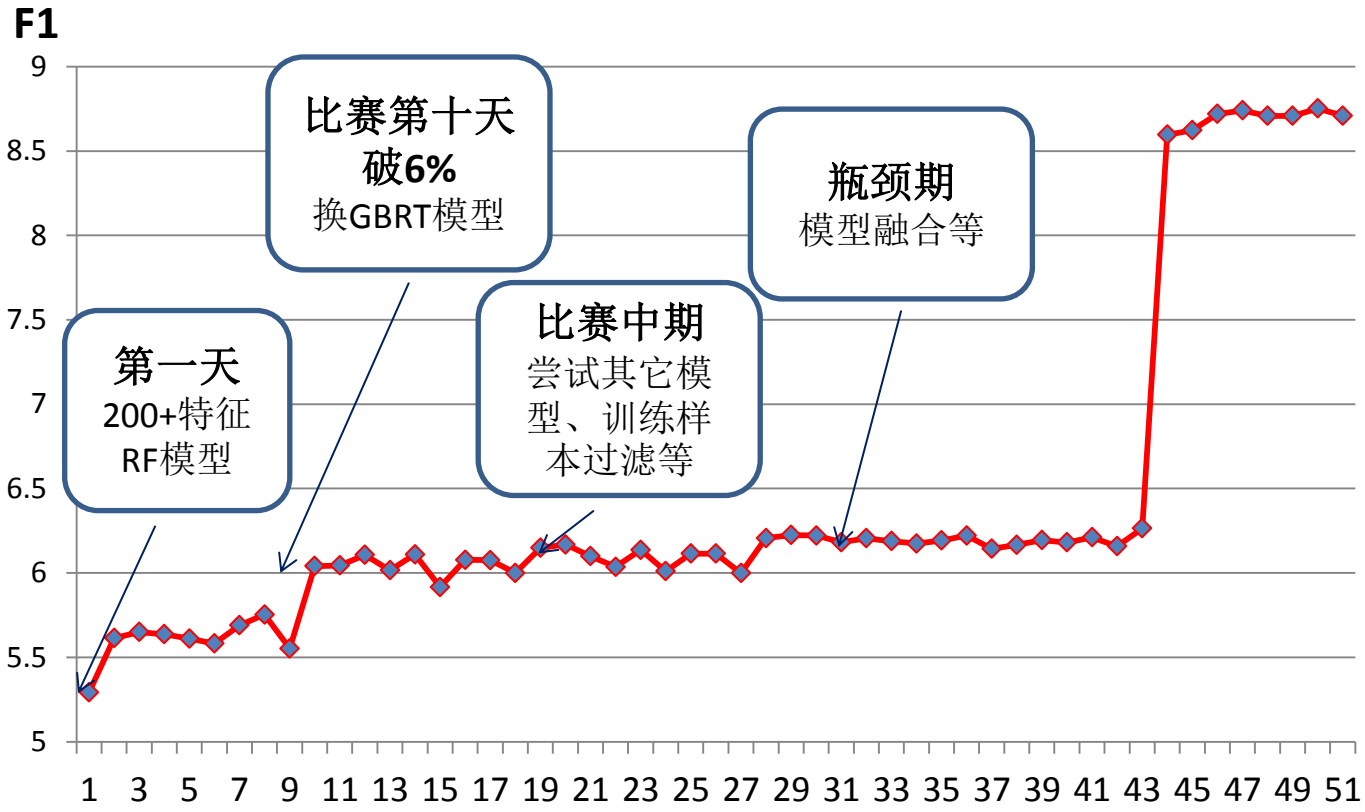
主代码手

参数调优

参赛历程

第二赛季

- Part 1
  - ✓ 榜首天数：26+3
  - ✓ 首次提交5.3%（修改提交数后5.6%）
- Part 2
  - ✓ 榜首天数：6
- 每天排名非1即2



参赛历程

又快又好，一步一个脚印，模型鲁棒性好

Part2 开始

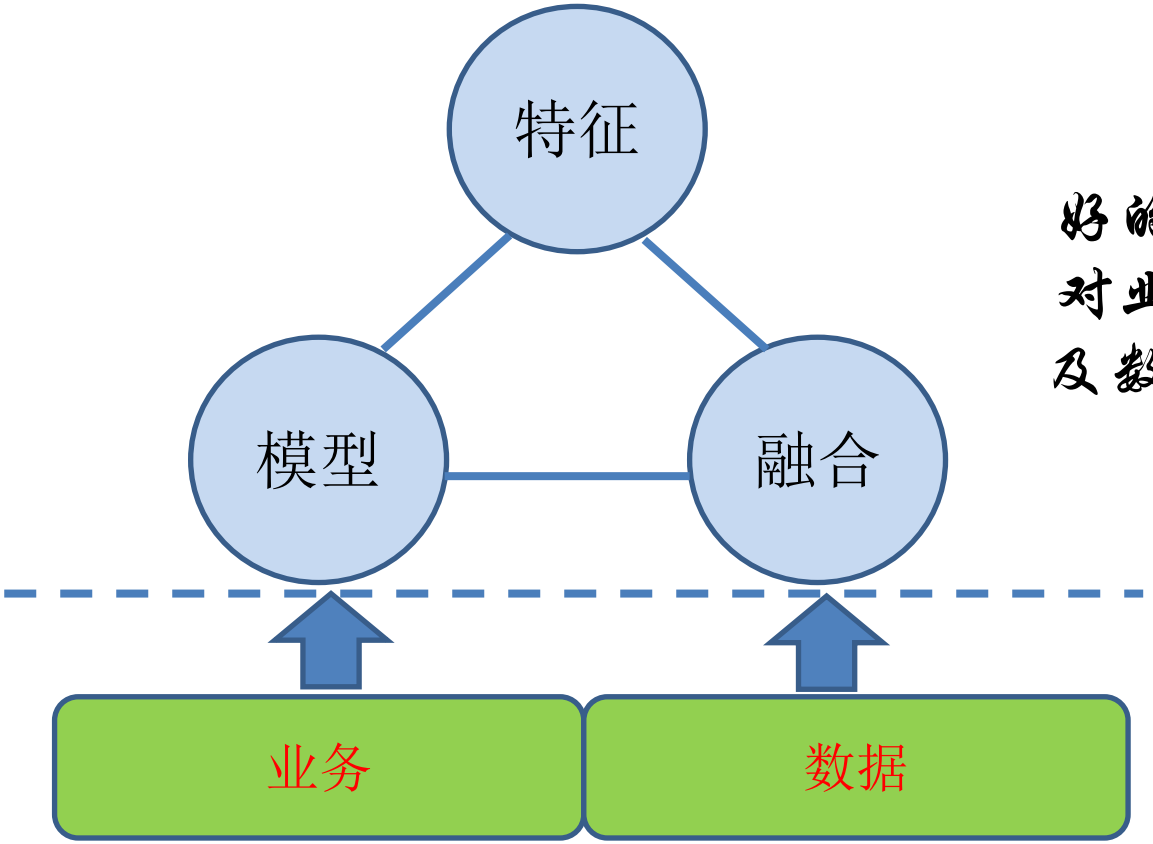
Rank	TeamName	School	F1Score	Accura	Recall	l
1	Sahara	清华大学中	8.60%	8.58%	8.61%	
2	CHLL	中国科学技	8.56%	9.86%	7.57%	
3	NEU-Smart	东北大学东	8.56%	8.50%	8.61%	
4	Constant-Penelo	华南理工大	8.52%	8.50%	8.54%	
5	北京仰望星空大学	中国科学院	8.50%	8.47%	8.53%	

Part2 结束

Rank	TeamName	School	F1Score	Accura	Recall
1	SecRet;WeaPon	南京理工南	8.78%	8.76%	8.80%
2	Sahara	清华大学中	8.75%	9.01%	8.51%
3	北京仰望星空大学	中国科学院	8.66%	8.64%	8.68%
4	NEU-Smart	东北大学东	8.64%	9.14%	8.19%
5	CHLL	中国科学技	8.63%	9.19%	8.14%

解决方案

铁人三项



两大支撑

好的解决方案源于
对业务的深入理解
及数据的细致分析！

时序日志特征提取

U:user

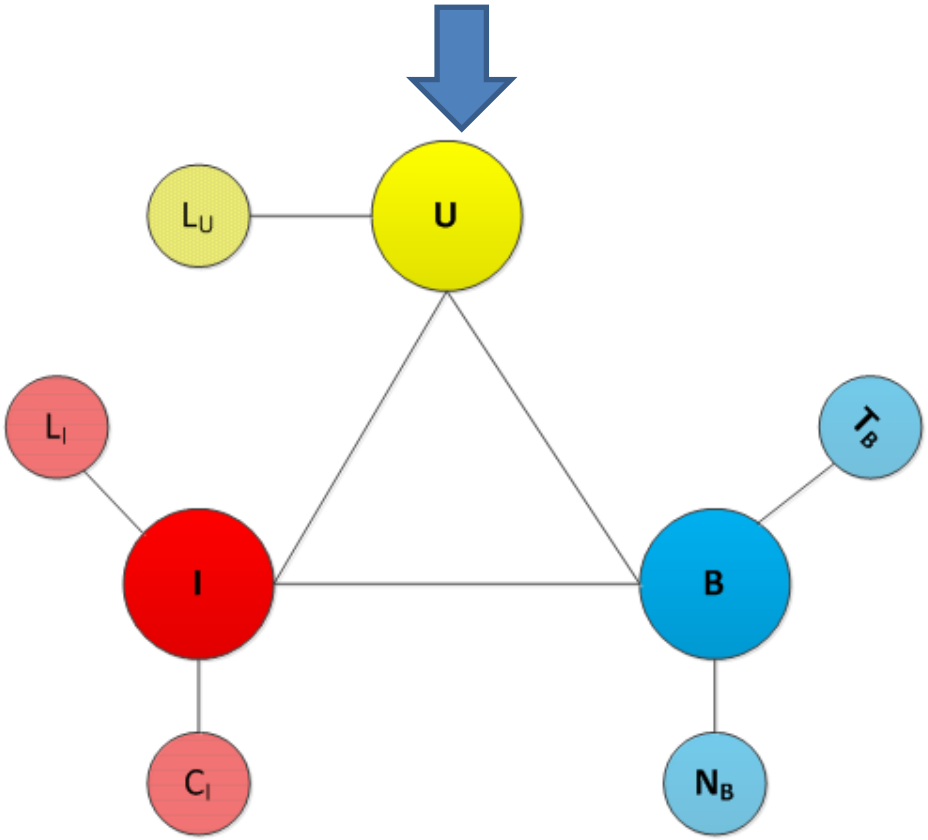
I:item

B:behavior

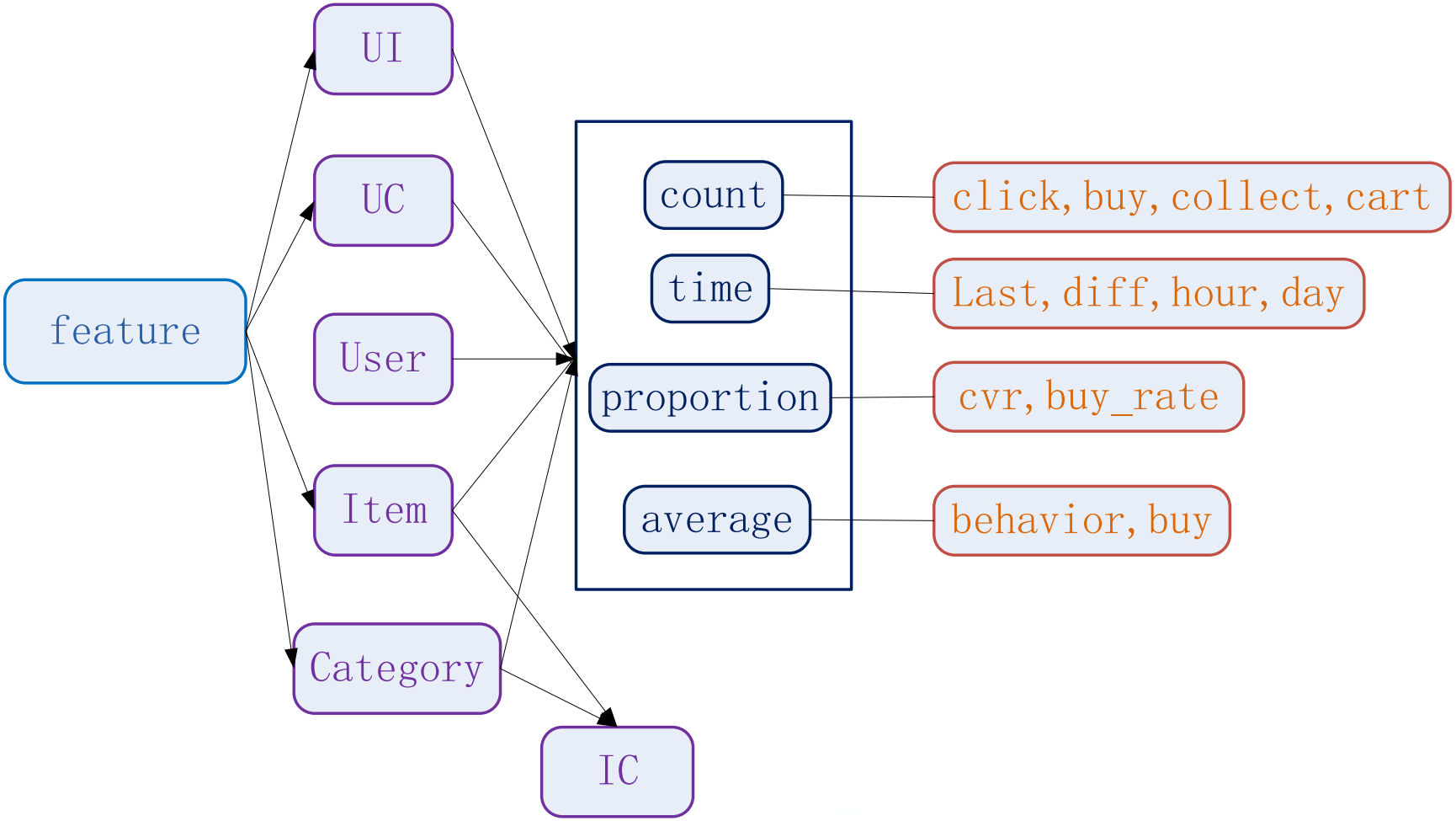
U/I/B是核心，

并且具有各自的属性

时序日志类通用的特征提取方法？



特征工程


feature
UI
UC
User
Item
Category
IC
count
time
proportion
average
click, buy, collect, cart
Last, diff, hour, day
cvr, buy_rate
behavior, buy

特征工程

为什么需要复合特征：

- **Item特征：**

  相对稀疏，信息量少

- **Category特征：**

  信息量大，但和item的购买预测相关度较小

- **IC特征：**

  复合特征

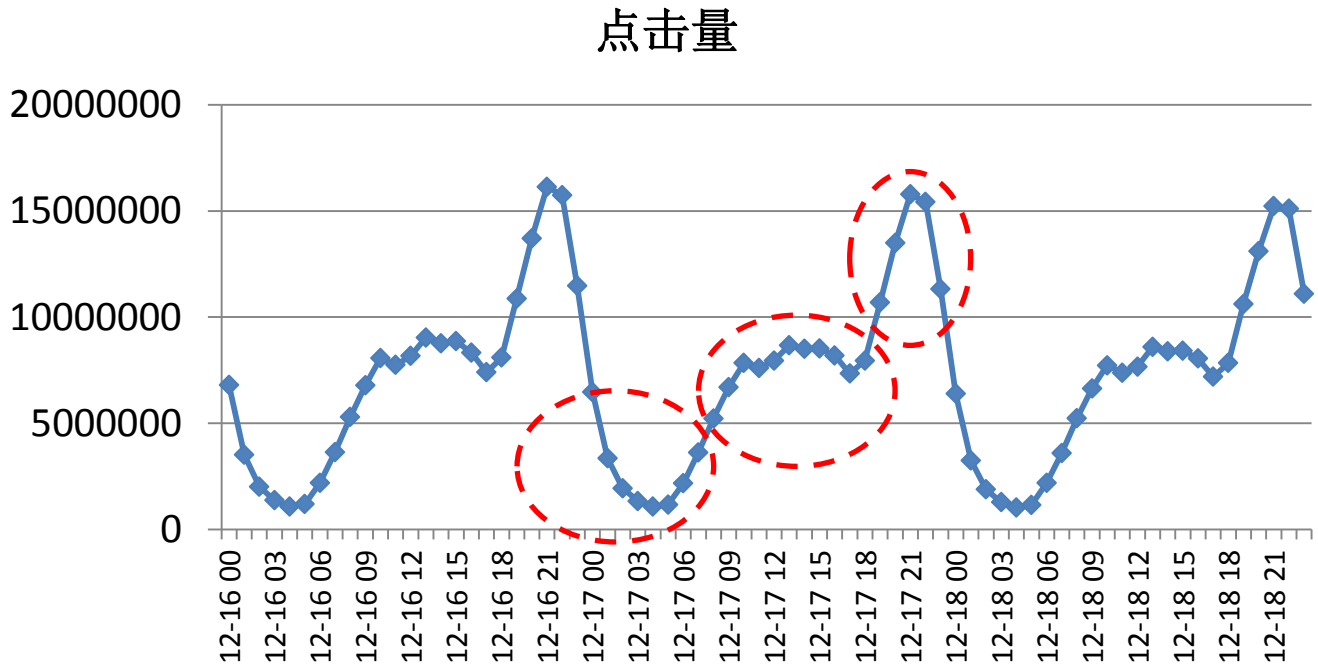
  互补Item、Category类特征的优劣

时序划分

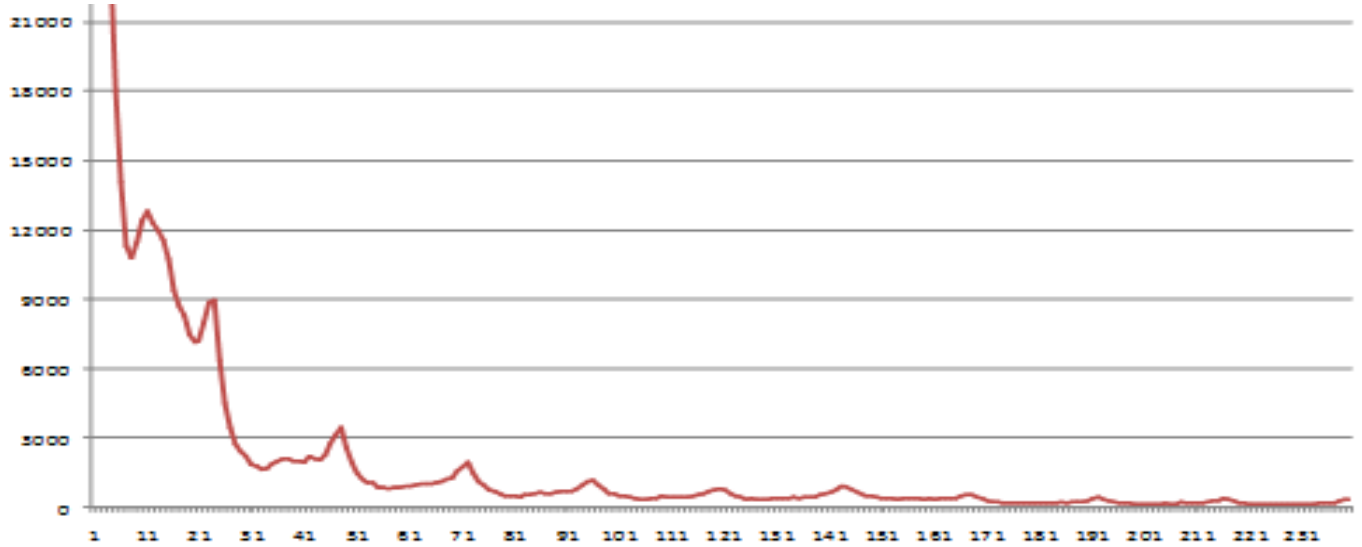
时序分段依据：

- 0:8
- 9:18
- 19~23

做细特征工程



样本选择—7天UI



购买-点击时间间隔图

- 1周（168个小时）以前的行为对当前的购买影响甚小

减少正样本的随机性

- 基于此结论进行采样，取最后7天出现过的UI的作为样本子集

特征工程—词云图


用户平均行为周期
商品双12后行为
点击变购率
用户-类目行为时间间隔
用户收藏夹
最近第3次行为数
商品在购物车中数量
用户-类目购买概率
最近第2次行为数
加购变购率
商品双12行为
最后工作日行为
用户点击--购买间隔
用户双12后行为
用户购物车
用户-相似商品行为
收藏变购率
类目双12行为
ui双12后行为
用户双12行为
商品平均行为周期
商品在收藏夹中数量
最近1次行为数
类目在收藏夹中数量

离线评测

- 离线评测是重要环节
- 减少对线上提交的过度依赖

方法：

- ✓ 特征天数/标签天数实验
- ✓ 训练集/测试集相隔多天/反转
- ✓ 多个离线评测指标：

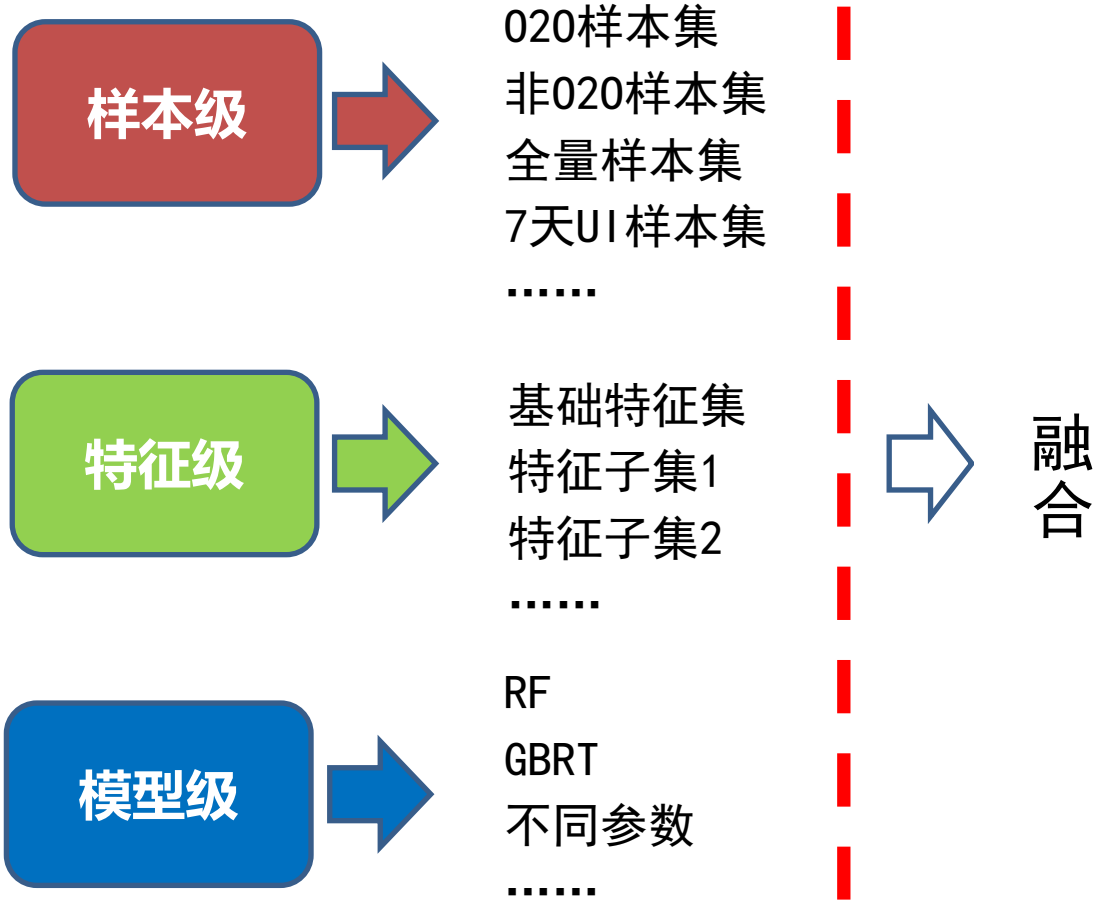
  模拟线上的50%子集的 F1

  增加AUC指标

靠谱的离线评测

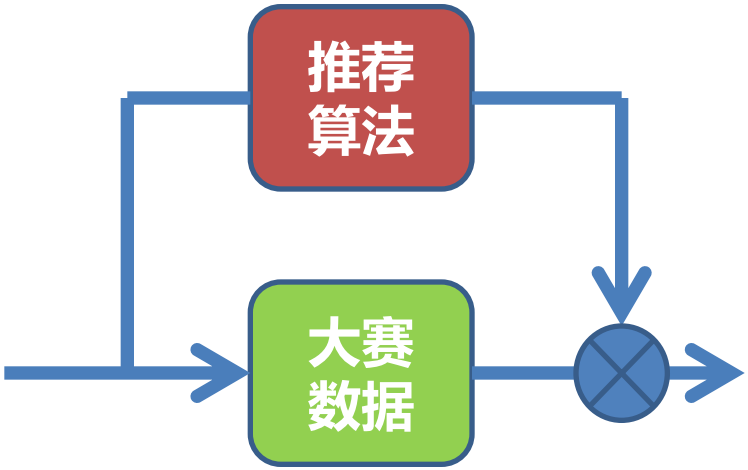
整体方案

- 不确定性：
  - ✓ 样本
  - ✓ 特征
  - ✓ 模型
- 多维度融合

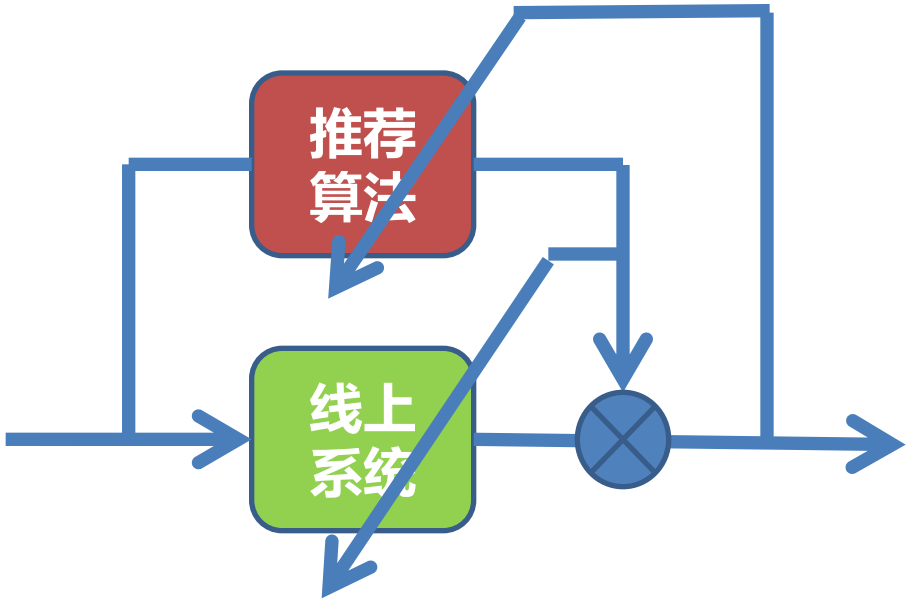


思考-与线上的区别

- 比赛/线上推荐区别
  - 从控制领域的反馈角度描述



- 比赛示意图



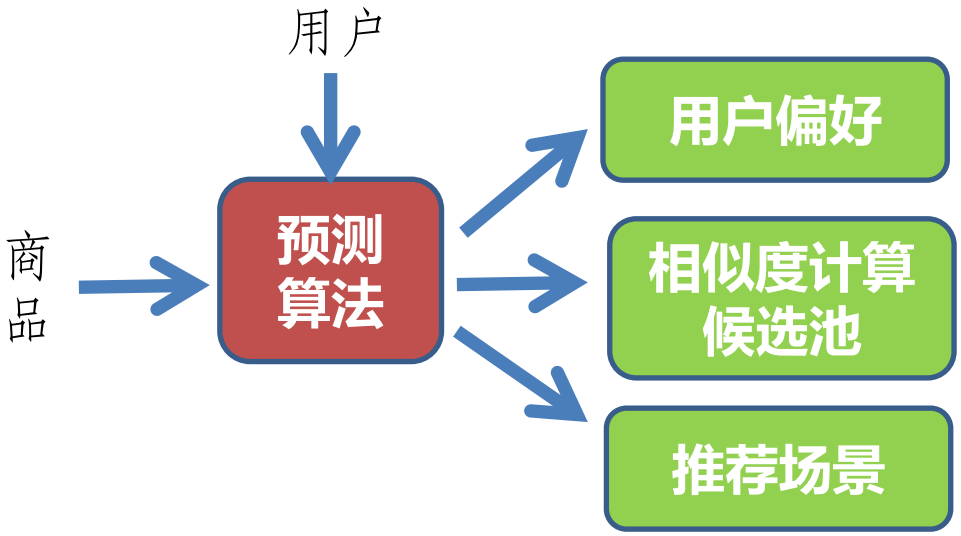
- 线上推荐示意图

思考-比赛的意义

- 预测的购买大部分为买过或近期有行为商品
- 比赛中无法模拟线上用户的交互过程

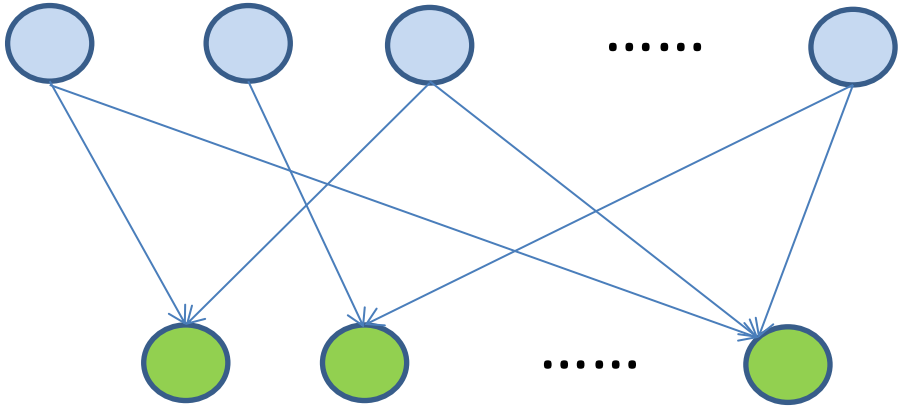
→ 比赛的意义？

意义体现在：



思考-LBS信息

- 位置信息加密
- 大量位置信息缺失
- 多位置信息

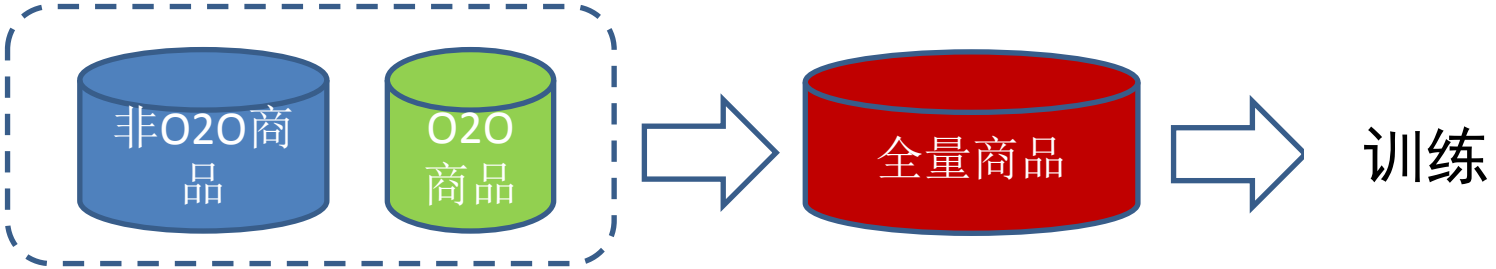


位置补全　位置恢复　社区发现

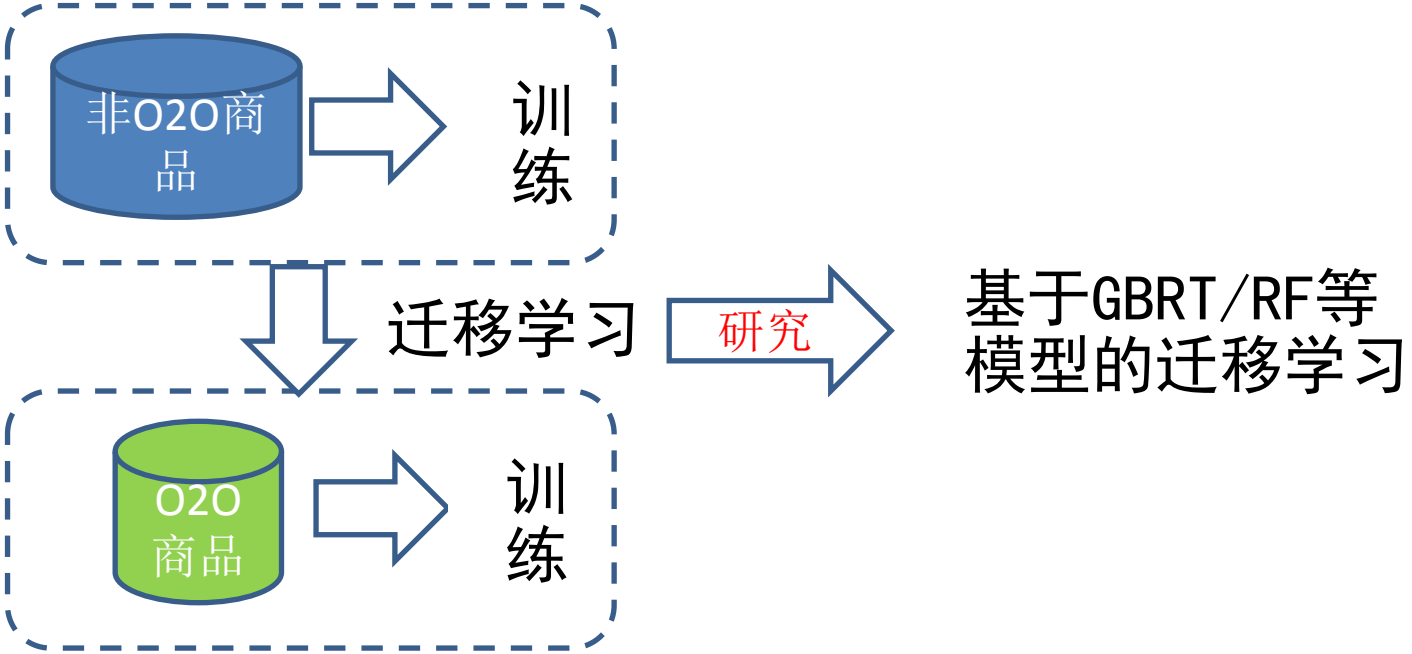
标签传播、谱聚类等

思考-迁移学习

- 常规

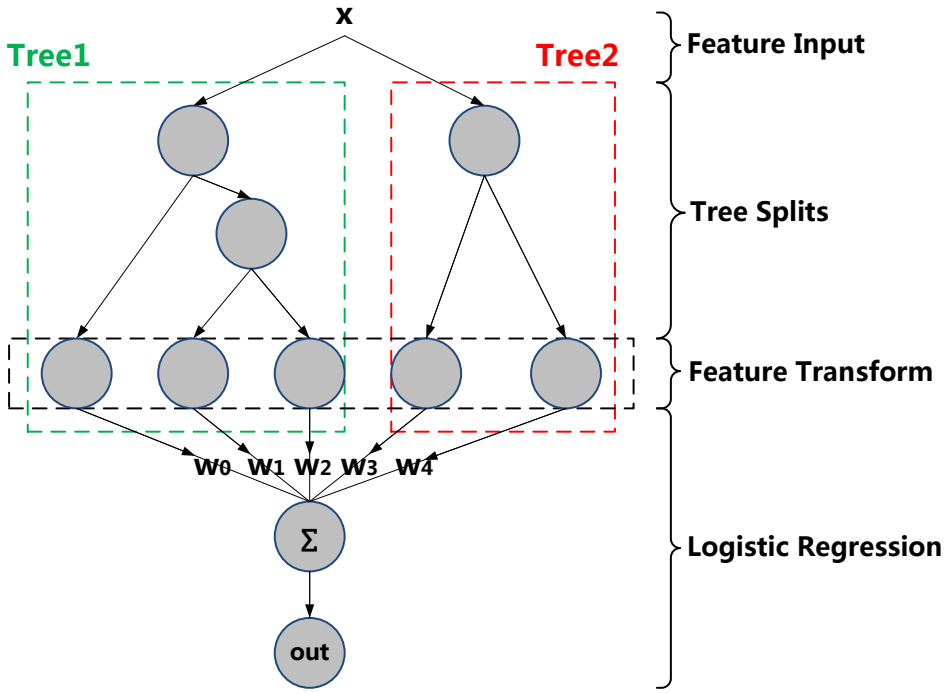


- 分而治之，协同优化



思考-LR特征离散化

- **连续特征离散化**
  - ✓ 受限于ODPS表列数限制，将连续特征部分分段离散
  - ✓ 效果得到一定提升
- **引入GBDT叶子节点离散向量**
  - ✓ 将连续特征先经过GBDT离散化，然后和离散特征一起放进LR学习
  - ✓ 受限于模型表的读取未实现



致谢

- 感谢阿里巴巴集团举办如此精彩的大数据竞赛！
- 感谢出题方及天池团队对竞赛的精心组织！
- 感谢一起参赛的小伙伴们！

新浪微博：
kevin_ning_thu