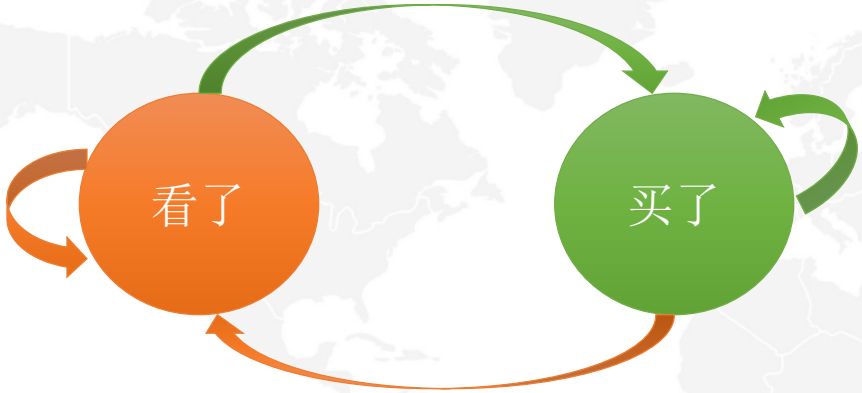
阿里巴巴大数据竞赛

天猫推荐算法大挑战

第二赛季　总决赛

参赛历程

问题思路



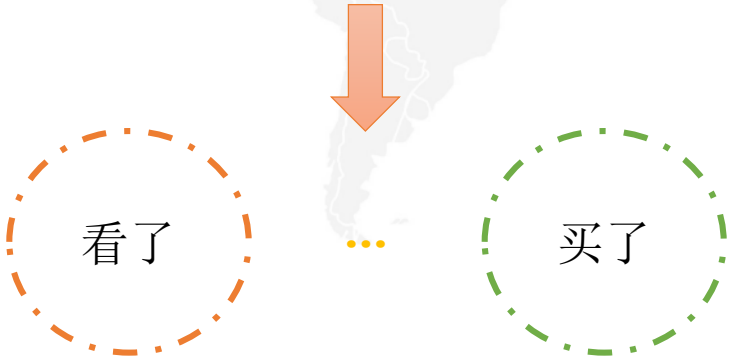
有交互

看了再看
看了再买
买了再买
买了再看

无交互

看了买了
还会看什么
还会买什么

转移成本



问题思路

Type	Label
click	0
buy	1
book	2
cart	3

Training set

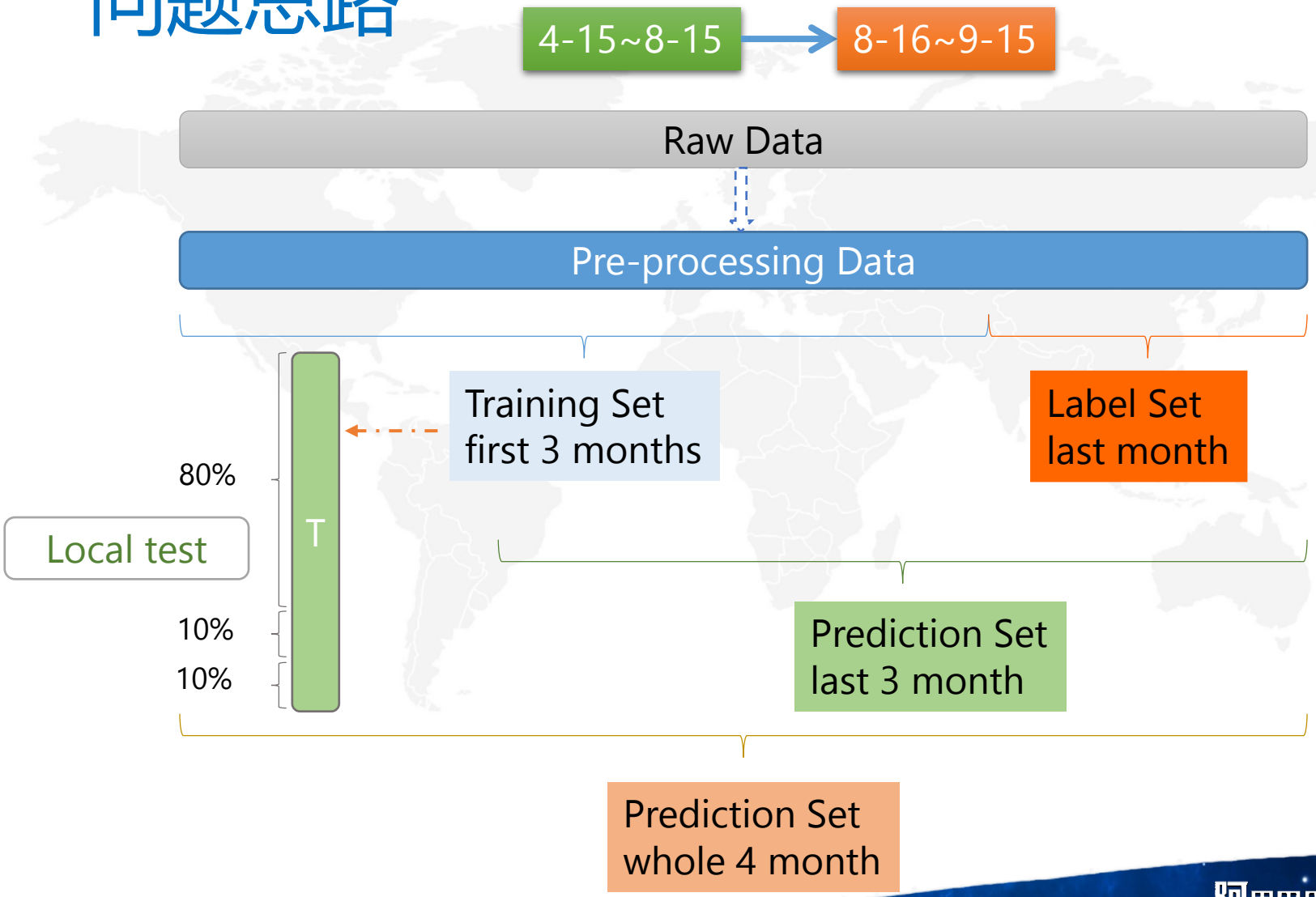
User-Brand	Is_buy	f1	f2	f3	.....	fn
⋮	0				⋮	
Useri-Brandj	1	0	2	4	⋮	1
⋮	⋮				⋮	

购买预测问题&有监督的二分类问题

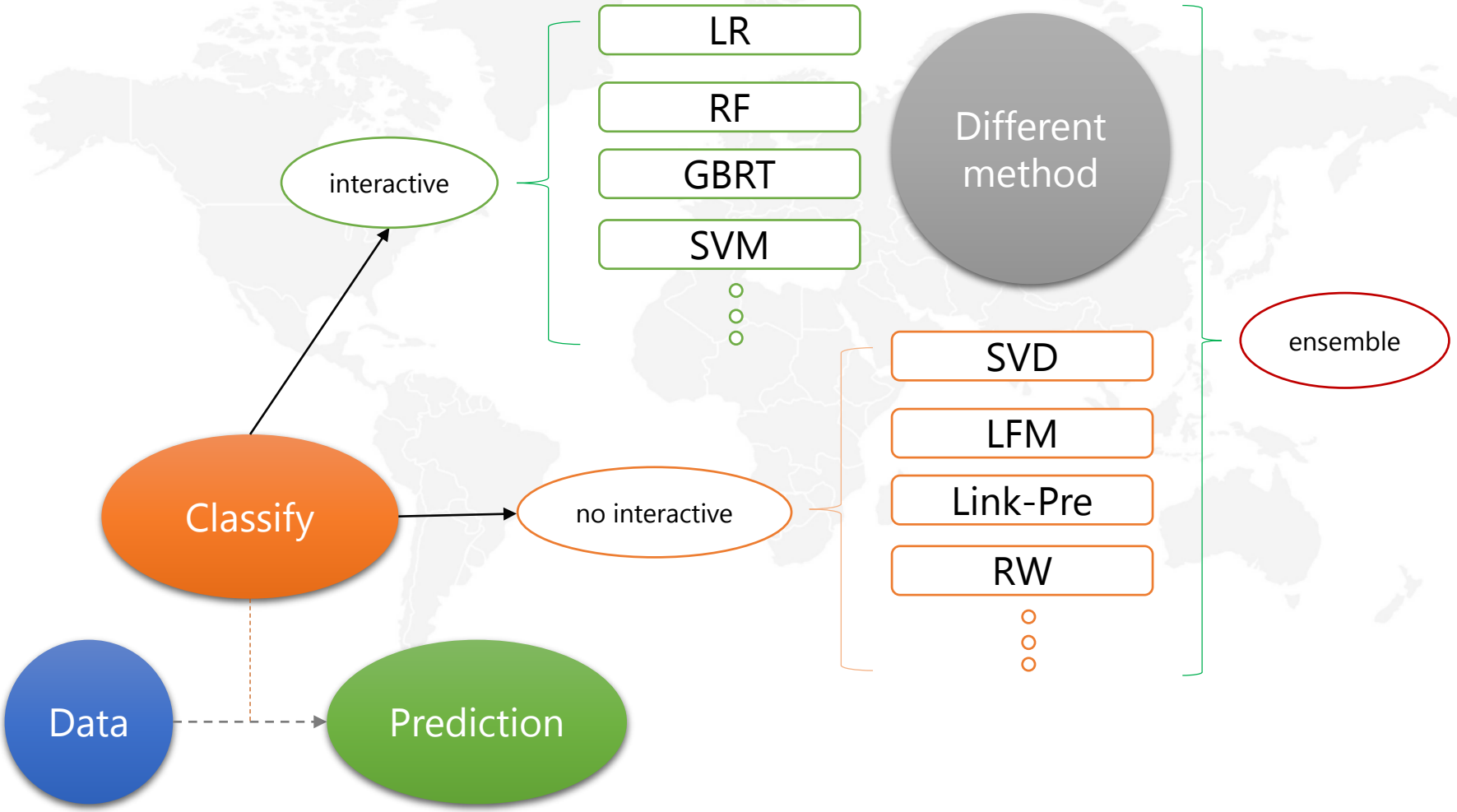
Prediction set

User-Brand	Buy_prob	f1	f2	f3	.....	fn
⋮	0.1				⋮	
$User_i$-$Brand_x$	0.9	0	1	2	⋮	1
⋮	⋮				⋮	

问题思路


4-15~8-15
8-16~9-15
Raw Data
Pre-processing Data
Training Set
first 3 months
Label Set
last month
80%
Local test
T
10%
10%
Prediction Set
last 3 month
Prediction Set
whole 4 month

问题思路


LR
RF
GBRT
SVM
Different method
interactive
ensemble
SVD
LFM
Link-Pre
RW
Classify
no interactive
Data
Prediction

优化方案

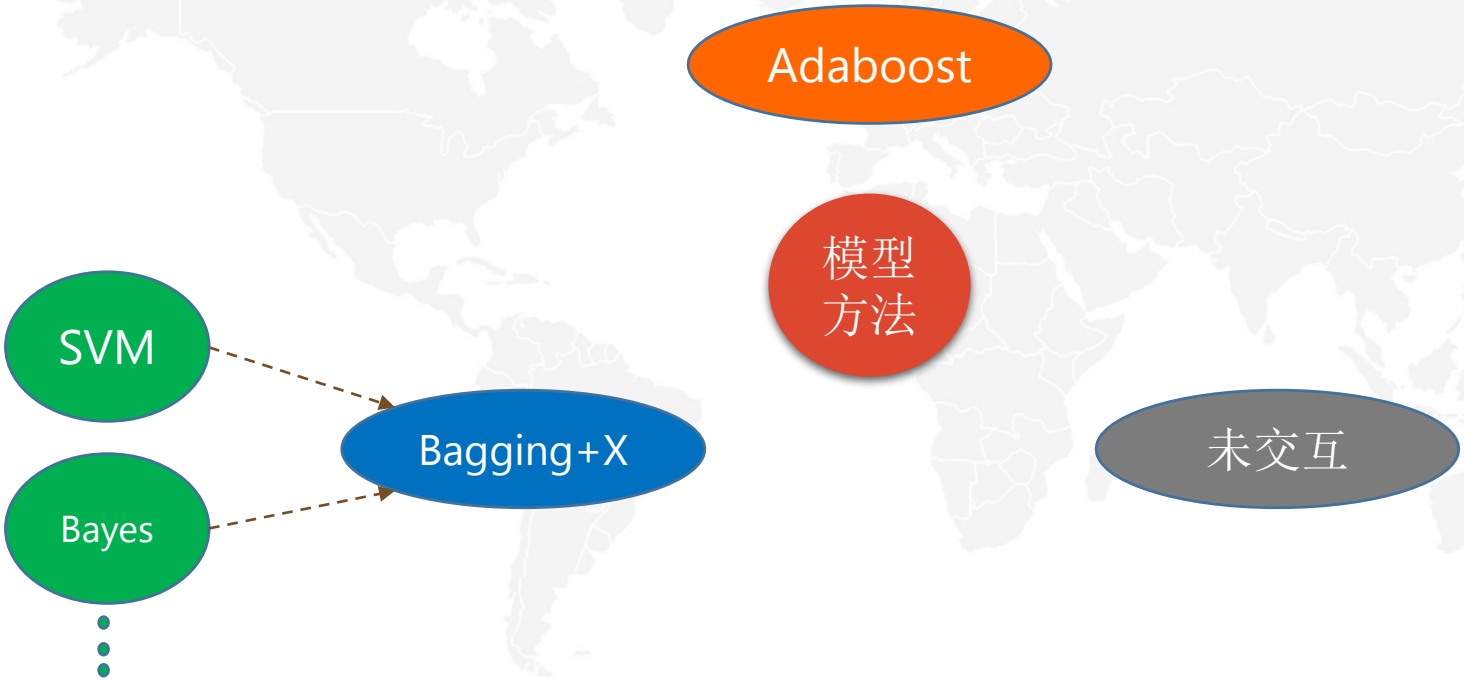
实现部分

缩小分布差异

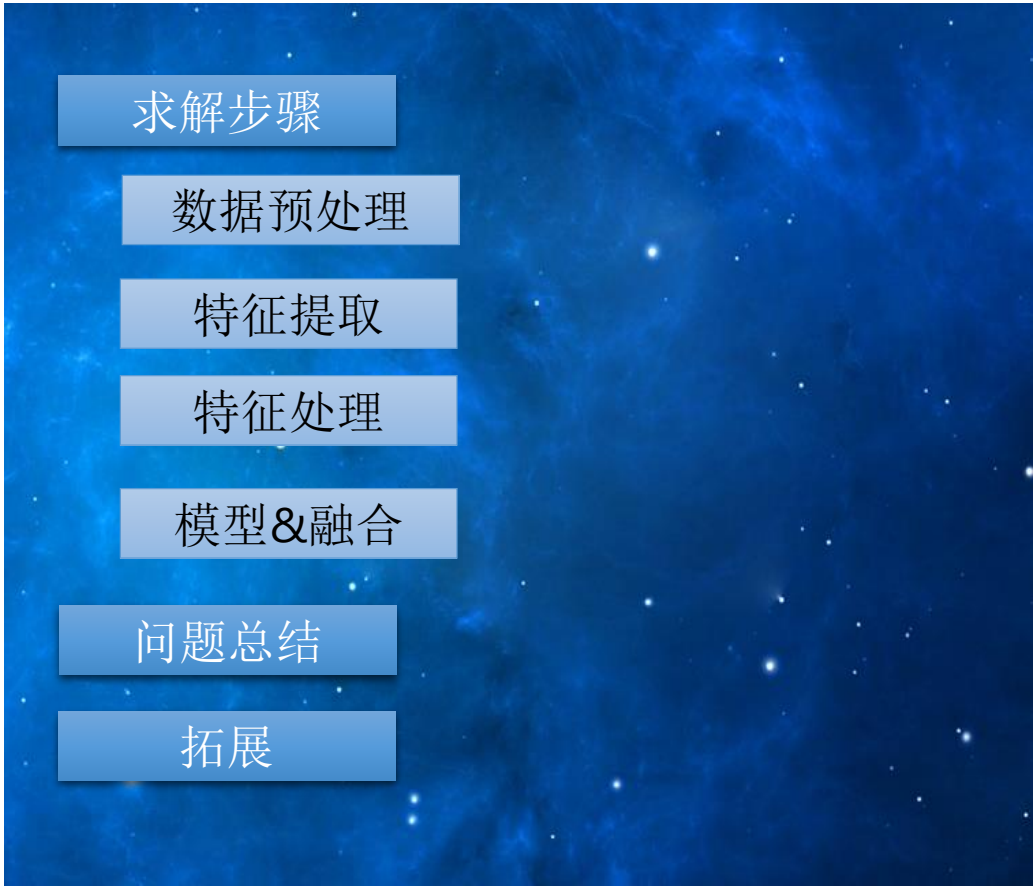
- 特征处理
  - 去除离群点
  - 归一/标准化
  - 自实现处理
- 方法
  - 结果处理
  - 模型融合
  - 类adaboost
  - 简单bagging
  - 分类
- 模型
  - 抽样比例
  - 参数设置
  - 特征筛选
  - 添加特征

优化方案

未实现部分

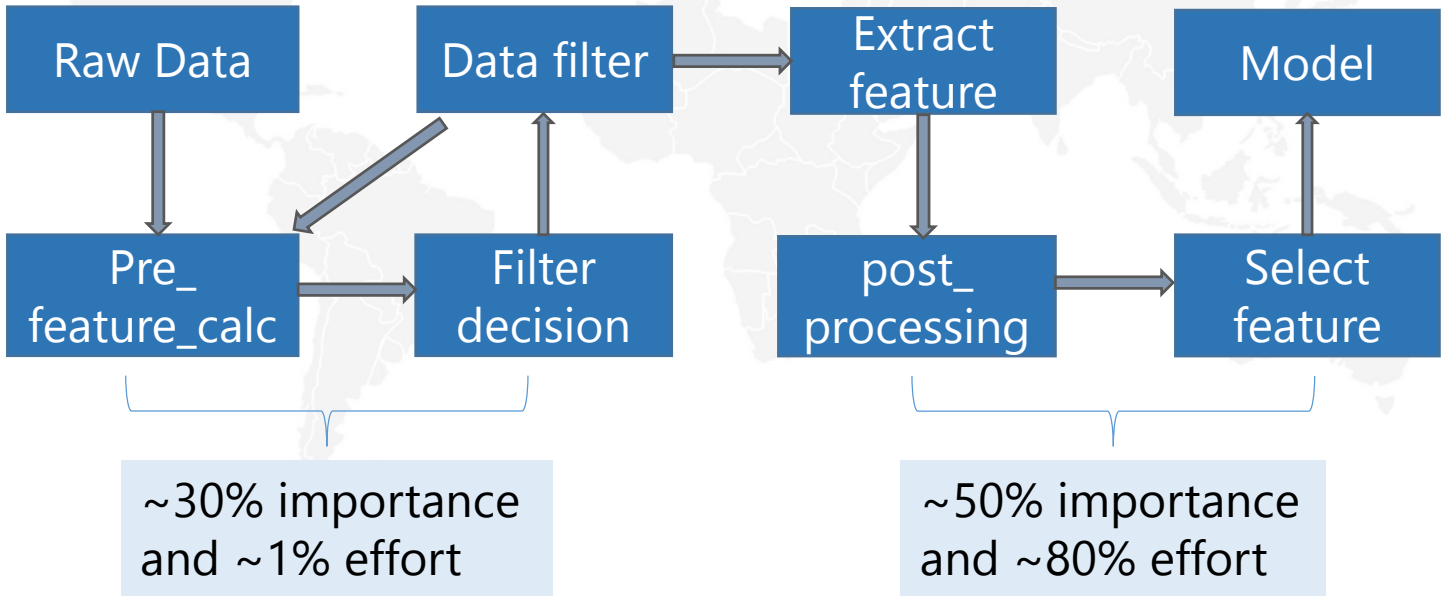


主要过程



求解步骤

- 问题求解流程：

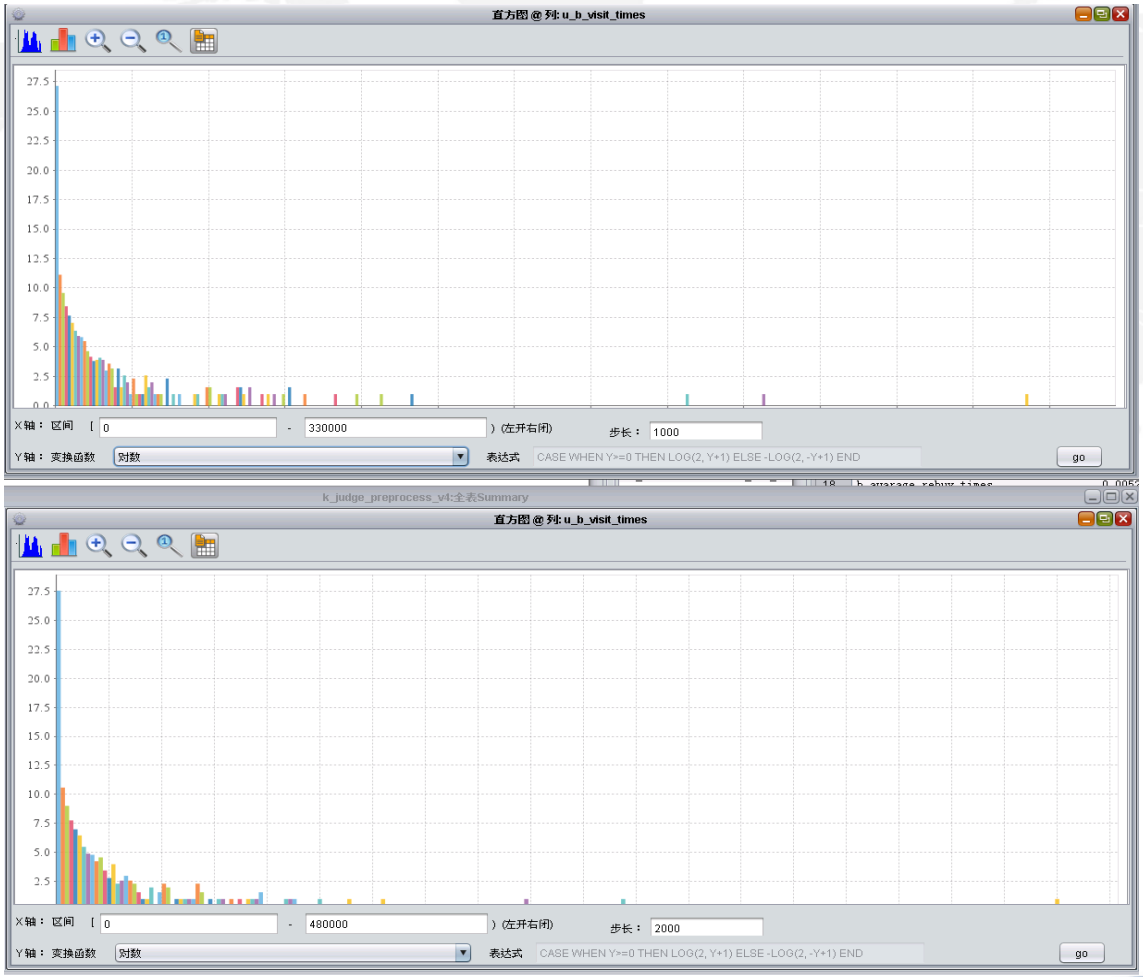

Less than 19% effort
And ~20% importance
Raw Data
Data filter
Extract feature
Model
Pre_ feature_calc
Filter decision
post_ processing
Select feature
~30% importance and ~1% effort
~50% importance and ~80% effort

求解步骤->数据预处理

- 日期转换：字符串->有意义的相对数值

- 基础特征计算：训练集&预测集
  - 目的：通过基础特征，比较两个集合的特征分布差异
  - 意义：根据差异，决定需要去除的离群样本集
  - 作用：尽量增加后续特征的泛化性，突出特征的作用

求解步骤->数据预处理


直方图 @ 列: u_b_visit_times
X轴： 区间 [0 - 330000) (左开右闭) 步长： 1000
Y轴： 变换函数 对数 表达式 CASE WHEN Y>=0 THEN LOG(2, Y+1) ELSE -LOG(2, -Y+1) END go
k_judge_preprocess_v4:全表Summary
直方图 @ 列: u_b_visit_times
X轴： 区间 [0 - 480000) (左开右闭) 步长： 2000
Y轴： 变换函数 对数 表达式 CASE WHEN Y>=0 THEN LOG(2, Y+1) ELSE -LOG(2, -Y+1) END go

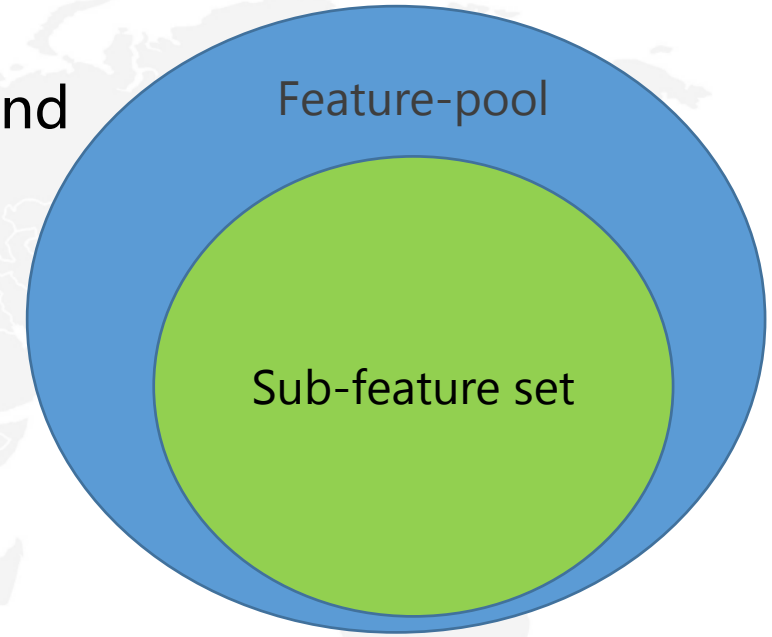

筛选方法
基本统计
直方图
均值
方差
...
训练集
预测集

求解步骤->特征提取

- Feature-pool:
  - 保存所有当前考虑到和可以求解的所有特征，而不关心这些特征中的某些部分最后是否有意义或者被选择
  - 所有特征构建围绕：品牌热门程度，用户购买能力，品牌购买周期，用户购买周期，用户重复性行为，用户行为偏好，用户交互价值，品牌价值，用户价值，行为贡献这几个总体方向去考虑
- *最终模型使用的特征是feature-pool的一个子集

求解步骤->特征提取

- Feature-pool: user,brand,user_brand
  - 基础特征：计数
    - 点击，购买，收藏，购物车
    - times,brands,users,days
  - 考虑时间因素：评分
  - 近期行为：



求解步骤->特征提取

- Feature-pool:
  - 时间特征： interactive&buy
  - 价值：存在总时间(有权，interactive&buy)
  - 周期性行为天数
  - 均值

求解步骤->特征提取

- Feature-pool:
  - 比值：
    - brand:转化率（购买/交互）
    - user:~~
    - user_brand：~~
    - ……
  - 平均值

求解步骤->特征处理

- 归一化:
  - 加快模型收敛速度，如LR
  - 统一训练集和预测集的标准
- 标准化：
  - 消除量纲，统一标准

求解步骤->模型&融合

- 模型:
  - 特征选择：
    - 与因变量（目标列）的相关系数大小
    - 特征的信息增益，gini增益，信息增益率，信息值大小
    - 特征间的相关性大小

求解步骤->模型&融合

- RF:
  - 原理
    - Bagging，多决策分类器，输出类别由个别树输出类别的众数决定。
  - ID3
    - 从信息增益最大（信息熵下降最快）的特征开始，递归执行。停止分裂的条件有信息熵为0，树深度达到要求，节点中样本个数少于指定值，熵下降幅度小于指定值，达到最大结点数；
  - CART
    - 划分依据为gini增益
  - C45
    - ID3的改进$\sqrt{M}$信息增益变成信息增益率

求解步骤->模型&融合

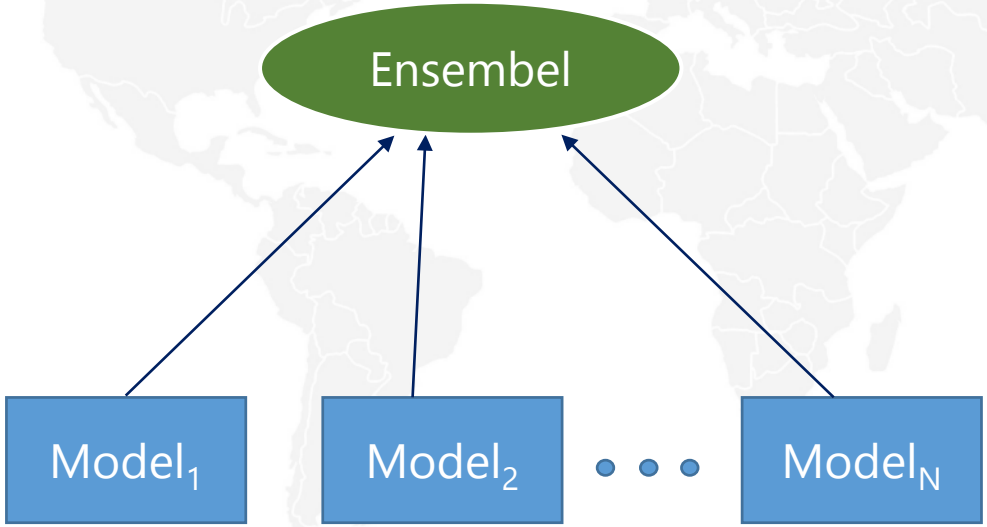
- GBRT:
  - 原理
    - Boosting，迭代决策树算法，代价函数为均方误差，结点分裂的目标是使得损失函数最小，后一棵树使用前(n-1)棵树的残差作为学习目标，最后累加所有决策树的结果作为最终结果。

求解步骤->模型&融合

- 思考:
  - 过拟合
    - 个人理解不需要刻意去减小过拟合，不一定是坏事（训练集上）
    - 这次的问题没有深深的过拟合，只有深深的分布不均和误用先验知识

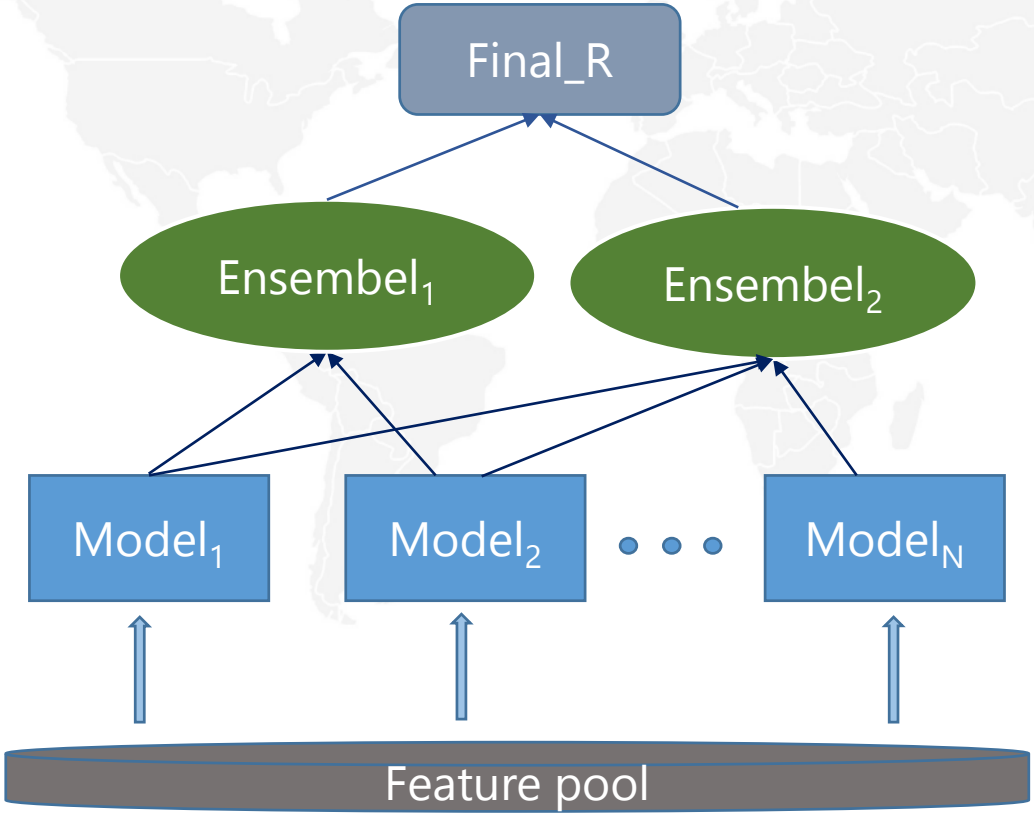
求解步骤->模型&融合

- 融合:
  - 群记录+数据观察+文献[3]



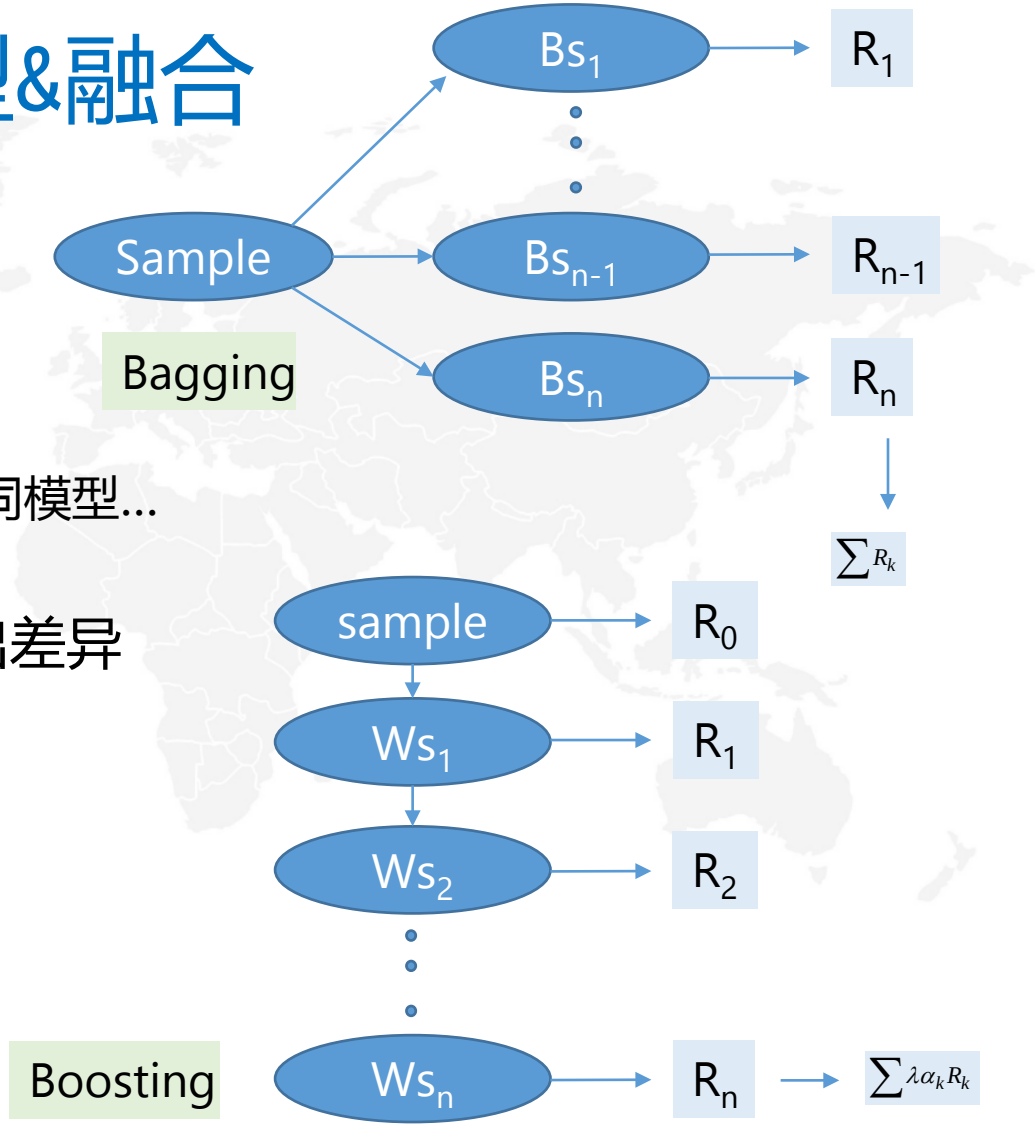
求解步骤->模型&融合

- 融合:二阶段融合[4]



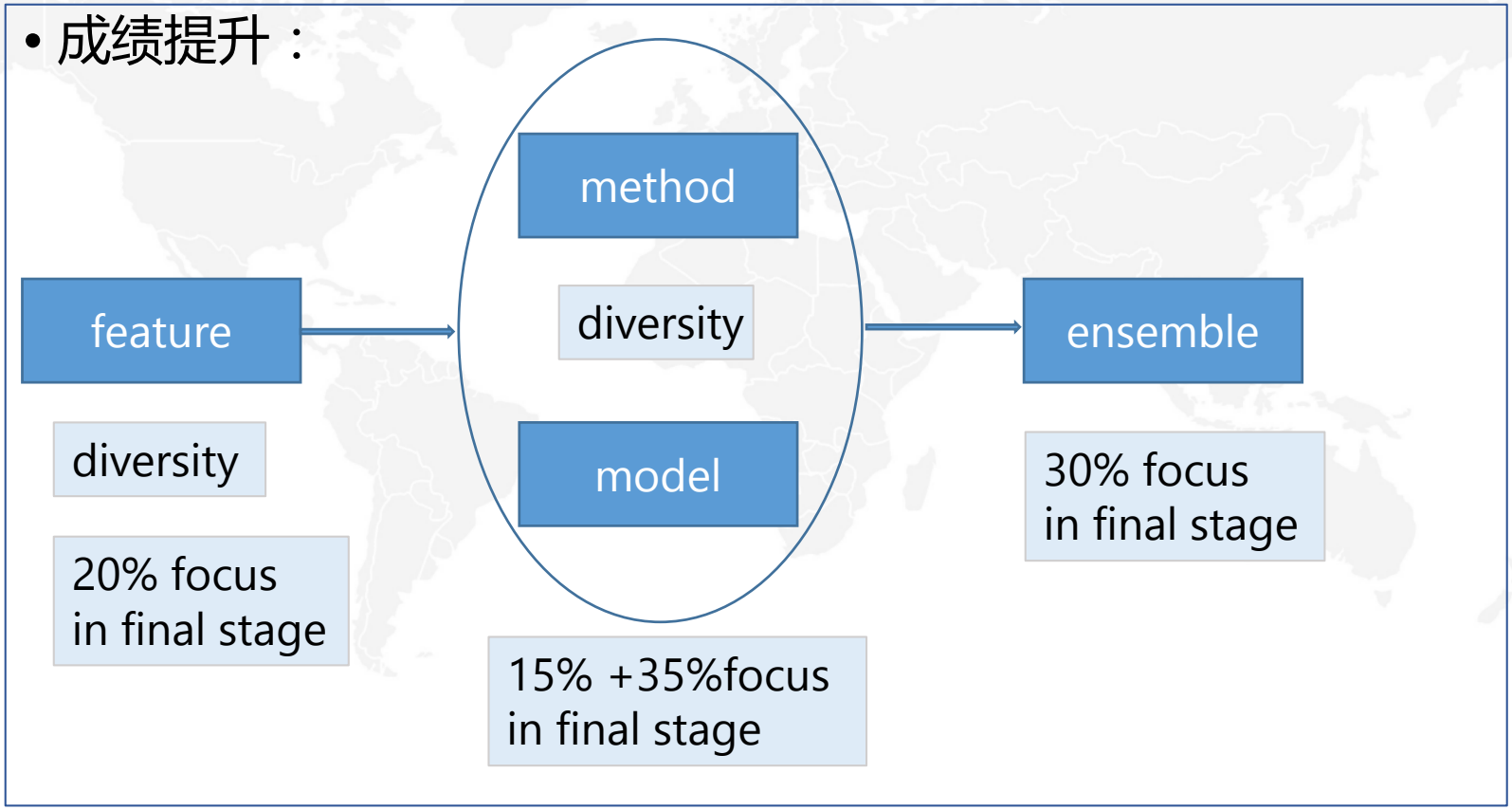
求解步骤->模型&融合

- 融合: 差异性
  - Bagging
  - Boosting
  - 不同特征, 不同采样, 不同模型...
- 结果合并: 降低模型输出差异
  - LR训练融合
  - 加权平均

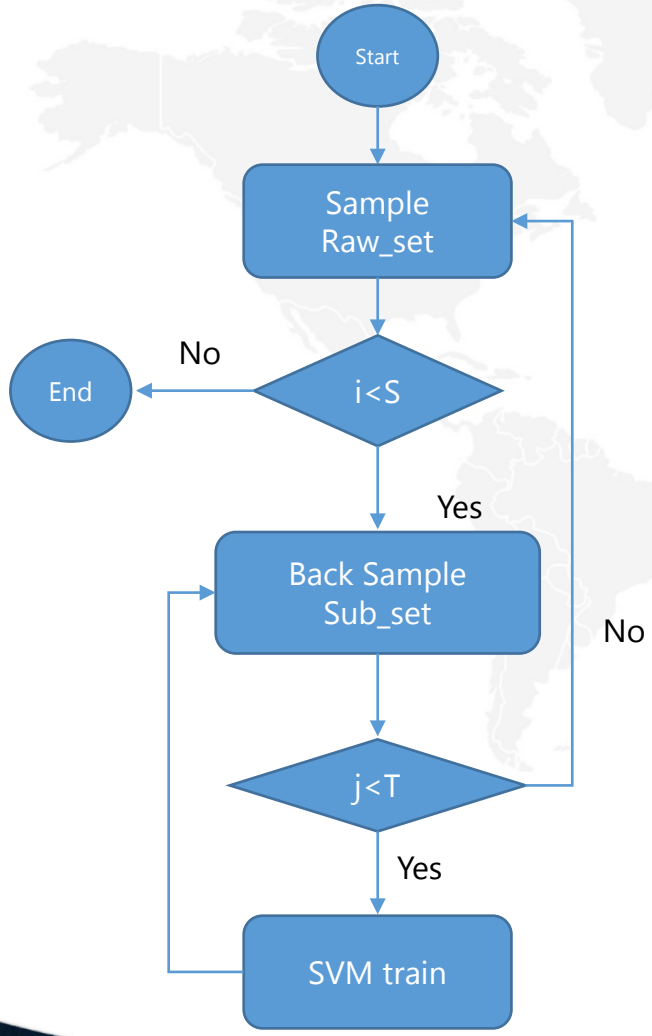


问题总结

- 成绩提升：


method
feature
diversity
ensemble
diversity
model
30% focus
in final stage
20% focus
in final stage
15% +35%focus
in final stage

拓展



Bagging&SVM

Boosting?

```
def bag_svm_train(S,T):
        for s in range(S):
                Sub_set[s]=sample(Raw_set)
                for i in range(T):
                        b_sub_set[i]=back_sample(Sub_set[s])
                        M_svm[s*T+i]=SVMtrain(b_sub_set[i])
        return M_svm
```

```
def pre_bag_svm(Model_SVM, Pre_set):
        L=len(Model_SVM)
        for i in range(L):
                pre_r[i]=Model_SVM[i](Pre_set)
        pre_bag=sum_norm(pre_r)
        return pre_bag
```

拓展

RF or GBRT Top-N prediction

User-Brand	Buy_prob
⋮	⋮
$User_i$-$Brand_x$	0.8
⋮	⋮

	$Brand_1$	$Brand_2$	.....	$Brand_n$
$User_1$	?	0.8	.....	0.9
$User_2$	0.7	?	.....	?
⋮			⋮	
$User_n$	?	0.7	.....	0.8

未交互部分采用SVD

大赛收获

大赛收获

感悟

- 信息收集
- 文献查阅
- 勤于思考
- 持之以恒

实际

- 数据处理
- 模型知识
- 泛化性影响
- 方法

大赛建议

大赛建议

参考文献

[1] Wang P., Zhou T.et al. Modeling correlated human dynamics with temporal preference. Phys. A, 2014, 398, 145-151.
[2] Schonlau M.. Boosted Regression (Boosting): An introductory tutorial and a Stata plugin. The Stata Journal, 5(3), 330-354.
[3] Zhou Z. H,. Ensemble Methods Foundations and Algorithms. 2012 by Taylor & Francis Group, LLC, 23-118.
[4] Wu K. W., Ferng C. S. et al. A Two-Stage Ensemble of Diverse Models for Advertisement Ranking in KDD Cup 2012.