阿里巴巴大数据竞赛

天猫推荐算法大挑战

第二赛季　总决赛

阿里巴巴大数据竞赛决赛分享

数据心跳
柯国霖
厦门大学

阿里巴巴大数据竞赛
天猫推荐算法大挑战
第二赛季 总决赛

内容

- 解题思路
- 三个要点
  - 特征
  - 模型
  - 融合

问题描述

- 给定天猫用户在4月15日到8月15日的品牌交互数据
- 一共四种行为：0(点击), 1(购买), 2(收藏), 3(购物车)

user_id	brand_id	type	visit_datetime
10944750	15761	0	4月24日
10944750	15761	0	4月24日
10944750	15761	0	4月24日
10944750	15761	0	4月24日
10944750	19673	0	7月5日
10944750	19673	0	7月5日

- 预测这些用户在8月16日到9月15日购买的品牌

问题分析

- 已交互推荐
  - 分类问题 -> 一个用户品牌对：买 or 不买
  - 回归问题(ranking) -> 一个用户品牌对的购买可能性评分
- 未交互推荐
  - 协同过滤
  - 关联分析
  - 流行推荐

未交互推荐

- 协同过滤
  - 购买代价大
  - 大部分购买是因为需求，而喜好 != 需求
  - 需求单一，相似推荐效果差
- 关联规则
  - 相关性 != 因果性
  - 品牌 != 商品
  - 关联样本太少
- 放弃未交互推荐
  - Gain太少，消耗时间大
  - Solo没精力

已交互推荐

- 二分类问题
  - 正样本：交互过->买
  - 负样本：交互过->没买

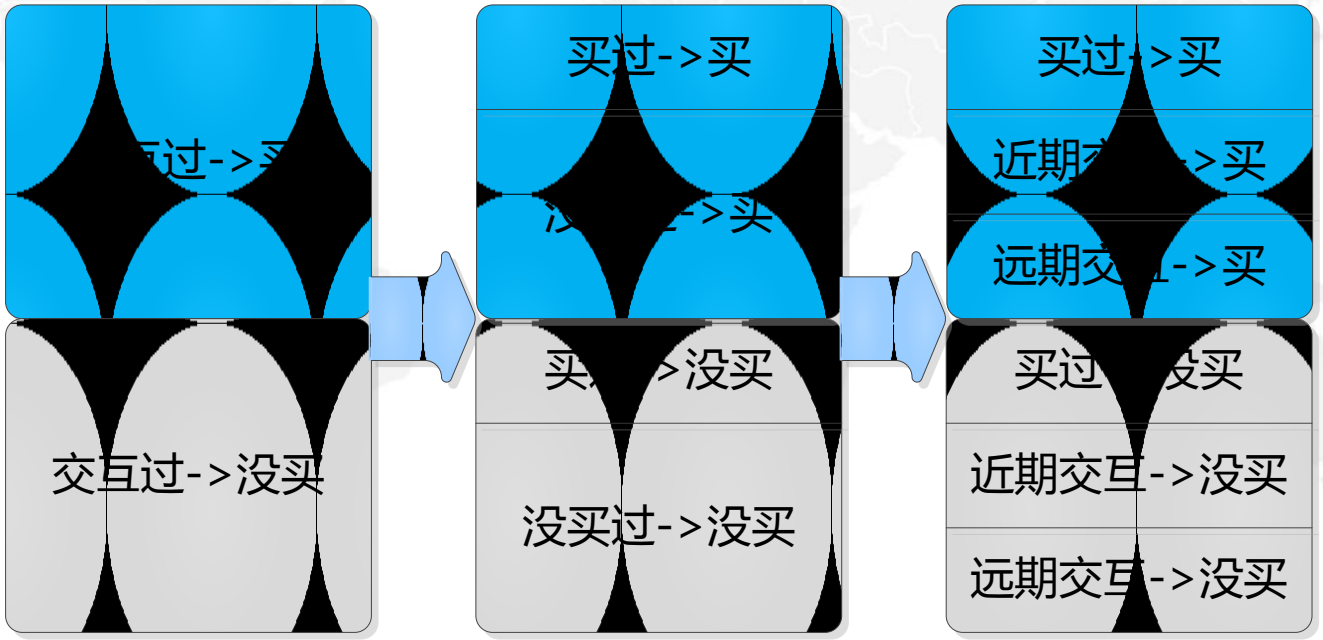
user_id	brand_id	is_buy	buy_probability
10944	1267	0	0.1453
10944	1588	0	0.0344
28444	1344	1	0.7888
28444	6888	0	0.2867

已交互推荐

- 购买的动机不一样
  - 忠实于某个品牌
  - 某个品牌口碑好
  - 促销
  - 需求
  - 心血来潮
  - ……

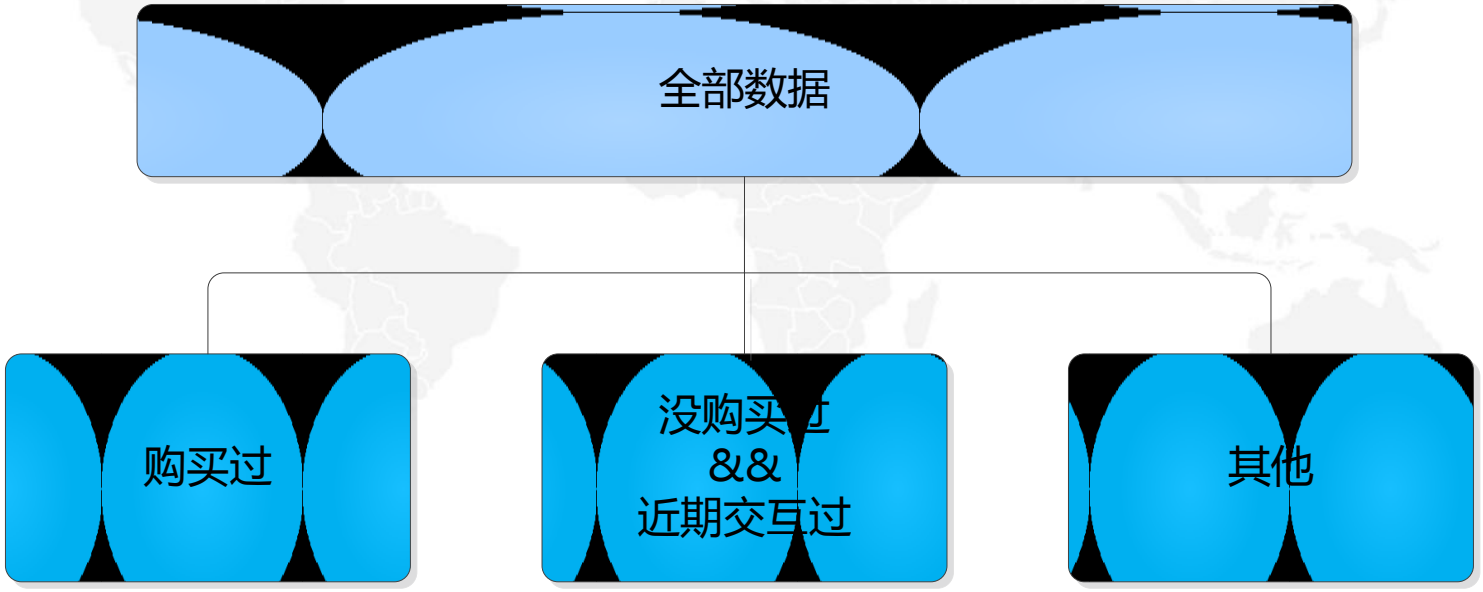
已交互推荐

- 二分类问题->多分类问题



数据拆分

- 数据切分成3块，分开训练



数据拆分

- 优点：
  - 符合实际问题
  - 充分利用3个instance
  - 主数据样本少 -> 速度快
  - 主数据噪声少 -> 精度高
  - 线上结果更优
- 缺点：
  - 子模型的预测量比例不好控制
    - 解决方法：线下调优

内容

- 解题思路
- 三个要点
  - 特征，决定UpperBound
  - 模型
  - 融合

特征设计

- 系统化设计：系统化，工程化地提取可能有用的特征
- 业务知识设计：根据业务知识，人工构造有效的特征

指标	系统化设计	业务知识设计
特征数量	多	少
信息量	大（噪声多）	较大（噪声少）
人力劳动	低	高
模型依赖度	高	低
单一特征刻画能力	弱	强
存在相似特征?	存在	不存在
适用模型	复杂模型	简单模型
适用领域	广	窄

特征设计

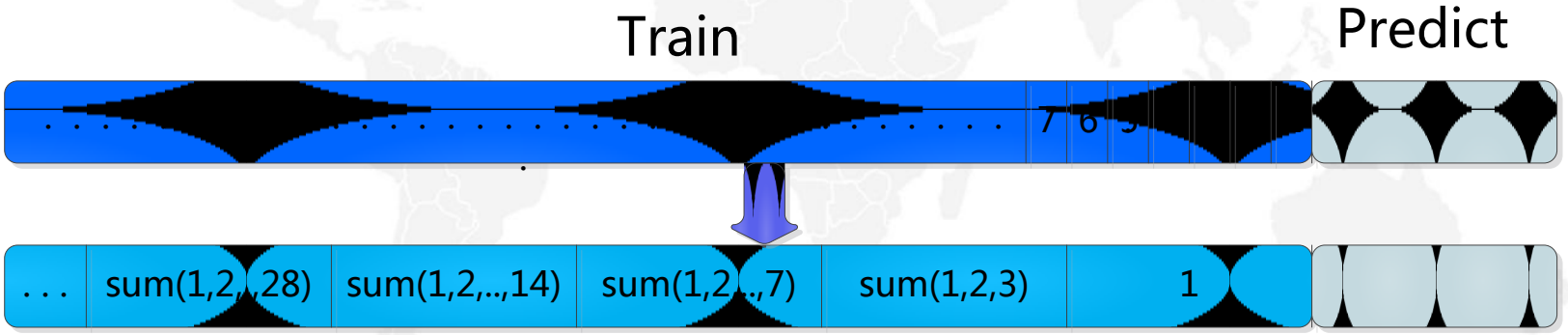
- 选择系统化设计
  - 业务知识不足
  - 人力劳动少
  - 特征变动少
  - solo思维较局限
  - 一般比赛做法

系统化特征设计

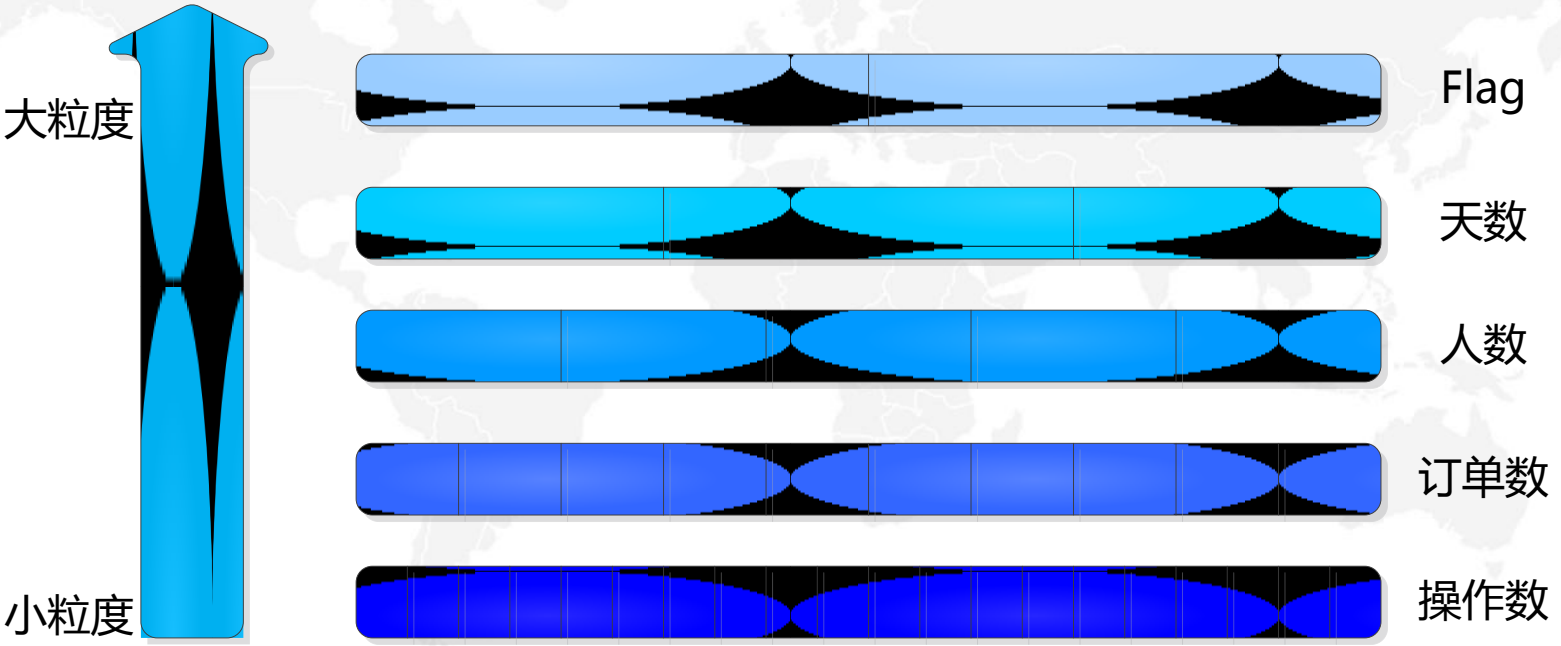
- 时间压缩
  - 降低特征维度
- 多粒度
  - 保证特征的信息量
- 交叉
  - 提高特征刻画能力

时间压缩

- 将序列信息分段压缩
- 近期密，远期粗，重叠
  - 1,3,7,14,28,56,max



多粒度

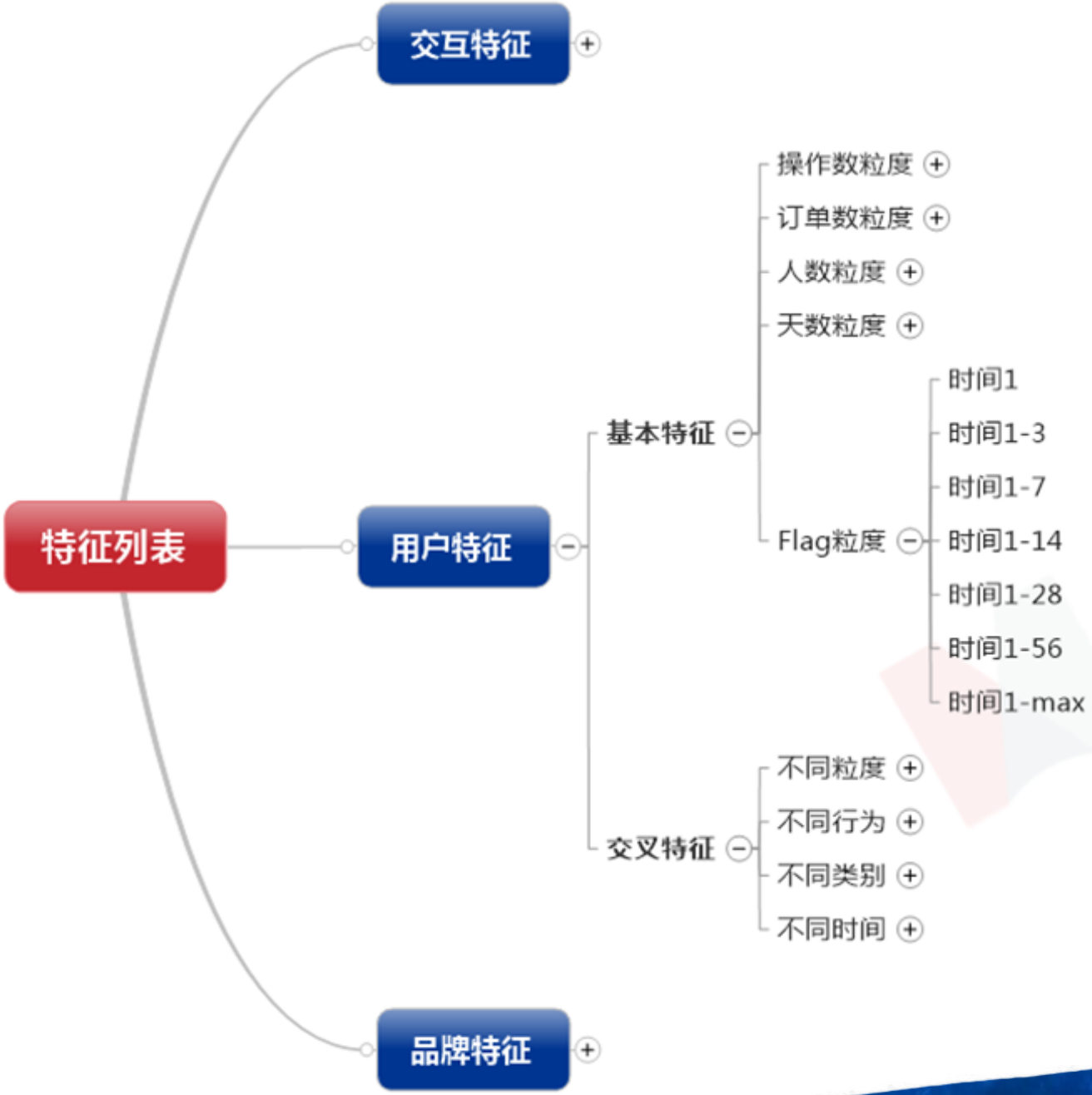


特征交叉

- 不同粒度的交叉
  - Brand的人均销量
  - Brand的日均销量
  - ……
- 不同行为的交叉
  - 转化率
  - ……

特征交叉

- 不同类别的交叉
  - 交互点击*转化率
  - 交互点击/用户点击
  - ……
- 不同时间的交叉
  - 增长率
  - 回头率
  - ……


特征列表
交互特征
用户特征
基本特征
操作数粒度
订单数粒度
人数粒度
天数粒度
Flag粒度
时间1
时间1-3
时间1-7
时间1-14
时间1-28
时间1-56
时间1-max
交叉特征
不同粒度
不同行为
不同类别
不同时间
品牌特征

数据清洗

- 目的：去掉明显的噪声，得到干净的特征统计

- 清洗对象：
  - 爬虫用户
    - 点击大于500(经验值)，且没有购买过
  - 异常用户
    - 没有点击，但有其他操作

特征处理

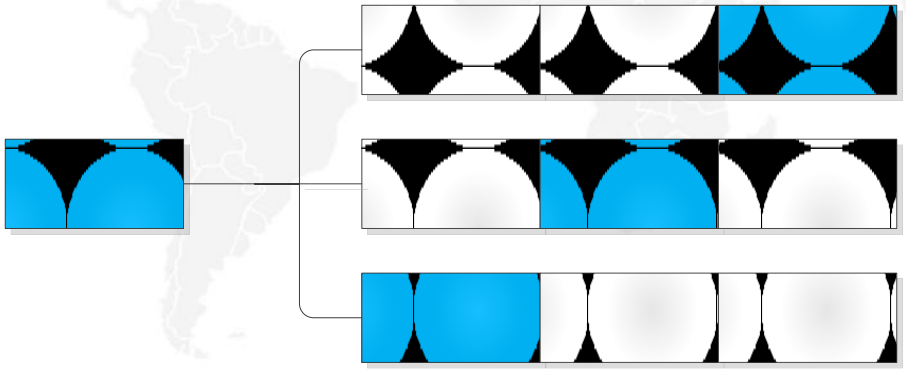
- 平滑
  - 数据缺失问题
  - Laplace平滑
    - $\frac{x}{y} \Rightarrow \frac{x+ab}{y+b}$

特征处理

- 小样本置信度
  - Wilson Score Interval
  - $$\frac{1}{1+\frac{1}{n}z^2}\left[\hat{\mathrm{p}}+\frac{1}{2n}z^2 \pm z\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})+\frac{1}{4n^2}z^2}\right]$$

特征处理

- 离散化
  - 扩展维度，解耦非线性
  - 品牌：冷门，普通，热门
  - 用户：普通，活跃



内容

- 解题思路
- 三个要点
  - 特征，决定UpperBound
  - 模型，决定接近UpperBound的程度
  - 融合

目的

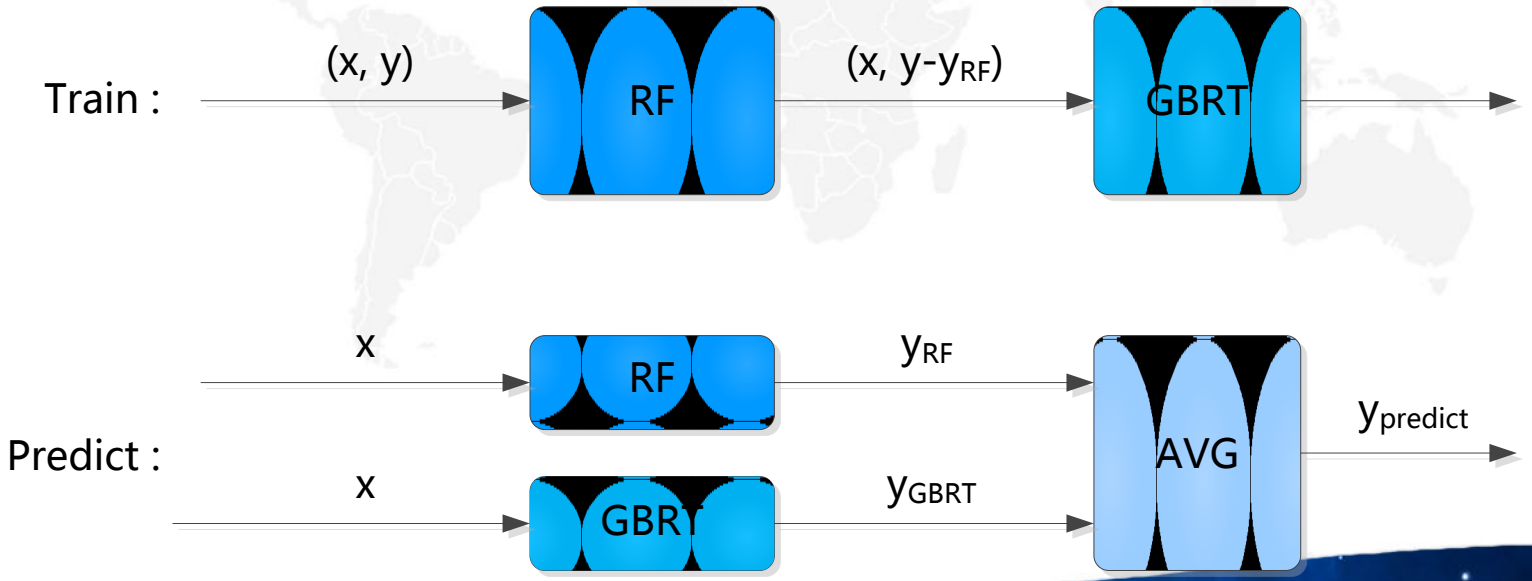
- 为融合准备的模型 -> 各种模型都要做到极致
  - 线性模型 vs 非线性模型
  - Bagging vs Boosting
  - 分类 vs 回归

基本模型

模型	输出	线性	Ensemble	速度	效果
逻辑回归(LR)	分类	线性	无	快	5.56%
随机森林(RF)	分类	非线性	Bagging	慢	6.01%
梯度渐进回归树(GBRT)	回归	非线性	Boosting	较快	6.09%

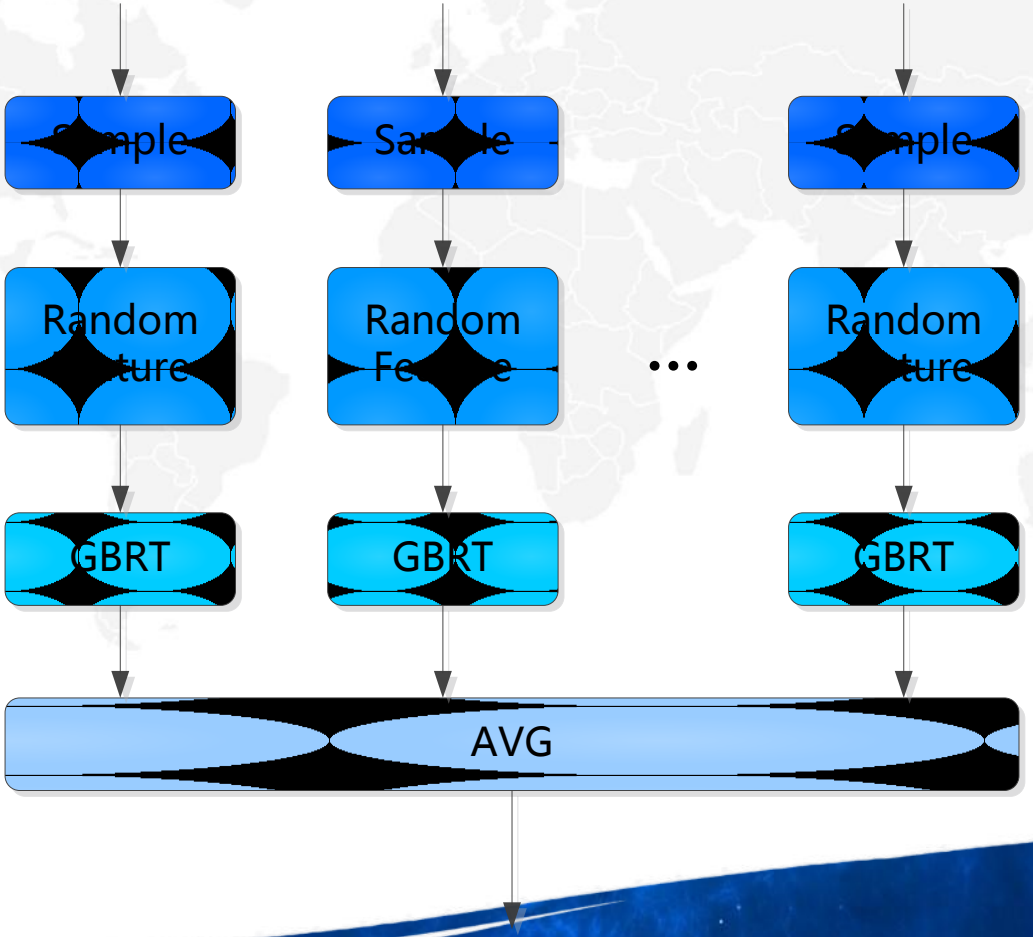
组合模型

- RF initial GBRT
  - 1. 先用RF训练，使用原始的目标值y训练，输出为y_{RF}
  - 2. GBRT训练，用RF初始化后的目标值$y-y_{RF}$训练
  - 3. 预测时，用1和2训练的RF和GBRT分别预测一次，最终结果取两者均值



组合模型

- Random GBRT Forest



内容

- 解题思路
- 三个要点
  - 特征，决定UpperBound
  - 模型，决定接近UpperBound的程度
  - 融合，更进一步接近UpperBound

融合关键

- 多样性

融合步骤

- 1. 在Local_Training集合上训练多个模型
- 2. 在Local_Test集合上调参，融合
- 3. 在Online_Training (Local_Training+Local_Test)集合上，训练单模型，使用2得到的参数
- 4. 使用2得到的融合参数进行融合

融合方法

- 人工指定
  - 对输出加权，穷举参数
  - 分别取top k，取并集
- 模型学习
  - 简单模型：逻辑回归，线性回归
  - 复杂模型：GBRT，RF_Initial_GBRT

做法多样性

- 抽样
  - 每个单模型都重新抽样训练(Bagging)
  - 不同抽样比例
- 不同长度的label区间(21天，28天，31天)
  - 扩展：滑窗构造
- 不同解决方案
  - 点击率预估 vs 购买率预估

最终使用

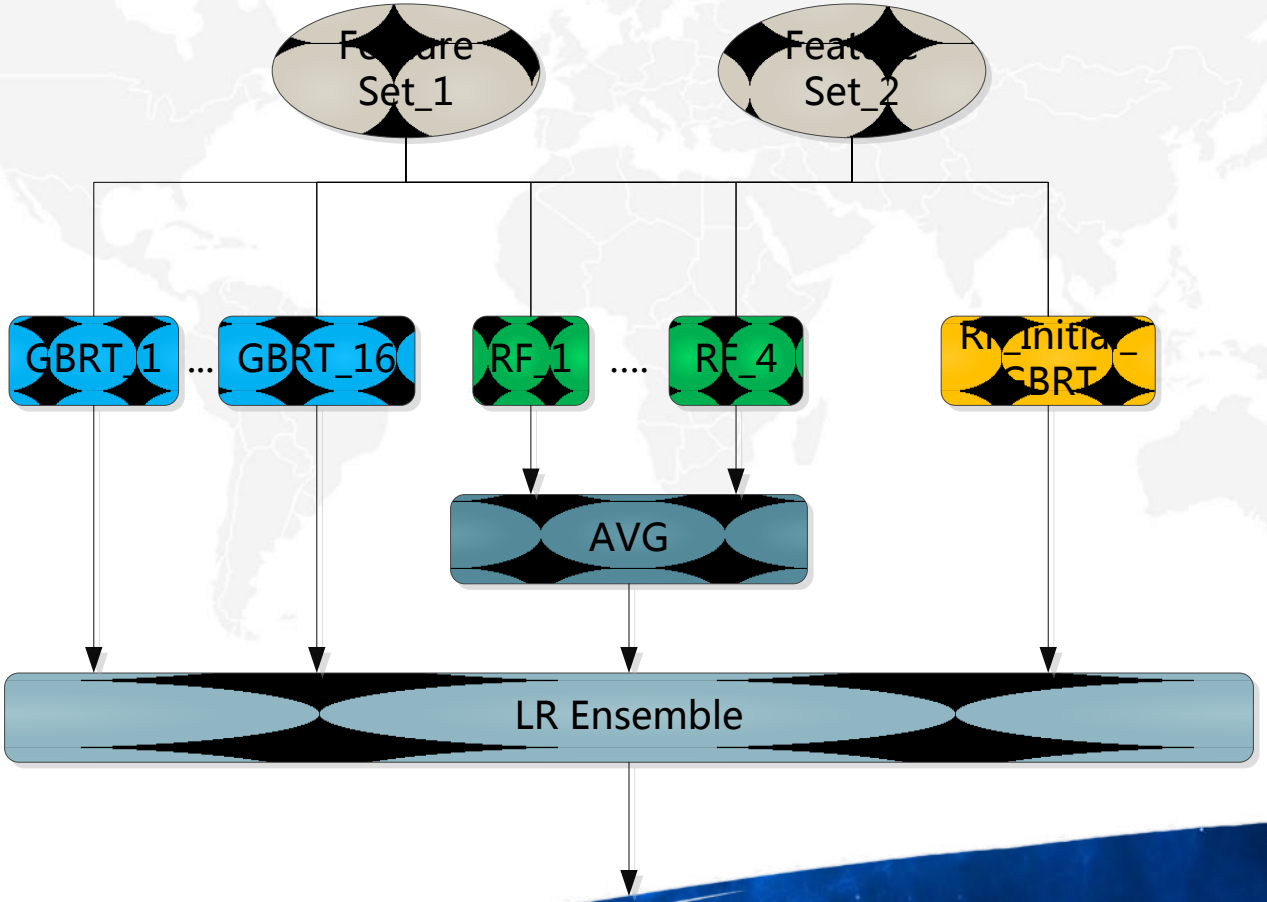
- 两组特征
  - 一组离散化，一组没有离散化
- 使用模型
  - LR (线性)
  - RF (非线性, bagging)
  - 多组参数不同的GBRT(非线性, boosting)
  - RF_Initial_GBRT
- 做法
  - 每个单模型都重新抽样训练(Bagging)
  - 不同长度的label区间(28天，31天)

最终使用

- 融合方法：LR
  - 模型多，人工指定效率低
  - 复杂模型融合线下过拟合(Time drift, 分布不一致)
  - LR 可以肉眼看系数，去掉异常模型

最终模型

- Buyed 子模型

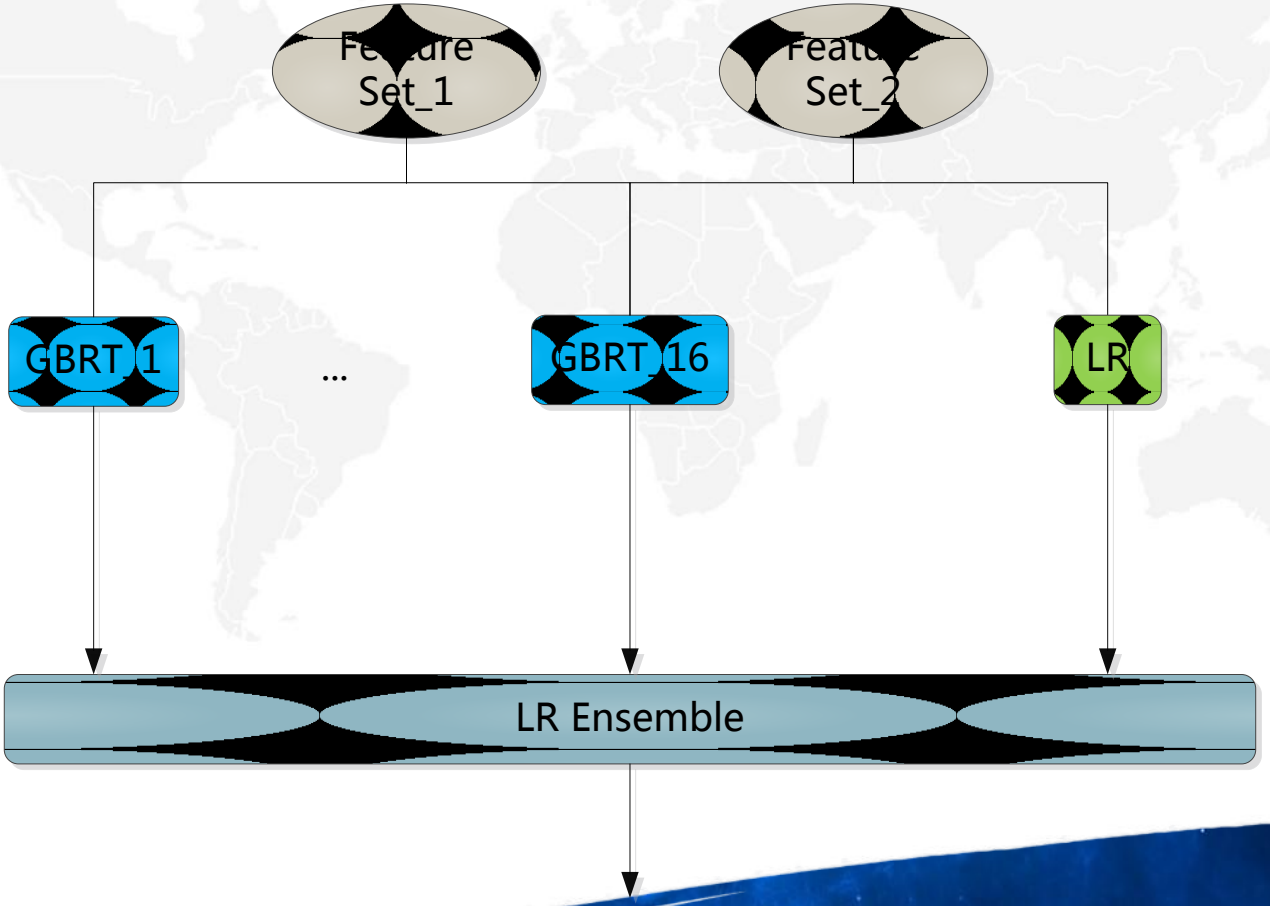


最终模型

- Short Term 子模型

最终模型

- Long Term 子模型



其他

- 融合效果取决于
  - 1 单模型的性能
  - 2 模型输出的重合度(Top K)
- 满足多样性的条件下，效果差的模型可以保留
- 去掉线下过拟合的模型
- 要对单模型的输出进行处理(z-score, ranking score)后再融合